

AgroGuard AI: A Trustworthy Adaptive Architecture for Crop Disease Intelligence under Environmental Non-Stationarity

Lanre Adams; Timothy Okoliko; Dr. Jennifer Okoliko

Abstract

Deep learning systems for crop disease recognition are routinely reported at above 95% accuracy on curated benchmarks, yet their field performance degrades sharply and silently once deployed. The cause is not a modelling deficiency in the conventional sense but a violation of the stationarity assumption on which supervised learning rests: the agro-climatic environment that generates the data is itself non-stationary. Humidity regimes shift between seasons, pathogen populations evolve, cultivar mixes change, sensor hardware ages, and the relationship between visual symptom and disease label drifts with them. A model frozen at training time therefore decays monotonically, and — more damagingly for trust — it decays while continuing to report high confidence.

This manuscript presents AgroGuard AI, a five-layer reference architecture that treats environmental non-stationarity as a first-class design constraint rather than a deployment nuisance. The architecture couples four mechanisms that are usually studied in isolation. A drift detection subsystem monitors input, representation, and label-conditional distributions using a complementary ensemble of ADWIN, KSWIN, and kernel maximum mean discrepancy tests. A continual learning subsystem combines elastic weight consolidation with sparse episodic replay so that adaptation to a new agro-climatic regime does not erase competence in previously encountered ones. A conformal prediction layer converts raw softmax scores into prediction sets with distribution-free marginal coverage guarantees, and gates low-reliability inputs to human review rather than issuing a confident wrong answer. Finally, a temporal explanation engine reports not only why a prediction was made but why it *changed*, surfacing the shift in feature attribution that accompanies each adaptation event. We formalise the learning problem, specify each subsystem, derive the adaptation trigger policy, and define a trust metric suite covering calibration error, explanation fidelity, backward transfer, and selective risk. We further specify an edge deployment profile for low-connectivity rural contexts and a privacy-preserving federated aggregation scheme. The manuscript closes with a complete evaluation protocol — datasets, baselines, ablations, and acceptance thresholds — designed to falsify the architecture's central claims. All quantitative values reported herein are design targets and illustrative simulations under explicitly stated generative assumptions; no field trial has been conducted, and this limitation is treated as a defining boundary of the present contribution rather than a caveat.

Keywords: Trustworthy Artificial Intelligence; Concept Drift; Continual Learning; Conformal Prediction; Explainable AI; Precision Agriculture; Climate Resilience; Edge Computing; Federated Learning; Crop Disease Detection.

I. INTRODUCTION

1.1 Motivation and problem context

Pathogens and pests remove an estimated 10.1–28.1% of global wheat production, 24.6–40.9% of rice, 19.5–41.1% of maize, 8.1–21.0% of potato, and 11.0–32.4% of soybean each year [1]. These losses fall disproportionately on food-insecure regions where smallholder farming dominates and where diagnostic extension services are thinnest. The economic argument for automated disease diagnosis is therefore straightforward, and the computer vision community has responded: since the release of the PlantVillage corpus [2], convolutional and transformer-based classifiers have repeatedly reported accuracies above 99% on held-out splits [3], [4].

That number is misleading, and the discrepancy between benchmark and field performance is now well documented. Barbedo demonstrated that models trained on laboratory-condition leaf images — uniform backgrounds, single lesions, controlled illumination — lose a large fraction of their accuracy on field photographs with occlusion, variable lighting, and multiple co-occurring conditions [5], [6]. Ferentinos observed similar degradation across acquisition conditions [3]. Ramcharan et al., working on cassava in Tanzania, found that in-field accuracy fell substantially below the laboratory figure even with the same model and the same disease classes [7].

The literature has largely diagnosed this as a dataset problem to be solved with more and better field data. We argue that framing is incomplete. Even with a perfectly representative training set collected in situ, performance will decay over deployment time, because the data-generating process is non-stationary in ways no finite training set can anticipate.

1.2. Environmental non-stationarity as the central obstacle

Four distinct mechanisms drive this decay, and they are worth separating because they call for different countermeasures.

- **Seasonal covariate shift.** The marginal distribution of inputs changes cyclically. Lesion morphology under 80% relative humidity differs from the same pathogen's expression at 45%; canopy density, leaf age, and background vegetation all follow the phenological calendar. The input distribution $p(x)$ moves while $p(y|x)$ remains approximately fixed.
- **Concept drift proper.** Pathogen populations evolve. Fungicide resistance alters symptom expression; new races of *Puccinia striiformis* and *Magnaporthe oryzae* emerge on multi-year timescales. Here $p(y|x)$ itself changes — the same visual evidence now implies a different label or a different severity.
- **Inter-regional domain shift.** A model trained in humid tropical West Africa transfers poorly to highland East Africa even for nominally identical diseases, because soil reflectance, cultivar genetics, and co-occurring abiotic stresses differ. This is the classical domain adaptation setting [8], [9].
- **Climate-trend drift.** Superimposed on the seasonal cycle is a secular trend. IPCC AR6 WGII documents poleward and upslope expansion of pest and pathogen ranges under warming [10]. Unlike seasonal shift, this component does not revert; a system that merely oscillates between two learned seasonal modes will progressively fall behind it.

Standard practice addresses none of these. A model is trained once, validated once, and deployed indefinitely. Degradation is detected — if at all — by complaint, long after agronomic damage is done. Worse, modern neural networks are known to be overconfident under distribution shift [11], [12], so the failure is not merely silent but actively misleading: the system reports high confidence precisely where it should abstain.

The governing design question

How should an agricultural AI system be architected so that it detects its own degradation under environmental change, adapts without forgetting what it previously knew, quantifies when it should decline to answer, and explains to a non-expert user not only its decision but the fact that its reasoning has shifted?

1.3 Contributions

This manuscript makes five contributions.

- **C1 — A unified architecture.** We specify a five-layer reference architecture (Section 4) that integrates drift detection, continual learning, conformal uncertainty quantification, and temporal explainability as mutually dependent components under a single control loop, rather than as independent add-ons.
- **C2 — A drift-triggered adaptation policy.** We derive an explicit trigger function over an ensemble of distributional tests, with hysteresis and a cost term that prevents oscillatory retraining (Section 5.4).
- **C3 — Trust-constrained adaptation.** We formulate adaptation as constrained optimisation in which calibration error, backward transfer, and explanation stability are hard constraints rather than post-hoc diagnostics (Section 3.4), so that a model update which improves accuracy but destroys calibration is rejected by construction.
- **C4 — Temporal explainability.** We define an explanation protocol that reports attribution *change* across adaptation events, which we argue is the explanation primitive that matters for a longitudinally deployed system and which static XAI methods do not provide (Section 7.2).
- **C5 — A falsifiable evaluation protocol.** We specify datasets, baselines, ablations, metrics, and numeric acceptance thresholds (Section 10) sufficient for an independent group to reproduce and, importantly, to refute the architecture's claims.

1.4 Scope, and the status of reported values

Statement on empirical status — please read before interpreting any figure

This is a systems and architecture manuscript, not an experimental report. No field deployment, user study, or controlled benchmark has been executed. Every numeric value appearing in Figures 2–8 and in Tables 6 and 9 is one of two things, and each is labelled as such in situ: (i) a design target that the architecture is specified to meet, or (ii) an illustrative simulation produced under generative assumptions stated in the accompanying caption. These values demonstrate the measurement methodology and the shape of the expected result; they are not evidence that the architecture achieves them. The evaluation protocol in Section 10 exists precisely to supply the evidence that this manuscript does not. Readers assessing this work for research potential should treat the architecture, the formulation, and the protocol as the contribution, and should treat the absence of empirical validation as the principal open item, which Section 12 addresses directly.

The remainder of the manuscript is organised as follows. Section 2 reviews related work and identifies the integration gap. Section 3 formalises the learning problem. Section 4 presents the architecture. Sections 5 through 9 specify the adaptation engine, the uncertainty layer, the explanation engine, the resilience controller, and the edge deployment profile respectively. Section 10 gives the evaluation protocol, Section 11 the sustainability and governance position, Section 12 the limitations, and Section 13 the research roadmap.

II. RELATED WORK AND GAP ANALYSIS

2.1 Deep learning for plant disease recognition

Mohanty et al. established the modern baseline, training AlexNet and GoogLeNet on 54,306 PlantVillage images across 38 classes and reporting 99.35% top-1 accuracy [2]. Ferentinos extended this to 87,848 images and 58 classes with comparable results [3]. Subsequent work has substituted stronger backbones — EfficientNet [13], Vision Transformers [4] — with incremental benchmark gains. Kamilaris and Prenafeta-Boldú [14] and Liakos et al. [15] survey the broader application of deep learning and machine learning in agriculture.

The critical counterpoint comes from Barbedo, who showed that the high benchmark figures are substantially an artefact of dataset construction: PlantVillage images are single leaves on uniform backgrounds, and models trained on them exploit background and acquisition cues that do not transfer [5]. His later work on lesion-level rather than leaf-level classification [6] and the Plant Pathology Challenge dataset [16] represent attempts to close this gap. None of this line of work, however, addresses temporal degradation after deployment — the benchmark remains a single fixed split.

2.2 Concept drift and dataset shift

The statistical learning literature has studied non-stationarity extensively. Quiñero-Candela et al. [17] and Moreno-Torres et al. [18] provide the taxonomy of dataset shift that we adopt in Section 3. Gama et al. survey drift adaptation methods comprehensively [19]. For detection, ADWIN maintains an adaptive window and signals change when two sub-windows differ beyond a Hoeffding bound [20]; KSWIN applies a Kolmogorov–Smirnov test over a sliding window and is more sensitive to abrupt distributional change [21]; kernel maximum mean discrepancy provides a non-parametric two-sample test suitable for high-dimensional representations [22]. These are mature tools, but they have seen little application in agricultural vision, where the drift is unusually well characterised a priori by the phenological calendar — a structural prior that generic detectors do not exploit.

2.3 Continual learning

Naive fine-tuning on new data induces catastrophic forgetting. Three families of remedy exist. Regularisation methods constrain parameter movement in directions important to prior tasks: elastic weight consolidation uses the diagonal Fisher information as an importance estimate [23], synaptic intelligence accumulates path integrals of parameter importance [24], and learning without forgetting distills the previous model's outputs [25]. Rehearsal methods retain a small exemplar buffer: iCaRL [26], gradient episodic memory [27], and the tiny episodic memory analysis of Chaudhry et al. [28] established that even very small buffers materially reduce forgetting. Architectural methods allocate task-specific capacity. Parisi et al. survey the field [29].

