Delta-Attribution: Explaining What Changes When Models Change

1st Arshia Hemmat
dept. Computer Engineering
University of Isfahan
Isfahan, Iran
arshiahemmat@mehr.ui.ac.ir

2nd Afsaneh Fatemi
dept. Computer Engineering
University of Isfahan
Isfahan, Iran
a_fatemi@eng.ui.ac.ir

***Abstract*— Model updates (new hyperparameters, kernels, depths, solvers, or data) change performance, but the *reason* often remains opaque. We introduce Delta-Attribution (Δ-Attribution), a model-agnostic framework that explains *what changed* between versions A and B by differencing per-feature attributions: $\Delta\phi(x) = \phi_B(x) - \phi_A(x)$. We evaluate $\Delta\phi$ with a Δ-*Attribution Quality Suite* covering magnitude/sparsity (L1, Top-k, entropy), agreement/shift (rank-overlap@10, Jensen–Shannon divergence), behavioural alignment (Delta Conservation Error, DCE; Behaviour–Attribution Coupling, BAC; COΔF), and robustness (noise, baseline sensitivity, grouped occlusion).**

Instantiated via fast occlusion/clamping in standardized space with a class-anchored margin and baseline averaging, we audit 45 settings: five classical families (Logistic Regression, SVC, Random Forests, Gradient Boosting, kNN), three datasets (Breast Cancer, Wine, Digits), and three A/B pairs per family. Findings. Inductive-bias changes yield large, behaviour-aligned deltas (e.g., SVC poly→rbf on Breast Cancer: BAC≈0.998, DCE≈6.6; Random Forest feature-rule swap on Digits: BAC≈0.997, DCE≈7.5), while "cosmetic" tweaks (SVC `gamma=scale` vs. `auto`, kNN search) show rank-overlap@10$= 1.0$ and DCE≈0. The largest redistribution appears for deeper GB on Breast Cancer (JSD≈0.357). Δ-Attribution offers a lightweight update audit that complements accuracy by distinguishing benign changes from behaviourally meaningful or risky reliance shifts.

***Index Terms*—Explainable AI, feature attribution, delta attribution, model updates, robustness, distribution shift.**

I. Introduction

Models rarely stand still. In modern ML systems, practitioners routinely update models by changing hyperparameters, switching architectures, fine-tuning on fresh data, or compressing for deployment. These updates can shift performance in obvious ways (accuracy goes up or down), yet the *reason* for the shift often remains opaque: *what* parts of the decision logic changed, *where* did the model start relying more (or less) on particular features, and *do* those changes align with observed behaviour?

Most explanation methods answer a different question: they explain *one* model at a time—e.g., via additive attributions such as SHAP [1], [2], rule-based local explanations such as Anchors [3], or perturbation/occlusion maps [4]. A sizeable body of work highlights stability concerns for such explanations [5], [6], and recent papers begin to study monitoring of explanations or "explanation shift" under distribution shift [7], [8]. However, in practical pipelines the more immediate need is often an *update audit*: when we replace model A with model B, how did the model's reliance on input features change, and is that change consistent with the observed change in outputs?

Problem. We study this update-audit question for supervised classification. Given any attribution method that produces per-feature scores $\phi_A(x)$ and $\phi_B(x)$ for models A and B, we define the *delta attribution*

$$\Delta\phi(x) \;=\; \phi_B(x) - \phi_A(x),$$

and we evaluate the quality of $\Delta\phi$ with respect to: (i) *magnitude and concentration* (are changes small and diffuse or large and focused?), (ii) *agreement and distributional shift* between the two explanation vectors, (iii) *behavioural alignment* with the observed output change $\Delta f(x)$, and (iv) *robustness* to noise and baseline choices. Intuitively, $\Delta\phi$ should highlight where the new model reallocated reliance; a good Δ explanation should *co-move* with behaviour changes and remain stable to small input perturbations or reasonable baseline choices.