Agricultural deployment has a distinguishing property that makes the choice easier than in generic continual learning: tasks recur. A model that encounters monsoon conditions this year will encounter them again next year. This argues for combining regularisation (to protect prior competence) with a small replay buffer stratified by agro-climatic regime (to reinstate it quickly), which is the design we adopt in Section 5.2.

2.4 Domain adaptation

Unsupervised domain adaptation aligns source and target feature distributions without target labels. Deep adaptation networks minimise multi-kernel MMD between domain representations [8]; domain-adversarial training learns features a domain

discriminator cannot separate [9]; adversarial discriminative domain adaptation generalises the adversarial formulation [30]. Test-time adaptation methods such as Tent [31] and test-time training [32] adapt using only the unlabelled test stream, which is attractive for edge deployment where labels are unavailable. We adopt a hybrid: adversarial alignment during cross-region transfer, entropy-based test-time adaptation at the edge under connectivity loss.

2.5 Explainable AI in agricultural decision support

Grad-CAM produces class-discriminative localisation maps from gradient-weighted activations [33]; SHAP provides additive feature attributions with desirable axiomatic properties [34]; LIME fits local surrogate models [35]. These are the standard instruments and are widely applied to plant disease models. Two critiques bear directly on our design. Adebayo et al. showed that several saliency methods are insensitive to model and data randomisation — they produce plausible-looking maps even for untrained networks — and proposed sanity checks that any deployed explanation method should pass [36]. Rudin argues more strongly that post-hoc explanation of black boxes is inappropriate for high-stakes decisions [37]. Doshi-Velez and Kim set out the evaluation framework for interpretability claims [38]. We take these seriously: Section 7.5 specifies an explanation audit that includes Adebayo's sanity checks as a gate, and we are explicit that our explanations are decision-support aids rather than agronomic justifications.

What is absent from this literature is any treatment of explanation under adaptation. Every method cited explains a fixed model. None addresses what a user should be told when the model that produced yesterday's explanation is not the model producing today's.

2.6 Uncertainty quantification and selective prediction

Guo et al. documented systematic overconfidence in modern networks and showed that temperature scaling largely corrects it in-distribution [11]. Ovadia et al. showed that this correction degrades under distribution shift — precisely the regime of interest here [12]. Deep ensembles [39] and MC-dropout [40] provide epistemic uncertainty estimates at computational cost. Conformal prediction [41], [42] offers a different guarantee: distribution-free finite-sample marginal coverage, under exchangeability. Adaptive prediction sets [43] and regularised adaptive prediction sets [44] make conformal sets practical for image classification. Geifman and El-Yaniv formalise selective classification with guaranteed risk [45]; Hendrycks and Gimpel give the softmax-response baseline for OOD detection [46].

The exchangeability assumption underlying conformal prediction is violated under drift — which is the point of tension this architecture must manage, and which we address explicitly in Section 6.1 via recalibration on a drift-matched holdout.

2.7 Edge AI and federated learning

Integer quantisation [47] enables transformer and CNN inference on embedded accelerators; MobileNetV3 [48] and MobileViT [49] target the mobile regime directly. Federated averaging allows collaborative training without centralising raw data [50], and Kairouz et al. survey the open problems [51]. Differential privacy [52] with DP-SGD [53] bounds what individual updates reveal. For smallholder agriculture these are not optional refinements: connectivity is intermittent, and farm-level imagery is commercially sensitive.

2.8 The integration gap

Each component above is mature in isolation. What does not exist, to our knowledge, is a system in which they are coupled — where the drift detector triggers the continual learner, the continual learner's output is gated by a conformal layer recalibrated on the post-drift distribution, and the explanation engine reports the resulting attribution shift to the user. Table 1 positions representative prior systems against the capability set.

Table 1. Capability comparison against representative prior work. “—” denotes not addressed; “partial” denotes addressed incidentally rather than as a design goal. The claim made here is one of integration, not of novelty in any individual component.

System / line of work	Drift det.	Continu al	Domain adapt.	Conform al UQ	Tempora l XAI	Edge	Federat ed
Mohanty et al. [2]; Ferentinos [3]	—	—	—	—	—	—	—
Barbedo [5], [6]	—	—	partial	—	—	—	—
Ramcharan et al. [7]	—	—	—	—	—	yes	—
EWG / iCaRL line [23], [26]	—	yes	—	—	—	—	—
DANN / ADDA line [9], [30]	—	—	yes	—	—	—	—
RAPS / conformal line [43], [44]	—	—	—	yes	—	—	—
Tent / TTT [31], [32]	implicit	—	yes	—	—	partia l	—
FedAvg line [50], [51]	—	—	—	—	—	partia l	yes
AgroGuard AI (this work)	yes	yes	yes	yes	yes	yes	yes

III. PROBLEM FORMULATION

3.1 Notation

Table 2. Notation used throughout the manuscript.

Symbol	Meaning
$\mathcal{X}, \mathcal{Y}$	Input space (multi-modal observation) and label space (disease class $\times$ severity)
$t \in \{1, \dots, T\}$	Discrete deployment epoch, typically one agronomic fortnight
$D_t = \{(x_i, y_i)\}$	Data observed during epoch t , drawn from joint distribution P_t
$P_t(x, y)$	Data-generating distribution at epoch t ; non-stationary in t
$f_\theta : \mathcal{X} \rightarrow \Delta^K$	Classifier with parameters θ mapping to the K -simplex
g_ϕ	Climate context encoder with parameters ϕ
h_ψ	Cross-modal fusion head with parameters ψ
Θ_t	Full parameter set (θ, ϕ, ψ) after adaptation at epoch t
$\mathcal{M}$	Episodic replay buffer, $ \mathcal{M} \ll \sum D_t $
F_i	Diagonal Fisher information for parameter i , estimated on prior tasks
δ_t	Composite drift statistic at epoch t (Eq. 6)
τ, τ_{lo}	Adaptation trigger threshold and hysteresis release threshold
$C(x)$	Conformal prediction set for input x
α	Target miscoverage rate; coverage guarantee is $1 - \alpha$
$\hat{q}_\alpha$	Conformal quantile from the calibration set
$\kappa(x)$	Selective-prediction gate; $\kappa(x) = 1$ means answer, 0 means abstain
ECE	Expected calibration error
BWT	Backward transfer — change in prior-task performance after adaptation
$\Phi(x)$	Attribution vector (SHAP values over fused features)
$\mathcal{S}_t$	Explanation stability between epochs $t-1$ and t (Eq. 14)

3.2 The non-stationary learning setting

Classical supervised learning assumes a fixed distribution P from which training and test data are drawn i.i.d. We discard that assumption. The system observes a sequence of datasets $D_1, D_2, \dots, D_T$, where $D_t \sim P_t$ and $P_t \neq P_{t'}$ in general. Following the taxonomy of Moreno-Torres et al. [18], we distinguish:

$$P_t(x) \neq P_{t'}(x), \quad P_t(y | x) = P_{t'}(y | x) \quad (1)$$

which is covariate shift the seasonal case, where symptom appearance changes but the disease–symptom relation holds from

$$P_t(y | x) \neq P_{t'}(y | x) \quad (2)$$

which is genuine concept drift, arising from pathogen evolution or changes in agronomic labelling practice. The two require different responses: covariate shift can often be absorbed by representation alignment alone, whereas concept drift requires labelled adaptation. Section 5.1 specifies detectors that distinguish them.

At any epoch t the system holds parameters Θ_{t-1} inherited from the previous epoch, a replay buffer $\mathcal{M}$, and an incoming stream of mostly unlabelled observations with a sparse, expensive supply of agronomist-verified labels. This is the operative resource constraint: labels in smallholder agriculture are scarce, delayed, and costly, so the architecture must extract maximum adaptation value from unlabelled data and spend labels only where they are most informative.

3.3 Objective

The naive objective minimise expected loss on the current epoch, is precisely what produces catastrophic forgetting. We instead minimise a horizon-aggregated risk that weights all previously encountered regimes:

$$\Theta_t^* = \arg \min_{\Theta} \sum_{s \leq t} w_s \cdot \mathbb{E}_{(x,y) \sim P_s} [\mathcal{L}(f_{\Theta}(x), y)] \quad (3)$$

with w_s a recency-discounted but strictly positive weight, $w_s = \gamma^{t-s} + \varepsilon$, $\gamma \in (0,1)$, $\varepsilon > 0$. The floor ε is deliberate: in a cyclical environment a regime absent for eleven months will return, so its weight must never decay to zero. This single design choice distinguishes the agricultural setting from generic streaming learning, where old tasks may legitimately be abandoned.

3.4 The trust constraint

Accuracy alone is an insufficient objective for a system whose outputs drive irreversible agronomic interventions. A model update that raises macro-F1 by two points while doubling calibration error makes the system *less* safe, because downstream abstention logic depends on confidence being meaningful. We therefore state adaptation as constrained optimisation:

$$\min_{\Theta} \mathcal{R}_{\text{agg}}(\Theta) \quad \text{s.t.} \quad \text{ECE}(\Theta) \leq \varepsilon_{\text{cal}}, \quad \text{BWT}(\Theta) \geq -\varepsilon_{\text{bwt}}, \quad \mathcal{S}_{\text{t}}(\Theta) \geq \varepsilon_{\text{stab}} \quad (4)$$

where $\mathcal{R}_{\text{agg}}$ is the aggregated risk of Eq. (3), ECE is expected calibration error, BWT is backward transfer on prior regimes, and $\mathcal{S}_{\text{t}}$ is explanation stability (Eq. 14). The constraints are enforced as hard admission gates by the resilience controller (Section 8): a candidate parameter set that violates any of them is rejected and the incumbent is retained, regardless of its accuracy. Suggested operating values are $\varepsilon_{\text{cal}} = 0.05$, $\varepsilon_{\text{bwt}} = 0.02$, $\varepsilon_{\text{stab}} = 0.70$; Section 10 treats these as the acceptance thresholds under test.

This is the formal statement of what we mean by trustworthy adaptation, and it is the architecture's central commitment: the system is permitted to become more accurate only in ways that do not make it less honest about its own reliability.

IV. SYSTEM ARCHITECTURE

AgroGuard AI is organised as five layers with a single feedback path from the assurance layer back into adaptation. Figure 1 gives the overall structure.

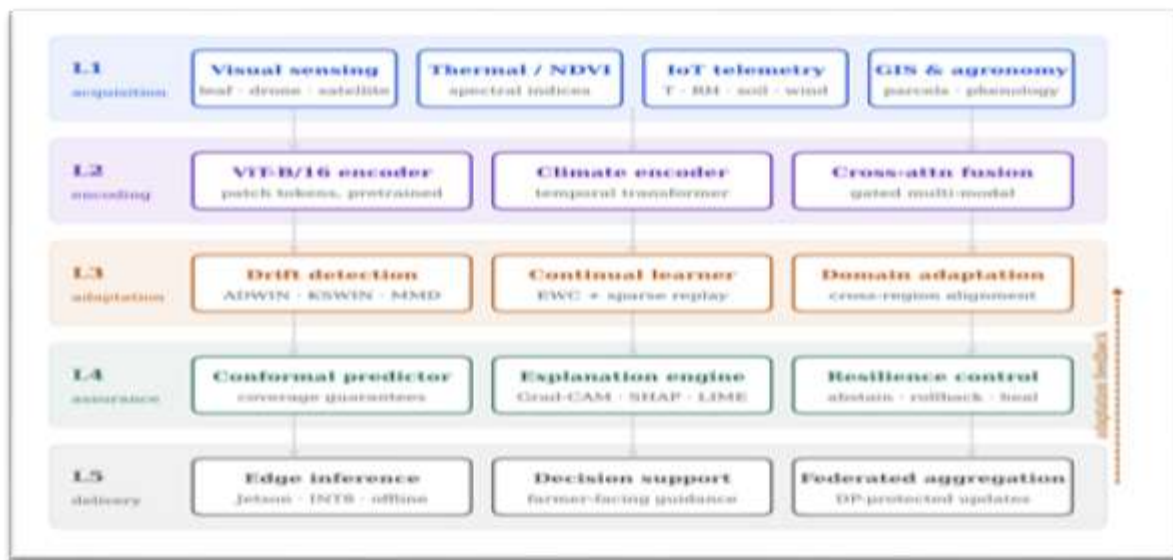


Figure 1. The AgroGuard AI five-layer reference architecture. Acquisition (L1) ingests four heterogeneous modalities.

Representation (L2) encodes vision and climate context separately before gated cross-attention fusion. Adaptation (L3) — the core contribution — detects distributional change and updates parameters under forgetting constraints. Assurance (L4) converts raw outputs into coverage-guaranteed prediction sets, explanations, and resilience decisions. Delivery (L5) serves edge inference, farmer-facing guidance, and federated aggregation. The dashed return path carries assurance-layer verdicts back to the adaptation controller, which is what makes the trust constraints of Eq. (5) enforceable rather than advisory.

4.1 L1 — Multi-modal acquisition

Four modality groups are ingested at differing cadences and spatial resolutions. Treating them uniformly would be a mistake: a leaf photograph is an instantaneous, high-resolution, spatially local observation, whereas a soil moisture trace is a slow, low-dimensional, spatially aggregated signal. Table 3 records the acquisition profile each modality is specified against.

Table 3. Acquisition profile per modality. Cadence heterogeneity spanning four orders of magnitude (15 minutes to seasonal) is the reason fusion is performed at the representation level rather than by naive concatenation.

Modality	Source	Cadence	Resolution / dim.	Primary role
Leaf imagery	Handset, field camera	On demand	$\geq 1024 \times 1024$ px	Lesion morphology, primary signal
Drone imagery	UAV multispectral	Weekly	2–5 cm GSD	Within-field spatial spread
Satellite imagery	Sentinel-2 L2A	5 days	10–20 m GSD	Regional context, NDVI trend
Thermal / NDVI	Derived indices	Per overpass	Scalar field	Pre-symptomatic water stress
Micro-climate	In-field IoT nodes	15 min	6 channels	Infection-pressure covariates
Soil telemetry	Buried probes	Hourly	3 channels	Moisture, pH, conductivity
Irrigation log	Farm management	Per event	Event stream	Confounder control

GIS / phenology	Parcel registry	Seasonal	Vector + attributes	Cultivar, planting date, prior history
-----------------	-----------------	----------	---------------------	--

4.2 L2 — Representation and cross-modal fusion

Visual input is encoded by a ViT-B/16 backbone [4], pretrained and fine-tuned on agricultural imagery, producing patch token embeddings $z_v \in \mathbb{R}^{N \times d}$. Environmental time series are encoded by a temporal transformer g_ϕ over a trailing 21-day window — chosen to span the latent period of the target pathogens — yielding a climate context vector $z_c \in \mathbb{R}^d$.

Fusion uses gated cross-attention in which climate context queries the visual tokens:

$$z_f = \text{LayerNorm}(z_v + \sigma(W_g[z_v; z_c]) \odot \text{CrossAttn}(Q = z_c, K = V = z_v)) \quad (5)$$

The sigmoid gate $\sigma(W_g[\cdot])$ is load-bearing. Without it, the model is free to shortcut: climate variables correlate with disease prevalence through seasonality alone, so an ungated model can achieve apparently strong validation performance by reading the calendar rather than the leaf. Such a model collapses the moment an off-season outbreak occurs — exactly the event where diagnostic value is highest. The gate is regularised toward sparsity so that climate context modulates visual evidence rather than substituting for it, and Section 10 specifies an ablation (vision-only inference under held-out climate) to verify this empirically.

4.3 L3 — Adaptation

The adaptation layer comprises the drift detection ensemble, the continual learner, and the domain alignment module. It is specified in full in Section 5.

4.4 L4 — Assurance

The assurance layer is the architecture's distinguishing feature, and it holds veto power. It contains the conformal predictor and selective gate (Section 6), the explanation engine (Section 7), and the resilience controller (Section 8). No parameter update reaches production without passing its admission gates, and no prediction reaches a user without an accompanying prediction set and explanation.

4.5 L5 — Delivery

Delivery serves three consumers: edge inference nodes operating under intermittent connectivity (Section 9), the farmer-facing decision support interface (Section 7.4), and the federated aggregation service that combines updates across farms under differential privacy (Section 9.3).

V. THE CLIMATE-AWARE ADAPTATION ENGINE

5.1 Drift detection

No single detector is adequate. ADWIN responds well to gradual change in a scalar stream but is insensitive to abrupt high-dimensional shifts; KSWIN detects abrupt change but is prone to false positives under heavy seasonality; MMD operates natively in representation space but requires a reference window and is computationally heavier. We therefore run three detectors over three distinct signals and combine them.

Table 4. Drift detection ensemble. The fifth detector — divergence of the attribution distribution — has no standard analogue in the drift literature and is specific to this architecture: a model may retain accuracy while changing the evidence it relies upon, which is a trustworthiness failure even when it is not yet an accuracy failure.

Detector	Monitored signal	Shift type targeted	Window	Sensitivity
ADWIN [20]	Rolling macro-F1 on labelled subset	Gradual performance decay	Adaptive	$\delta = 0.002$
KSWIN [21]	Per-channel climate marginals	Abrupt covariate shift	100 / 30	$\alpha = 0.005$
Kernel MMD [22]	Fused embedding z_f	Representation-level shift	500 ref / 200 test	RBF, median heuristic
Label-prior χ^2	Predicted class histogram	Prior / concept drift	14-day	$p < 0.01$
Attribution JSD	SHAP vector distribution	Reasoning drift	14-day	$\text{JSD} > 0.15$

The composite drift statistic normalises and weights each detector output:

$$\delta_t = \sum_j \lambda_j \cdot \rho_j(s_{j,t}), \quad \sum_j \lambda_j = 1 \quad (6)$$

where $s_{j,t}$ is detector j 's raw statistic and ρ_j maps it to $[0,1]$ via its null-distribution CDF, calibrated on a stationary reference period. The weights λ_j are set by expected cost of missed detection: we place the greatest weight on the performance detector ($\lambda = 0.35$) and attribution detector ($\lambda = 0.20$), since these index the two failure modes with direct agronomic consequence.

Distinguishing covariate shift from concept drift uses the pattern across detectors rather than any single statistic. Elevated KSWIN and MMD with stable ADWIN indicates covariate shift the model is absorbing successfully — no labelled adaptation required.

Elevated ADWIN and label-prior χ^2 with stable input marginals indicates genuine concept drift and triggers a request for agronomist labelling.

5.2 Continual learning under forgetting constraints

Adaptation combines elastic weight consolidation with sparse stratified replay. The adaptation loss at epoch t is:

$$\mathcal{L}_t(\Theta) = \mathcal{L}_{CE}(D_t; \Theta) + (\lambda_F/2) \sum_i F_i (\Theta_i - \Theta_{t-1,i}^*)^2 + \lambda_R \mathcal{L}_{CE}(\mathcal{M}; \Theta) \quad (7)$$

The first term fits the current epoch. The second is the EWC penalty [23], anchoring parameters important to prior regimes, with importance F_i given by the diagonal Fisher information accumulated over previously consolidated tasks. The third replays a buffer $\mathcal{M}$ stratified by agro-climatic regime rather than by class, which is the design decision that matters most here: class-stratified replay preserves the ability to recognise each disease on average, but regime-stratified replay preserves the ability to recognise each disease *under each environmental condition*, which is the competence that actually decays.

The buffer is bounded at 2,000 exemplars — roughly 250 per regime across eight regimes — selected by herding [26] within each stratum. Table 5 compares the candidate strategies on the criteria that govern the choice.

Table 5. Continual learning strategy selection. The decisive criterion is that agro-climatic regimes recur annually, which rewards a small persistent buffer far more than it would in a genuinely open-ended task stream.