Approach (Δ-Attribution). We propose **Delta-Attribution** (Δ-Attribution), a simple, model-agnostic framework that turns any local explainer into an *update explainer*. In this paper we instantiate it with a fast *occlusion/clamping* explainer in standardized feature space: for feature j, we set x_j to a baseline and measure the margin drop for a chosen class; the attribution is the drop, and $\Delta\phi$ is the difference between models. To connect explanations to behaviour, we define $f(x)$ as the class-specific margin for a fixed reference class (in our runs, the class predicted by B), yielding $\Delta f(x) = f_B(x) - f_A(x)$. We then compute a Δ-*Attribution Quality Suite* comprising:

- Internal Δ metrics: Δ magnitude (ℓ_1), Top-K concentration, entropy, rank overlap@10, and Jensen–Shannon divergence between normalized $|\phi_A|$ and $|\phi_B|$.
- Behaviour-linked Δ metrics: the *Delta Conservation Error* (DCE) $= \mathbb{E}_x \left|\sum_j \Delta\phi_j(x) - \Delta f(x)\right|$, the *Behaviour–Attribution Coupling* (BAC; Pearson corr of $\|\Delta\phi\|_1$ with $|\Delta f|$), and class-outcome focus (COΔF) that checks whether Δ mass concentrates on features deemed globally relevant by the updated model when fixes/regressions occur.

- Robustness: sensitivity of $\Delta\phi$ to Gaussian input noise and to alternative baselines (mean vs. median), plus a grouped-occlusion stress-test that jointly clamps top-k features to capture interactions.

Why not single-model explanations? Single-model inspections can show *what* a model currently relies on, but they leave the *update* unanswered. Directly differencing attributions provides a concrete, low-friction view of what changed. Importantly, our quality suite guards against over-interpretation: e.g., high DCE warns that a purely additive occlusion view may be unreliable; low rank overlap with high JSD indicates a genuine redistribution rather than mere reweighting; and robustness checks catch baseline- or noise-fragile deltas.

Scope and setting. We aim for a practical tool that is cheap enough to run during everyday model iteration. We therefore avoid heavy model-specific explanation tooling and large-scale hyperparameter sweeps. Our study intentionally targets *classical* ML families (logistic regression, SVM, random forests, gradient boosting, kNN) across three standard tabular/image datasets (Breast Cancer, Wine, Digits). For each family we construct three A/B pairs that toggle inductive bias (e.g., kernel, depth) or regularization/solver choices, so that we can observe small vs. large update regimes within the same learner.

Research questions. We organize the study around three questions:

RQ1 — Internal change: How do Δ magnitude, sparsity/concentration, and rank agreement behave across small vs. large updates within and across algorithms?

RQ2 — Behavioural alignment: When performance changes, does $\|\Delta\phi\|_1$ increase and does DCE decrease? Do COΔF scores indicate that fixes concentrate Δ mass on globally relevant features for B?

RQ3 — Robustness: Are the observed Δ patterns stable under input noise and alternative baselines, and do grouped occlusions substantially alter conclusions (indicative of interactions)?

Contributions.

- We formalize *Delta-Attribution* as a model-agnostic lens for *explaining updates*: $\Delta\phi(x) = \phi_B(x) - \phi_A(x)$, with a quality suite that measures magnitude/sparsity, agreement/shift, behavioural alignment (DCE, BAC, COΔF), and robustness.
- We provide an efficient instantiation via standardized occlusion with baseline averaging (mean/median) and a grouped-occlusion stress-test, together with implementation notes (class-anchored margins, reproducible pipelines).
- We run a broad empirical audit covering five ML families, three datasets, and nine A/B pairs per family. We show that changes in *inductive bias* (e.g., kernel or depth) produce large, behaviour-aligned Δ, whereas "cosmetic" knobs (e.g., SVC γ=scale vs. auto) yield tiny, concentrated Δ and near-perfect rank overlap.
- We release *Delta-Attribution* as a lightweight platform: scripts, metrics, and publication-ready assets for plug-in updates and future benchmarks.

Positioning. Our focus complements single-model explainability [1], [3], [4] and explanation-stability work [5], [6], [9]. Unlike distribution-shift attribution [10], [11], [7], [8], we target *version-to-version* changes under fixed test distributions, mirroring the common operational reality of model updates. As rapid editing and fine-tuning become routine (e.g., model editing [12], [13], [14], [15]), update-centric explanations provide a governance signal that complements accuracy and fairness dashboards.