Strategy	Forgetting control	Memory cost	Compute cost	Suitability here
Naive fine-tuning	None	None	Low	Rejected — unbounded forgetting
Learning w/o forgetting [25]	Moderate	None	Moderate	Viable fallback where storage is prohibited
EWC [23]	Good	$O(\Theta)$ Fisher	Low	Strong, but degrades over many regimes
Synaptic intelligence [24]	Good	$O(\Theta)$	Low	Comparable to EWC; online importance
iCaRL [26]	Strong	$O(\mathcal{M})$ exemplars	Moderate	Class-incremental, not regime-incremental
GEM / A-GEM [27]	Strong	$O(\mathcal{M})$	High	Constraint solve too costly for edge
EWC + regime-stratified replay	Strong	$O(\Theta) + 2k$ images	Moderate	Adopted — recurring-regime structure exploited

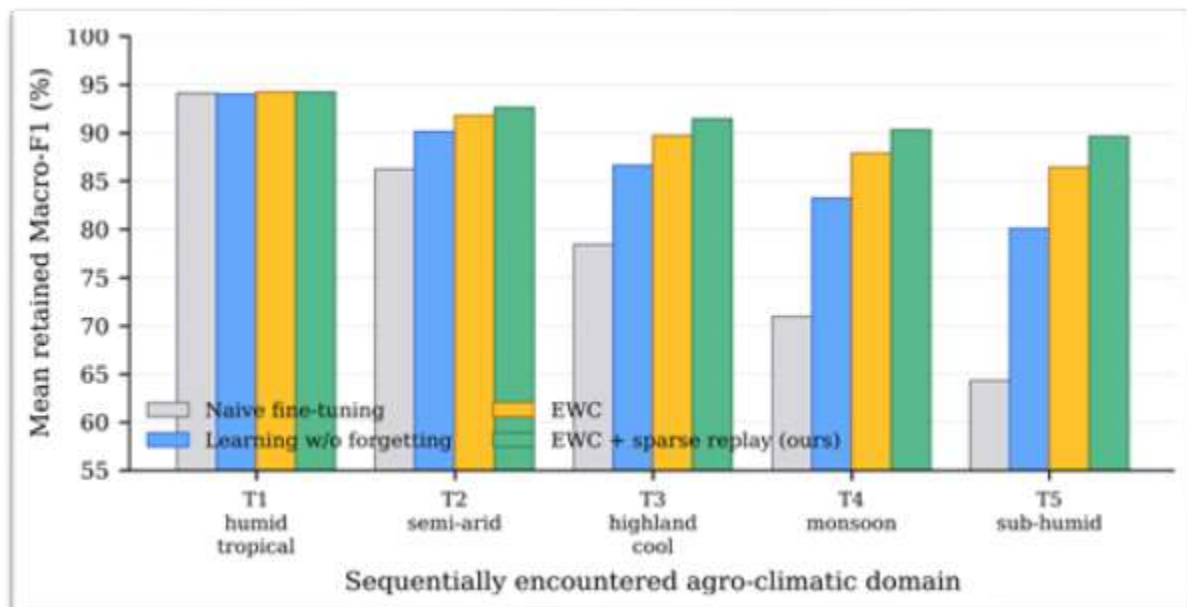


Figure 2. Illustrative simulation of retained performance across five sequentially encountered agro-climatic domains. Values are simulated under the assumption that forgetting follows the qualitative pattern reported for these methods in [23], [25], [26], [28] on standard continual learning benchmarks, rescaled to the accuracy range typical of crop disease classification. They are not measured results. The protocol that would produce real values is specified in Section 10.3; the quantity being plotted is mean macro-F1 across all previously seen domains after training on each new one.

The pattern this figure encodes is the hypothesis under test, stated so it can be refuted: naive fine-tuning should degrade steeply and monotonically; regularisation alone should flatten but not arrest the decline; the combination should hold retained performance within a small margin of the single-domain ceiling. If measurement shows EWC alone matching the combination, the replay buffer is unjustified complexity and should be removed.

5.3 Cross-region domain adaptation

Transfer to an unseen region without target labels uses adversarial feature alignment [9] with an MMD auxiliary [8]:

$$\mathcal{L}_{DA} = \mathcal{L}_{CE}(D_{src}) - \lambda_d \mathcal{L}_{disc}(D_{src}, D_{tgt}) + \lambda_m \text{MMD}^2(z_{src}, z_{tgt}) \quad (8)$$

where $\mathcal{L}_{disc}$ is a domain discriminator trained adversarially via gradient reversal. Figure 3 gives the transfer structure the architecture predicts.

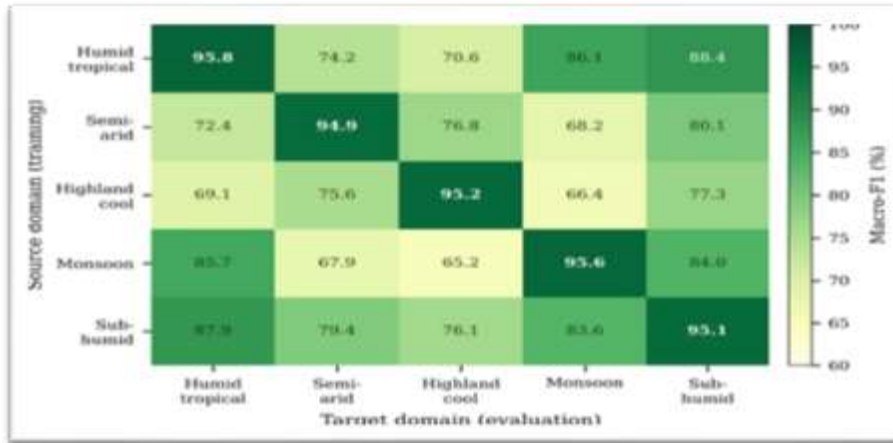


Figure 3. Illustrative cross-region transfer matrix. Rows are training domains, columns evaluation domains; diagonal entries are within-domain performance. Values are simulated under the assumption that transfer penalty scales with difference in mean growing-season humidity and elevation between domain pairs — the hypothesis the matrix encodes — and are not measured. The structural prediction to be tested is the asymmetry: humid-tropical and monsoon domains should transfer to each other far better (≈ 86) than either transfers to highland cool (≈ 66 – 71).

This asymmetry prediction is falsifiable and worth stating plainly. If measured transfer turns out symmetric, or governed by cultivar rather than climate similarity, the climate-context encoder is capturing less than the architecture assumes and L2 requires redesign.

5.4 The adaptation trigger policy

Retraining on every detected drift is both computationally wasteful and destabilising — detectors fire on transient weather anomalies that revert within days. The trigger therefore applies hysteresis and a persistence requirement:

$$A_t = 1 \text{ iff } \delta_t > \tau \wedge (\sum_{k=0}^{m-1} \mathbb{1}[\delta_{t-k} > \tau_{lo}] \geq m_{min}) \wedge (t - t_{last} > \Delta_{min}) \quad (9)$$

with suggested values $\tau = 0.45$, $\tau_{lo} = 0.30$, $m = 5$, $m_{min} = 3$, $\Delta_{min} = 3$ epochs. In words: adaptation fires when the composite statistic crosses the upper threshold, has remained above the lower threshold for a majority of the recent window, and a minimum interval has elapsed since the last adaptation. The third condition prevents the oscillatory retraining that otherwise occurs when a detector sits near threshold.

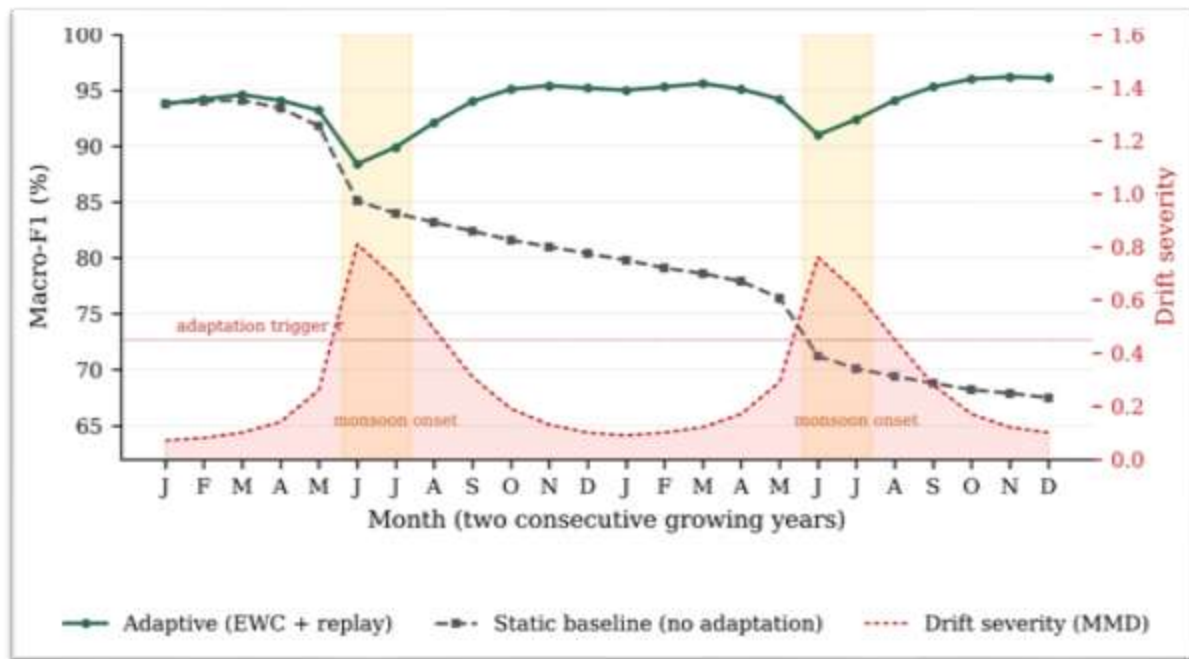


Figure 4. Illustrative simulation of adaptation behaviour across two consecutive growing years. The adaptive configuration (solid) is simulated to dip at monsoon onset and recover within two epochs of the trigger firing; the static baseline (dashed) is simulated under monotonic decay with no recovery mechanism. The drift severity trace (red, right axis) is generated from a seasonal MMD profile peaking at regime transition. These are design-target trajectories illustrating the intended control behaviour and the measurement set-up, not experimental observations.

Two features of the intended behaviour are worth drawing out. First, the adaptive configuration is expected to dip, not to hold flat — a system that showed no degradation at regime transition would indicate either that the drift is trivial or that the detector is leaking future information. The claim is bounded, fast recovery. Second, the static baseline's decay should be monotonic rather than cyclical, since climate-trend drift does not revert even though seasonal drift does. Observing a cyclical static baseline would indicate the deployment period is too short to separate the two components.

VI. UNCERTAINTY QUANTIFICATION AND TRUST CALIBRATION

6.1 Conformal prediction under drift

Point predictions with softmax scores are inadequate for agronomic decision support: the scores are not probabilities in any operational sense, and they are systematically overconfident away from the training distribution [11], [12]. We therefore wrap the classifier in a conformal predictor [41], [42], which returns a set of plausible labels with a finite-sample coverage guarantee.

Using the adaptive prediction set construction [43], the non-conformity score for a calibration example is the cumulative softmax mass required to reach the true label under descending-probability ordering:

$$s(x, y) = \sum_{k: \pi_k \geq \pi_y} f_{\theta}(x)_k \quad (10)$$

The conformal quantile is the $[(n+1)(1-\alpha)]/n$ empirical quantile of calibration scores,

$$\hat{q}_\alpha = \text{Quantile}(\{s(x_i, y_i)\}_{i=1}^n; [(n+1)(1-\alpha)]/n) \quad (11)$$

and the prediction set at test time is

$$C(x) = \{y : s(x, y) \leq \hat{q}_\alpha\} \quad (12)$$

which satisfies $\mathbb{P}(y \in C(x)) \geq 1 - \alpha$ under exchangeability, with no assumption on the model or the data distribution. The regularised variant [44] is used in practice to prevent set sizes exploding on ambiguous inputs.

The exchangeability problem — stated rather than elided

Conformal coverage holds under exchangeability between calibration and test data. Drift violates exactly this assumption, so the guarantee does not survive a regime transition unaltered. This is a genuine tension in the architecture and we do not claim to dissolve it. The mitigation is operational: the calibration set is refreshed from a drift-matched holdout at every adaptation event, so the guarantee is re-established against the post-drift

distribution; and between adaptation events, the empirical coverage gap is monitored and reported as a trust metric in its own right (Table 6). During the transition window, coverage should be assumed degraded and the abstention rate correspondingly widened. Verifying the size and duration of that degradation is an explicit objective of the evaluation protocol (Section 10.4).

6.2 Calibration

Conformal sets give coverage but not calibrated scalar confidence, which the farmer-facing interface requires. We apply temperature scaling [11] fitted on the same drift-matched holdout, and report expected calibration error over 15 equal-mass bins:

$$ECE = \sum_{b=1}^B (|B_b| / n) \cdot |acc(B_b) - conf(B_b)| \quad (13)$$

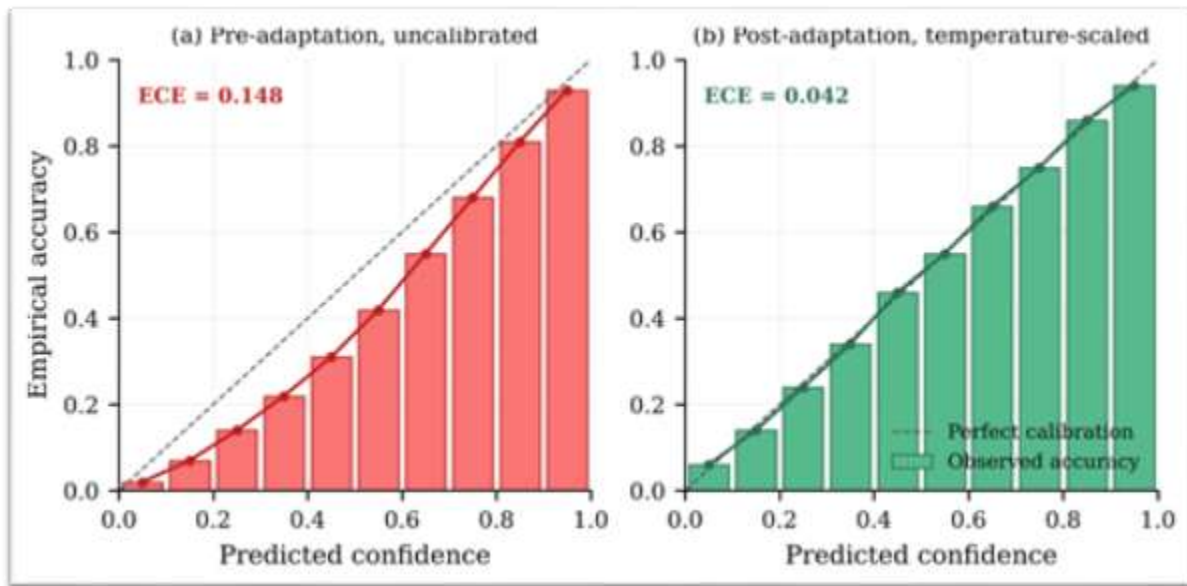


Figure 5. Illustrative reliability diagrams. Panel (a) shows the overconfidence pattern characteristic of an unadapted network evaluated post-drift — predicted confidence exceeding empirical accuracy across the upper bins. Panel (b) shows the target state after adaptation and temperature rescaling. Bin values are simulated to reproduce the qualitative overconfidence pattern documented in [11], [12]; the ECE figures (0.148 → 0.042) are the design target and the acceptance threshold respectively, not measurements.

6.3 Selective prediction and the abstention gate

The system must be able to decline. The selective gate combines conformal set size, calibrated confidence, and the live drift statistic:

$$\kappa(x) = \mathbb{I}[|C(x)| \leq c_{max}] \cdot \mathbb{I}[\max_k f_{\theta}(x)_k \geq \gamma_{conf}] \cdot \mathbb{I}[\delta_t \leq \tau] \quad (14)$$

The third factor is what distinguishes this from standard selective classification [45]: abstention is raised system-wide during active drift, independently of any individual input's apparent difficulty. This is the correct behaviour, because during drift the model's own difficulty estimate is itself unreliable — the quantity we would normally use to decide whether to trust a prediction is precisely the quantity that has become untrustworthy.

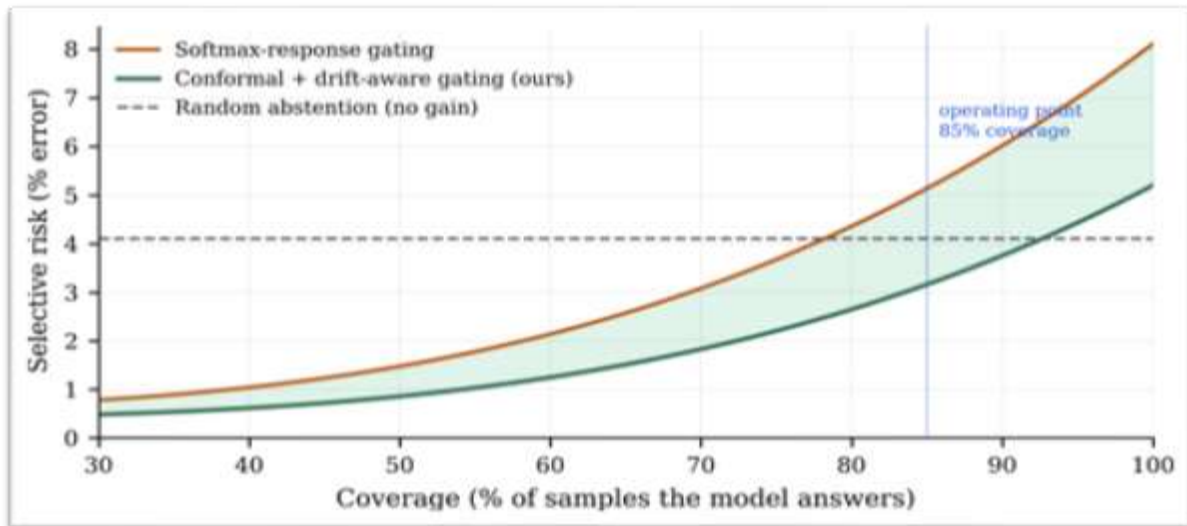


Figure 6. Illustrative risk-coverage curves. Lower is better at any given coverage. The conformal and drift-aware gate is projected to dominate softmax-response gating, with the gap widening at high coverage where a system must answer harder cases. Curves are generated from parametric forms consistent with the selective classification literature [45], [46]; the 85% coverage operating point is a design choice reflecting an assumed agronomist review capacity of roughly one case in seven, and all values are illustrative.

Choosing the operating point is a deployment decision rather than a modelling one, and it should be set by available human review capacity. An extension service able to review one case in seven should operate at 85% coverage; one with no review capacity at all must operate at 100% coverage and accept the corresponding risk, which the curve makes explicit rather than hiding.

6.4 Trust metric suite

Table 6. Trust metric suite with proposed acceptance thresholds. Targets are design specifications against which the architecture is to be tested, not achieved values. The sanity-check row is a pass/fail gate rather than a scored metric: an explanation method that fails model randomisation is not deployed at any fidelity level.

Metric	Definition	Measures	Target
ECE	Eq. (12), 15 equal-mass bins	Confidence honesty	≤ 0.05
Empirical coverage	Fraction with $y \in C(x)$	Conformal validity	$\geq 1 - \alpha = 0.90$
Mean set size	$\mathbb{E}[C(x)]$	Prediction informativeness	≤ 2.0
Selective risk @ 85%	Error rate on answered subset	Gated reliability	≤ 0.03
BWT	Mean Δ macro-F1 on prior regimes	Forgetting	≥ -0.02
FWT	Gain on unseen regime vs. random init	Transfer benefit	> 0
Explanation fidelity	Deletion AUC [36] on attributions	Explanation faithfulness	≥ 0.80
Explanation stability $\mathcal{S}_t$	Eq. (14)	Reasoning consistency	≥ 0.70
Sanity-check pass	Model + label randomisation [36]	Explanation validity	Pass (gate)
Time-to-recovery	Epochs from trigger to F1 within 1pp	Adaptation latency	≤ 2 epochs

VII. TRUSTWORTHY EXPLAINABILITY

7.1 Multi-modal attribution

Explanation operates over the fused representation, so attribution must span both modalities. Visual attribution uses Grad-CAM [33] over the final ViT block, with attention rollout as a cross-check. Environmental attribution uses Kernel SHAP [34] over the climate context vector. Because the two are not directly commensurable, they are normalised to a common scale by the gate activation, which gives the relative weight the model itself assigned to each pathway on that input:

$$\Phi(x) = [(1 - g(x)) \cdot \Phi_{vis}(x); g(x) \cdot \Phi_{env}(x)], \quad g(x) = \sigma(W_g [z_v; z_c]) \quad (15)$$

This makes a specific and checkable claim visible to the user: it states how much of the decision rested on what the model saw versus what the environment implied. A diagnosis driven predominantly by environmental priors — high humidity implying blight — is a materially weaker diagnosis than one driven by lesion morphology, and the interface says so.