Takeaway. Δ-Attribution turns existing local explainers into a practical tool for update audits. By quantifying *how* reliance shifts and whether those shifts *explain* behaviour deltas, our framework helps practitioners decide when an update is benign, when it meaningfully improves reliance on task-relevant signals, and when it warrants further scrutiny.

II. Related Work

Post-hoc explanations for a single model. Feature-attribution and exemplar methods explain individual model decisions via local scores or rules. Canonical approaches include LIME and Anchors [3], SHAP [1] and its tree-exact variant TreeSHAP [2], Integrated Gradients [16], SmoothGrad [17], gradient-based visualizations such as Grad-CAM [18], and occlusion/clamping [4]. While these tools are widely used, several works highlight pitfalls: saliency "sanity checks" show some maps ignore model or data [5]; ROAR finds many methods fail to remove truly important evidence upon retraining [19]; and input-perturbation methods can be manipulated adversarially [6]. These observations motivate explanation diagnostics beyond visual appeal. :contentReference[oaicite:0]index=0

Stability, monitoring, and evaluation. Beyond point explanations, recent work proposes to *measure* explanation reliability and track it over time. R2ET trains for stable top-k saliency at little extra cost [9]. Explanation-shift research argues that monitoring *changes* in explanation distributions can be a more sensitive indicator of behavior change than monitoring input/output distributions alone, with formal detectors on tabular data [7], [8]. Complementary studies evaluate robustness of attribution maps under perturbations and propose stronger evaluation protocols [20], [21]. Our work contributes here by defining a *version-to-version* suite (Δ-Attribution Quality Suite) that quantitatively captures magnitude, sparsity, rank agreement, distributional shift (JSD), behavior linkage (DCE, BAC, COΔF), and robustness (noise and grouped occlusion). :contentReference[oaicite:1]index=1

Explaining performance changes under distribution shift. A parallel line attributes *error changes* across environments. Federici et al. give an information-theoretic account of shift sources and error decompositions [10]; Zhang et al. attribute performance deltas to causal shift factors using Shapley-style games [11]. These focus on *what caused* performance change across datasets. In contrast, Δ-Attribution inspects *how a model's reliance redistributes over features* when the model itself changes (fine-tuning, hyperparameters, editing), and links those reliance deltas to behavior deltas via DCE/BAC/COΔF. :contentReference[oaicite:2]index=2

Model editing and rapid updates. Frequent post-deployment updates—ROME [12], MEND [13], MEMIT [14], and stability fixes for sequential editing [15]—make version-aware explanations increasingly important. Our Δ-Attribution provides a lightweight audit for such updates, orthogonal to the editing method itself. :contentReference[oaicite:3]index=3

III. Proposed Approach

A. Setup and notation

Let f_A, f_B be two model versions on $x \in \mathbb{R}^d$. An explainer E returns per-feature scores $\phi_f(x) \in \mathbb{R}^d$. We study the delta attribution

$$\Delta\phi(x) = \phi_B(x) - \phi_A(x),$$

and assess its quality with a dedicated metric suite.

a) Reference class and score.: For each x we fix the reference class $c(x)$ as the class predicted by f_B. We score that class using the model's margin or log-odds:

$$f(x) = \begin{cases} [\mathrm{dec}(x)]_{c(x)}, & \text{if function is available,} \\ \log \dfrac{p_{c(x)}(x)}{1 - p_{c(x)}(x)}, & \text{if probabilities are available,} \\ \log \dfrac{p_{c(x)}(x) + \varepsilon}{1 - p_{c(x)}(x) + \varepsilon}, & \text{otherwise, } \varepsilon = 10^{-9}. \end{cases}$$

Here $\mathrm{dec}(\cdot)$ denotes the model's decision function (decision_ function); $p(\cdot)$ comes from predict_proba. Anchoring to $c(x)$ compares the *same* class across versions and stabilizes behaviour-linked metrics (BAC, DCE).