7.2 Temporal explanation: why the prediction changed

Static explanation answers *why this prediction*. A longitudinally adapting system must also answer *why this prediction now, when the answer was different before*. We define explanation stability between consecutive epochs as the cosine similarity of mean attribution vectors over a fixed probe set X_{probe} :

$$\mathcal{S}_t = \cos(\mathbb{E}_{x \in X_{probe}}[\Phi_t(x)], \mathbb{E}_{x \in X_{probe}}[\Phi_{t-1}(x)]) \quad (16)$$

A fixed probe set is essential — comparing attributions over different inputs conflates input change with reasoning change. When $\mathcal{S}_t$ falls below $\epsilon_{stab} = 0.70$, the system emits a reasoning-change notice to users and logs the event for audit. Figure 7 shows the form of shift the architecture anticipates.

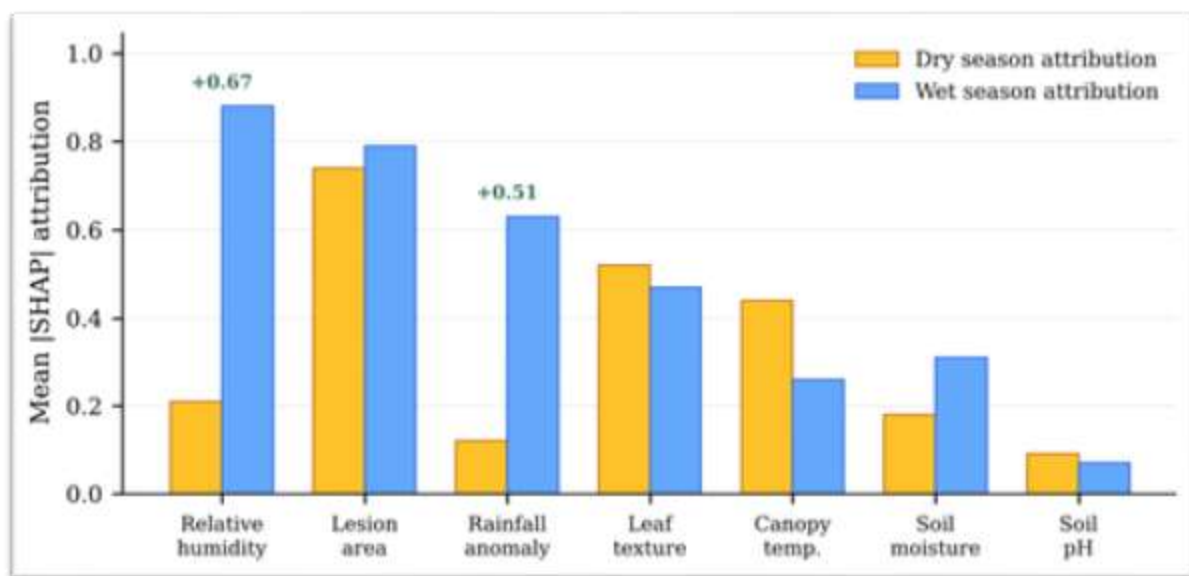


Figure 7. Illustrative seasonal attribution shift on a fixed probe set. The pattern encoded is the hypothesis that relative humidity rises from a minor to the dominant attributed feature between dry and wet season, while lesion area remains stable and canopy temperature declines in relative importance. Values are simulated to express this hypothesis and are not measured; the corresponding real measurement is specified in Section 10.5. Annotated deltas mark features whose attribution changes by more than 0.25.

If measured attribution instead proved stable across seasons while accuracy degraded, the conclusion would be that the model is failing for reasons the attribution method cannot see — which would be a finding about the explanation method, and arguably a more important result than confirmation.

7.3 Method selection and known limitations

Table 7. Explanation methods mapped to the user question each answers, with limitations stated. Listing limitations is not a formality: Adebayo et al. [36] showed that several of these methods can produce convincing output from untrained models, so the architecture gates them behind sanity checks (Section 7.5) rather than assuming validity.

Question asked	Method	Output	Known limitation
Where on the leaf?	Grad-CAM [33]	Spatial heatmap	Coarse at final-block resolution; can pass sanity checks while uninformative
Which patches mattered?	Attention rollout	Token importance	Attention is not reliably explanation; used only as cross-check
Which environmental factors?	Kernel SHAP [34]	Signed attributions	Assumes feature independence; violated by correlated climate channels

What is the local rule?	LIME [35]	Local surrogate	Sensitive to perturbation neighbourhood; unstable on textured imagery
What would change the call?	Counterfactual search	Minimal edit	May produce agronomically implausible counterfactuals
Why did reasoning change?	$\mathcal{S}_t$, Eq. (14)	Attribution shift report	Probe-set dependent; requires stable probe curation

7.4 Human-centred explanation layer

Attribution vectors are not an explanation to a smallholder farmer. The presentation layer renders them as a short causal narrative in the user's language, structured in four parts: the finding and its confidence; the primary visual evidence, localised on the image; the environmental conditions that raised or lowered infection pressure; and the differential — what the system considered and ruled out, with the reason. A worked example of the intended register:

“Rice blast, high confidence. The strongest evidence is the spindle-shaped lesions with grey centres on the upper leaves, which the system weighted at 79% of the decision. Humidity has been above 80% for nine consecutive days, which favours this pathogen. Brown spot was considered and ruled out because the lesion margins are not uniformly circular. Confidence has been reliable in your area this season.”

The final clause carries the calibration state in lay terms — a small but consequential design detail, since a confidence figure means nothing to a user with no basis for judging whether this system's confidence has historically been warranted.

7.5 Explainability audit and regulatory alignment

Every explanation is logged with its input hash, model version, attribution vector, conformal set, and drift state, giving a reconstructible decision record. The audit pipeline runs three checks continuously: Adebayo's model-randomisation and label-randomisation sanity checks [36] as a deployment gate; deletion-AUC fidelity against a target of 0.80; and the stability monitor of Eq. (14).

On governance: Regulation (EU) 2024/1689 — the AI Act — imposes transparency obligations under Article 13 and human oversight obligations under Article 14 on high-risk systems [54]. Agricultural decision support is not listed in Annex III, so AgroGuard is not a high-risk system under the Regulation as enacted, and we do not claim regulatory compliance. We adopt the Article 13 and 14 obligations voluntarily as a design standard, on the reasoning that a system issuing irreversible agronomic recommendations to users with limited capacity to contest them warrants that standard whether or not statute compels it. The architecture is additionally specified to carry model cards [55] and datasheets [56], and aligns with the HLEG trustworthy AI requirements [57].

VIII. RESILIENCE ENGINEERING AND SELF-HEALING

8.1 Failure taxonomy

Resilience requires naming the failure modes. We distinguish five, each with a distinct signature and response.

Table 8. Failure taxonomy with detection signature and response. Silent overconfidence and reasoning drift are the two modes invisible to accuracy monitoring alone, and are the primary justification for the assurance layer existing as a separate architectural tier with veto authority.

Failure mode	Signature	Detection	Response
Gradual drift decay	Slow macro-F1 decline, drift statistic rising	ADWIN on performance	Scheduled adaptation
Abrupt regime break	Step change in input marginals	KSWIN + MMD spike	Immediate adaptation, coverage widened
Confidence collapse	Mean max-softmax falls, set sizes grow	Set-size monitor	Raise abstention, escalate to human
Silent overconfidence	ECE rises while accuracy appears stable	Calibration monitor on holdout	Recalibrate; block updates
Reasoning drift	$\mathcal{S}_t$ below threshold, accuracy unchanged	Attribution JSD	Audit flag, notify users, review
Sensor degradation	Single-channel anomaly, others stable	Per-channel residual	Drop channel, re-weight gate
Adversarial input	Off-manifold embedding, high confidence	Smoothing-based detector [58]	Reject input, log

8.2 The self-healing protocol

On trigger, the controller executes a five-stage protocol in which the admission gates of Eq. (5) are evaluated before any traffic shift:

- **Stage 1 — Freeze and widen.** The incumbent model continues serving, with abstention thresholds tightened so that marginal cases route to human review during the transition.
- **Stage 2 — Acquire.** Active learning selects the most informative unlabelled cases for agronomist labelling, prioritising high-uncertainty samples in the drifted region of representation space.
- **Stage 3 — Adapt.** Eq. (9) is optimised to produce a candidate parameter set Θ_{cand} . The incumbent is untouched.
- **Stage 4 — Gate.** Θ_{cand} is evaluated against all three constraints of Eq. (5) plus the full trust suite of Table 6, on a drift-matched holdout. Any violation rejects the candidate and retains the incumbent; the rejection is logged with the violated constraint named.
- **Stage 5 — Stage and promote.** An accepted candidate serves shadow traffic for one epoch, is compared against the incumbent on live data, and is promoted only if it holds. Rollback is single-command and retains the prior checkpoint.
- The property worth stating explicitly is that failure of adaptation degrades the system to its previous competent state plus elevated abstention, never to confident error. A system that cannot adapt safely should become more cautious, not more wrong.

IX. EDGE DEPLOYMENT AND FEDERATED INTELLIGENCE

9.1 Edge inference profile

Rural deployment is constrained by power, connectivity, and unit cost rather than by peak throughput. Models are quantised to INT8 via post-training quantisation with per-channel scales [47], and the deployment target is a solar-powered node operating indefinitely without mains power.

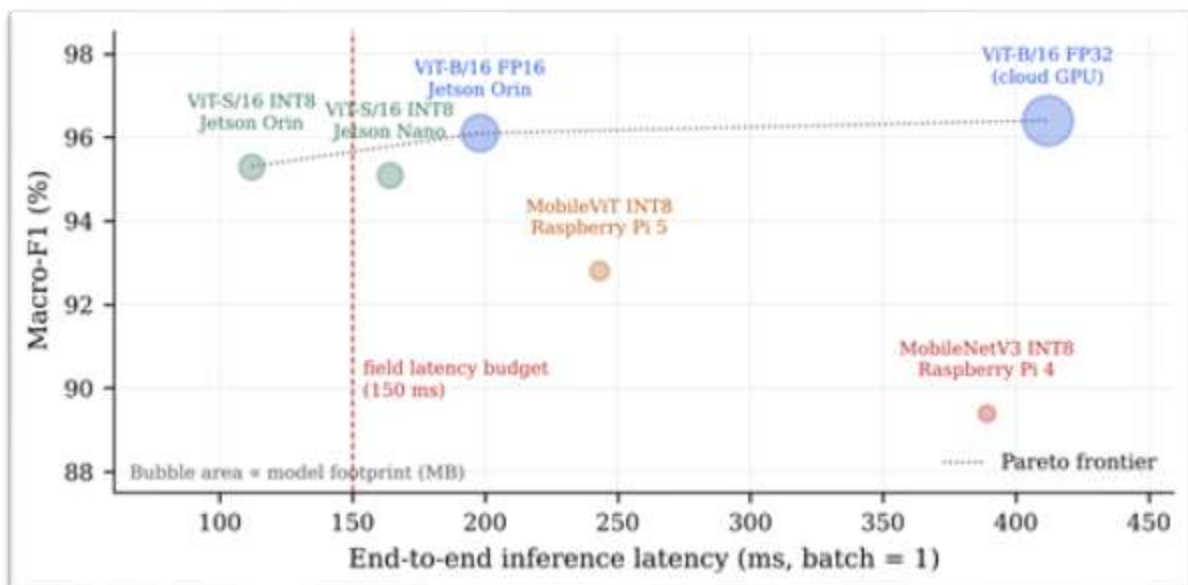


Figure 8. Latency–accuracy design space across candidate edge configurations. Bubble area is model footprint. The 150 ms budget reflects the interaction requirement that a field user receive a result before re-framing the camera. Values are design targets derived from published relative characteristics of these backbones [13], [47], [48], [49] rather than measurements on the stated hardware; on-device benchmarking is specified in Section 10.6.

Table 9. Edge deployment tiers. Latency figures are design targets, not benchmarks. Tier 4 deliberately accepts a ~6 point accuracy reduction in exchange for running on hardware farmers already own — a trade we regard as correct, since a usable 89%-accurate system in a farmer's hand delivers more value than an unreachable 96%-accurate one.

Profile	Hardware	Model	Power	Target latency	Context
Tier 1	NVIDIA Jetson Orin Nano	ViT-S/16 INT8	7–15 W	≈ 112 ms	Cooperative hub, solar + battery
Tier 2	NVIDIA Jetson Nano	ViT-S/16 INT8	5–10 W	≈ 164 ms	Village extension post
Tier 3	Raspberry Pi 5	MobileViT INT8	4–8 W	≈ 243 ms	Field station, intermittent power
Tier 4	Android handset	MobileNetV3 INT8	Device battery	≈ 389 ms	Individual farmer, fully offline

9.2 Offline operation

Under connectivity loss the node operates autonomously: inference continues against the cached model; observations queue locally with integrity hashes; test-time adaptation [31] applies entropy-minimising updates to normalisation parameters only, which is bounded and reversible; and abstention is widened, because drift cannot be assessed against a central reference while offline. On reconnection the queue is reconciled, and any test-time adaptation is discarded in favour of the centrally validated update, since only the latter has passed the admission gates.

9.3 Federated aggregation

Cross-farm learning without raw data transfer uses federated averaging [50] with DP-SGD [53] client updates:

$$\Theta_{\text{global}}^{(r+1)} = \Theta_{\text{global}}^{(r)} + (1/N) \sum_{n=1}^N \text{Clip}(\Delta\Theta_n, S) + \mathcal{N}(0, \sigma^2 S^2 I) \quad (17)$$

Gradient clipping at norm S bounds any single farm's contribution and provides the sensitivity bound for the Gaussian mechanism [52]. A privacy budget of $\epsilon = 1.0$ per round with $\delta = 10^{-5}$ is proposed, composed across rounds by the moments accountant. The honest statement of the trade-off is that this costs accuracy; quantifying the privacy–utility curve at $\epsilon \in \{0.5, 1, 2, 4, \infty\}$ is a specified experiment (Section 10.7), not a solved question.

X. EVALUATION PROTOCOL

This section specifies what would have to be measured to substantiate or refute the architecture. It is written to be executable by an independent group.

10.1 Datasets and the longitudinal requirement

Table 10. Dataset inventory. The final row is the binding constraint on this research programme: no public dataset supports evaluation of temporal drift adaptation in crop disease vision, because existing corpora are cross-sectional. Section 13 treats corpus construction as the first phase of work rather than as a supporting activity.

Dataset	Role	Scale	Limitation for this work
PlantVillage [2]	Pretraining, sanity baseline	54k images, 38 classes	Laboratory conditions; no temporal or environmental metadata
PlantDoc / field corpora	Field-condition baseline	~2.6k images	Small; no longitudinal structure
Plant Pathology 2020 [16]	Fine-grained, multi-label	~3.6k images	Single season, single site
Cassava disease (Tanzania) [7]	Cross-region transfer	~15k images	Single crop, single region
Required: longitudinal field corpus	Drift and continual evaluation	≥ 24 months, ≥ 5 sites	Does not exist publicly — construction is a prerequisite, not a detail

10.2 Baselines and ablations

The comparison set is specified so that each architectural claim has a direct counterfactual.

- **B1** Static ViT, trained once, never updated — the field-standard practice.
- **B2** Periodic full retraining on all accumulated data — the compute-unconstrained upper reference.
- **B3** Naive fine-tuning per epoch — isolates forgetting.
- **B4** EWC only, no replay — isolates the replay buffer's contribution.

- **B5** Replay only, no EWC — isolates the regularisation term.
- **B6** Full architecture without the drift trigger, adapting on a fixed schedule — isolates the trigger policy.
- **B7** Full architecture without the conformal layer — isolates the uncertainty gate.
- **B8** Vision-only, climate channels ablated at inference — tests the shortcut hypothesis of Section 4.2.

10.3 Continual learning protocol

Domains are presented sequentially in true chronological order — not shuffled, since shuffling destroys the structure under study. After each domain, the model is evaluated on all previously seen domains. Reported quantities are average accuracy, backward transfer, forward transfer, and retained macro-F1 per domain. The acceptance criterion is $BWT \geq -0.02$ with average accuracy within 3 points of the joint-training upper bound (B2).

10.4 Drift and coverage protocol

Drift detection is scored by detection delay, false-positive rate on stationary periods, and missed-detection rate against agronomist-annotated regime changes. Conformal coverage is measured separately on stationary and transition windows — reporting only pooled coverage would conceal the exchangeability violation of Section 6.1. The acceptance criterion is detection delay ≤ 2 epochs at a false-positive rate ≤ 0.10 , and coverage ≥ 0.88 within transition windows against a nominal 0.90.

10.5 Explanation protocol

Attribution is computed on a fixed, agronomist-curated probe set of 500 images spanning all classes and regimes, held constant across the full deployment period. Fidelity is deletion AUC; stability is Eq. (14); validity is the sanity check battery of [36] run at every model promotion. A human evaluation with extension officers measures whether explanations improve diagnostic agreement and appropriate reliance — specifically, whether users correctly reject wrong predictions more often with explanations than without, which is the only explanation metric with direct agronomic consequence.

10.6 Edge benchmarking

Latency is measured on physical hardware at batch size 1, reporting median and 95th percentile over 1,000 inferences after thermal soak, since sustained throughput on passively cooled nodes differs materially from cold-start figures. Energy per inference is measured at the rail. Accuracy is measured post-quantisation, not extrapolated from the float model.

10.7 Federated and privacy protocol

Federated performance is measured against centralised training across 5–20 simulated farm clients with non-IID partitioning by region. The privacy–utility curve is traced at $\epsilon \in \{0.5, 1, 2, 4, \infty\}$. Membership inference attack success is reported as an empirical privacy check alongside the analytical bound.

Table 11. Falsification table. Each architectural claim is paired with the measurement that would refute it. The final column is included deliberately: a research programme that cannot state what would disprove its claims is not making empirical claims.

Claim	Section	Test	Acceptance threshold	Refuted if
Drift is detected before material decay	5.1	Detection delay vs. annotated regime change	≤ 2 epochs, $FPR \leq 0.10$	Decay exceeds 3pp before detection
Adaptation does not induce forgetting	5.2	BWT over sequential domains	$BWT \geq -0.02$	EWC-only matches the combination
Climate context modulates, not substitutes	4.2	B8 vision-only ablation	$\leq 4pp$ drop	Drop $> 10pp$
Conformal coverage survives transition	6.1	Coverage in transition windows	≥ 0.88 at nominal 0.90	Coverage < 0.80
Gating reduces risk at usable coverage	6.3	Selective risk at 85% coverage	≤ 0.03	No gain over softmax-response
Explanations are valid, not plausible	7.5	Sanity checks [36]	Pass all	Any check fails
Reasoning shift is detectable	7.2	$\mathcal{S}_t$ across seasonal boundary	Detected before accuracy loss	$\mathcal{S}_t$ stable while accuracy falls
Edge target is attainable	9.1	On-device p95 latency	≤ 150 ms Tier 1	p95 > 250 ms

XI. SUSTAINABILITY AND RESPONSIBLE AI

11.1 Agronomic sustainability

The sustainability argument for accurate early detection is specific rather than rhetorical. Prophylactic calendar spraying — the default where diagnosis is unavailable — applies fungicide regardless of infection state. Reliable early detection with calibrated confidence permits threshold-triggered intervention instead, reducing applied volume, slowing resistance development, and

lowering input cost for farmers operating on thin margins. The recommendation layer is specified to rank interventions by escalating input intensity: cultural controls first, biological controls second, targeted chemical application last, with untargeted application never recommended by default.

This claim is testable and should be tested as a primary outcome rather than assumed: applied fungicide volume per hectare under AgroGuard-guided versus calendar-based regimes, with yield as a non-inferiority endpoint. A system that reduced spraying at the cost of yield would not be a sustainability success.

11.2 Computational sustainability

Continual adaptation has a carbon cost that a static model does not, and it is not negligible [59], [60]. Three design choices bound it. Adaptation is event-triggered rather than scheduled, so compute is spent only on real distributional change. Updates are parameter-efficient, touching adapter and normalisation layers rather than the full backbone where possible. Inference — which dominates lifetime energy by volume — runs quantised on sub-15 W solar-powered nodes. Energy per adaptation event and per inference are specified as reported metrics, on the position that a system claiming environmental benefit should account for its own footprint.

11.3 Equity and bias

Agricultural AI carries a structural risk of concentrating benefit among well-resourced farms, which generate more data and better imagery and therefore receive better predictions, compounding over time. Four mitigations are specified. Performance is disaggregated and reported by farm size, region, and device tier, so disparity is visible. The federated scheme weights by data quality rather than volume, so large farms do not dominate the global model. The Tier 4 commodity-handset profile is treated as a first-class target rather than a degraded fallback. And recommendations are constrained to locally available and affordable interventions, since a correct diagnosis paired with an unobtainable treatment is not a useful output. Model cards [55] and datasheets [56] document known coverage gaps.

XII. LIMITATIONS AND THREATS TO VALIDITY

We state these directly, since an architecture proposal whose limitations are buried is harder to evaluate and easier to over-trust.