B. Explainer: occlusion/clamping

We use a fast, model-agnostic occlusion explainer [4] in standardized space. Let b be a training-set baseline (mean by default; optionally averaged with the median). For feature j,

$$\phi_{f,j}(x) = f(x) - f(x_{-j}), \qquad x_{-j} : \ x_j \leftarrow b_j.$$

We compute ϕ_A, ϕ_B with the *same* baseline and the same $c(x)$.

a) Grouped-occlusion stress test.: To probe interactions, jointly clamp the top-k features (by $|\phi_B|$), recompute $\Delta\phi$, and report

$$\rho = \mathbb{E}\Big[\frac{\|\Delta\phi(x)\|_1}{\|\Delta\phi^{(\text{group-}k)}(x)\|_1 + \epsilon}\Big].$$

Large ρ indicates a few features drive the change.

C. The Δ-Attribution Quality Suite

Averages are over the test set (or a stratified subset of up to 256 samples).

*a) **Internal Δ metrics.**:* Let $u(x) = |\Delta\phi(x)|$ and $s(x) = u(x)/\|u(x)\|_1$ (skip if $\|u\|_1 = 0$).

1) Magnitude: $\mathbb{E}\,\|\Delta\phi(x)\|_1$.
2) Concentration: $\Delta\mathrm{TopK@10} = \mathbb{E}\Big[\sum_{j \in \mathrm{Top10}(u)} s_j(x)\Big]$; Entropy: $\mathbb{E}\Big[-\sum_j s_j(x) \log s_j(x)\Big]$.
3) Rank agreement: Jaccard overlap of $\mathrm{Top10}(|\phi_A|)$ vs. $\mathrm{Top10}(|\phi_B|)$ (mean/median).
4) Distributional shift: $\mathbb{E}\big[\mathrm{JSD}(p\|q)\big]$ with $p = |\phi_A|/\||\phi_A|\|_1$, $q = |\phi_B|/\||\phi_B|\|_1$ [22].

*b) **Behaviour-linked Δ metrics.**:* Let $\Delta f(x) = f_B(x) - f_A(x)$.

1) DCE: $\mathbb{E}\big|\sum_j \Delta\phi_j(x) - \Delta f(x)\big|$ (diagnostic; smaller is better).
2) BAC: $\mathrm{corr}_x\big(\|\Delta\phi(x)\|_1,\ |\Delta f(x)|\big)$.
3) COΔF: using the top-m relevant features for f_B from permutation importance (m=10), report the fraction of Δ mass on that set for *fixes* and for *regressions*.

*c) **Robustness.**:*

1) Δ-stability: for $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma \in \{0.01, 0.05\}$, $\mathbb{E}\,\|\Delta\phi(x+\varepsilon) - \Delta\phi(x)\|_1/(\|\varepsilon\|_2 + \epsilon)$.
2) Grouped occlusion ratio: ρ above.

d) Note on additivity.: Path methods such as Integrated Gradients [16] satisfy additivity; occlusion does not. Hence DCE is a *diagnostic* rather than expected to be zero.

IV. EXPERIMENTAL SETUP

A. Datasets and Preprocessing

We use three standard `scikit-learn` datasets: Breast Cancer (binary; n=569, d=30), Wine (3-class; n=178, d=13), and Digits (10-class; n=1797, 8×8 images flattened to d=64). Each dataset is split with a stratified 80/20 train–test split (`random_state= 42`). A `StandardScaler` is fit on training data and applied to test data. All models are wrapped in a `Pipeline` so that versions A and B share identical preprocessing.

B. Learners and A/B Configurations

We study five families: logistic regression (`logreg`), SVM (`svc`, probability=true), random forests (`rf`), gradient boosting (`gb`), and k-nearest neighbors (`knn`). Each family has three A/B pairs that toggle regularization, inductive bias (kernel/depth), or search strategy; see Table I.