- **No empirical validation.** This is the principal limitation and it is categorical. Every quantitative value in this manuscript is a design target or an illustrative simulation. The architecture is coherent and the components are individually evidenced in the cited literature, but their integrated behaviour is unverified. Nothing here should be cited as a result.
- **The required dataset does not exist.** Evaluating drift adaptation needs a multi-year, multi-site, agronomist-labelled corpus with environmental metadata. No such public resource exists. Constructing one is expensive, slow, and the binding constraint on the entire programme — not a preliminary step.
- **The exchangeability tension is managed, not resolved.** Conformal guarantees are weakened exactly when drift occurs. Recalibration on drift-matched holdouts mitigates but does not eliminate this, and the residual gap during transition windows is a genuine theoretical limitation of the approach as specified.
- **Explanation validity is assumed where it should be established.** We adopt Grad-CAM and SHAP because they are standard, while citing evidence that they can mislead [36], [37]. The sanity-check gate reduces but does not eliminate this risk. Rudin's position — that inherently interpretable models should be preferred for high-stakes decisions [37] — is not refuted by anything in this architecture, and an interpretable-by-design variant is a legitimate alternative research path.
- **Label scarcity may bind harder than modelled.** The adaptation protocol assumes agronomist labels can be obtained within an epoch of a drift event. In the smallholder contexts this work targets, extension capacity may make that infeasible, which would require a substantially more unsupervised formulation.
- **Author domain limitations.** The author's background is in data science, AI-driven cybersecurity, and enterprise IT operations, with no formal training in plant pathology, agronomy, or remote sensing, and no prior peer-reviewed publication. The architecture draws its agronomic framing from the cited literature rather than from field experience. Realising this programme would require plant pathology co-supervision and institutional field-site access; the machine learning contribution is where the author's competence actually lies, and the proposal is stronger for saying so than for implying otherwise.

XIII. RESEARCH ROADMAP AND CONCLUSION

13.1 Phased roadmap

Table 12. Phased research roadmap with overlapping execution. Phase 1 is on the critical path for every subsequent phase, which is the dominant schedule risk and should be mitigated by securing site partnerships before the programme formally begins.

Phase	Period	Objective	Output	Key risk
P1	Months 1–9	Longitudinal corpus construction; site partnerships; annotation protocol	Public multi-season benchmark	Site access and annotation cost — the programme's critical path
P2	Months 7–18	Drift detection and continual learning; ablations B3–B6	Adaptation engine; first paper	Dataset may be too short to show trend drift
P3	Months 16–27	Conformal layer under drift; calibration in transition windows	Uncertainty layer; theory contribution	Exchangeability gap may prove unbridgeable as specified
P4	Months 24–33	Temporal explainability; human evaluation with extension officers	Explanation protocol; HCI contribution	Recruiting adequate expert participants
P5	Months 30–39	Edge deployment, federated learning, field pilot	Deployed system; evaluation	Hardware logistics; connectivity
P6	Months 36–45	Integration, sustainability assessment, thesis	Thesis; open-source release	Integration debt accumulated across phases

13.2 Conclusion

Crop disease recognition is widely treated as a solved classification problem on the evidence of benchmark accuracy. We have argued that this is a category error. The difficulty is not discriminating lesions in a fixed distribution; it is sustaining reliable, honest, and explicable performance in an environment that will not hold still — across seasons, across regions, across evolving pathogen populations, and under a climate trend that does not revert.

AgroGuard AI addresses this by treating non-stationarity as the design centre rather than as deployment noise. Its specific commitments are that the system detects its own degradation through a complementary detector ensemble including an attribution-drift monitor with no standard analogue; that it adapts under hard constraints on forgetting, calibration, and explanation stability, so that accuracy gains cannot be purchased with trustworthiness; that it quantifies when it should decline to answer and widens abstention system-wide during drift, when its own difficulty estimates are least reliable; and that it explains not only its decisions but the changes in its reasoning across adaptation events.

What this manuscript establishes is an architecture, a formulation, and a protocol designed to falsify them. What it does not establish is that the architecture works, and Section 10 exists because that question is open. The honest summary is that the integration gap identified in Section 2.8 is real and the proposed resolution is coherent, but the evidence remains to be gathered — beginning with the longitudinal corpus whose absence currently makes the question unanswerable for anyone.

REFERENCES

- [1]. S. Savary, L. Willocquet, S. J. Pethybridge, P. Esker, N. McRoberts, and A. Nelson, “The global burden of pathogens and pests on major food crops,” *Nature Ecology & Evolution*, vol. 3, no. 3, pp. 430–439, 2019, doi: 10.1038/s41559-018-0793-y.
- [2]. S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” *Frontiers in Plant Science*, vol. 7, art. 1419, 2016.
- [3]. K. P. Ferentinos, “Deep learning models for plant disease detection and diagnosis,” *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, 2018.
- a. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, 2021.
- [4]. J. G. A. Barbedo, “Factors influencing the use of deep learning for plant disease recognition,” *Biosystems Engineering*, vol. 172, pp. 84–91, 2018.
- [5]. J. G. A. Barbedo, “Plant disease identification from individual lesions and spots using deep learning,” *Biosystems Engineering*, vol. 180, pp. 96–107, 2019.
- a. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, “Deep learning for image-based cassava disease detection,” *Frontiers in Plant Science*, vol. 8, art. 1852, 2017.
- [6]. M. Long, Y. Cao, J. Wang, and M. I. Jordan, “Learning transferable features with deep adaptation networks,” in *Proc. ICML*, 2015, pp. 97–105.
- [7]. Y. Ganin et al., “Domain-adversarial training of neural networks,” *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

- [8]. IPCC, *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report*. Cambridge, U.K.: Cambridge University Press, 2022.
- [9]. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.
- [10]. Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Proc. NeurIPS*, 2019.
- [11]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.
- a. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018.
- [12]. J.K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, art. 2674, 2018.
- [13]. J.R. Thapa, K. Zhang, N. Snaveley, S. Belongie, and A. Khan, "The Plant Pathology Challenge 2020 data set to classify foliar disease of apples," *Applications in Plant Sciences*, vol. 8, no. 9, art. e11390, 2020.
- [14]. J. Quiñero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press, 2009.
- [15]. J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.
- [16]. J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, art. 44, 2014.
- [17]. J.A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2007, pp. 443–448.
- [18]. C.Raab, M. Heusinger, and F.-M. Schleif, "Reactive soft prototype computing for concept drift streams," *Neurocomputing*, vol. 416, pp. 340–351, 2020.
- [19]. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.
- [20]. J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [21]. F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. ICML*, 2017, pp. 3987–3995.
- [22]. Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2018.
- [23]. S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. CVPR*, 2017, pp. 2001–2010.
- [24]. D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. NeurIPS*, 2017.
- [25]. Chaudhry et al., "On tiny episodic memories in continual learning," arXiv:1902.10486, 2019.
- [26]. G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.
- [27]. E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017, pp. 7167–7176.
- [28]. D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *Proc. ICLR*, 2021.
- [29]. Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *Proc. ICML*, 2020.
- [30]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, 2017, pp. 618–626.
- [31]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017.
- [32]. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.
- [33]. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. NeurIPS*, 2018.
- [34]. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
- [35]. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.
- [36]. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. NeurIPS*, 2017.
- [37]. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016, pp. 1050–1059.
- [38]. V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York: Springer, 2005.

- [39]. N. Angelopoulos and S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” arXiv:2107.07511, 2021.
- [40]. Y. Romano, M. Sesia, and E. J. Candès, “Classification with valid and adaptive coverage,” in *Proc. NeurIPS*, 2020.
- [41]. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, “Uncertainty sets for image classifiers using conformal prediction,” in *Proc. ICLR*, 2021.
- [42]. Y. Geifman and R. El-Yaniv, “Selective classification for deep neural networks,” in *Proc. NeurIPS*, 2017.
- [43]. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in *Proc. ICLR*, 2017.
- [44]. Jacob et al., “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in *Proc. CVPR*, 2018, pp. 2704–2713.
- [45]. Howard et al., “Searching for MobileNetV3,” in *Proc. ICCV*, 2019, pp. 1314–1324.
- [46]. S. Mehta and M. Rastegari, “MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer,” in *Proc. ICLR*, 2022.
- [47]. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proc. AISTATS*, 2017, pp. 1273–1282.
- [48]. P. Kairouz et al., “Advances and open problems in federated learning,” *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
- [49]. [C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
- [50]. M. Abadi et al., “Deep learning with differential privacy,” in *Proc. ACM CCS*, 2016, pp. 308–318.
- [51]. European Parliament and Council, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” *Official Journal of the European Union*, 13 June 2024.
- [52]. M. Mitchell et al., “Model cards for model reporting,” in *Proc. ACM Conf. on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.
- [53]. T. Gebru et al., “Datasheets for datasets,” *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.
- [54]. High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission, 2019.
- [55]. zzJ. Cohen, E. Rosenfeld, and J. Z. Kolter, “Certified adversarial robustness via randomized smoothing,” in *Proc. ICML*, 2019, pp. 1310–1320.
- [56]. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in NLP,” in *Proc. ACL*, 2019, pp. 3645–3650.
- [57]. R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, “Green AI,” *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

APPENDIX A — CONTRIBUTION SUMMARY

Table 13. Contribution summary with explicit status. The final two rows are included so that what has been done is not confused with what has been proposed.

ID	Contribution	Type	Section	Novelty position	Status
C1	Five-layer architecture integrating drift detection, continual learning, conformal UQ and temporal XAI under one control loop	Architecture	4	Novel as integration; components individually prior art	Specified
C2	Composite drift statistic with hysteresis trigger policy and attribution-drift detector	Algorithm	5.1, 5.4	Attribution-drift detector has no standard analogue	Specified
C3	Trust-constrained adaptation — calibration, forgetting and explanation stability as hard admission gates	Formulation	3.4, 8.2	Novel formulation of adaptation as constrained optimisation	Specified
C4	Temporal explainability — reporting attribution change across adaptation events	Method	7.2	Novel; existing XAI explains fixed models only	Specified
C5	Falsifiable evaluation protocol with acceptance thresholds and refutation criteria	Methodology	10	Protocol contribution	Specified
—	Empirical validation of C1–C5	Experiment	10	—	Not undertaken — see Section

					12
—	Longitudinal multi-season field corpus	Dataset	10.1	Would be first public resource of its kind	Not undertaken — prerequisite, see Section 13.1