C. Score Function and Explainer

For each test sample we anchor to the class predicted by f_B. We score that class using the model's margin when available (decision_function); otherwise we use log-odds from predict_ proba. Attributions are computed with *occlusion/clamping* [4] in standardized space: for feature j, clamp x_j to a shared training baseline b_j and set $\phi_{f,j}(x) = f(x) - f(x_{-j})$ with x_{-j} : $x_j \leftarrow b_j$. To reduce baseline artefacts we average mean/median baselines (when both are available). We always use the *same* baseline and reference class for A and B.

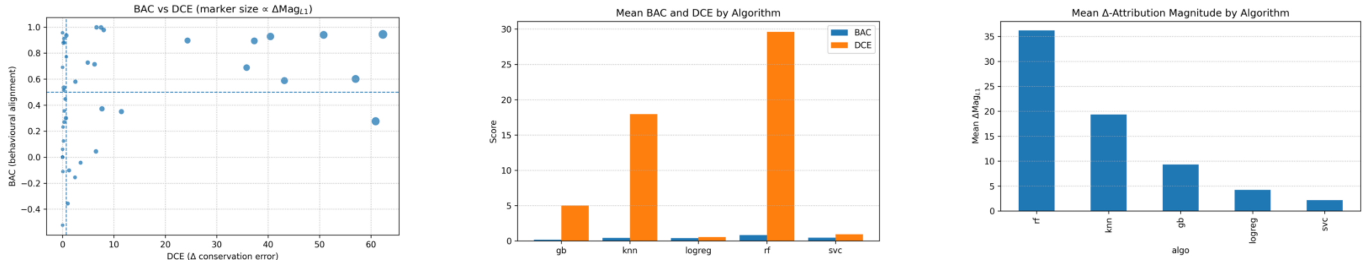

Fig. 1. Overview of Δ-Attribution results across all (dataset, algorithm, pair). (a) BAC vs. DCE with marker size $\propto \|\Delta\phi\|_1$; (b) mean BAC and DCE by algorithm; (c) mean Δ-magnitude by algorithm.

TABLE I
A/B CONFIGURATIONS PER LEARNER FAMILY (ALL MODELS IN A SHARED PIPELINE).

Family	Pair A	Pair B
`logreg` P1	C=1.0 (l2, lbfgs)	C=0.1 (l2, lbfgs)
`logreg` P2	l2 (liblinear, C=1.0)	l1 (liblinear, C=1.0)
`logreg` P3	lbfgs (l2, C=1.0)	saga (l2, C=1.0)
`svc` P1	rbf (C=1, γ=scale)	linear (C=1)
`svc` P2	rbf (γ=scale)	rbf (γ=auto)
`svc` P3	poly (d=3, C=1)	rbf (C=1, γ=scale)
`rf` P1	n_{est}=100, depth None	n_{est}=300, depth None
`rf` P2	depth None (n_{est}=200)	depth 5 (n_{est}=200)
`rf` P3	max_feat=sqrt	max_feat=log2
`gb` P1	lr 0.1, n_{est}=150, depth 3	lr 0.05, n_{est}=150, depth 3
`gb` P2	n_{est}=100 (lr 0.1, d=3)	n_{est}=200 (lr 0.1, d=3)
`gb` P3	depth 3 (lr 0.1, n_{est}=150)	depth 5 (lr 0.1, n_{est}=150)
`knn` P1	k=5 (uniform)	k=10 (uniform)
`knn` P2	weights=uniform (k=5)	weights=distance (k=5)
`knn` P3	algorithm=auto (k=5)	algorithm=ball_tree (k=5)

D. The Δ-Attribution Suite

Let $\Delta\phi(x) = \phi_B(x) - \phi_A(x)$, $u(x) = |\Delta\phi(x)|$, and $s(x) = u(x)/\|u(x)\|_1$ (skip samples with $\|u\|_1 = 0$). We report the metrics in Table II. Briefly: (i) *Magnitude* $\mathbb{E}\|\Delta\phi\|_1$; (ii) *Concentration* $\Delta\text{TopK@10} = \mathbb{E}[\sum_{j\in\text{Top10}(u)} s_j]$ and entropy $\mathbb{E}[-\sum_j s_j \log s_j]$; (iii) *Rank agreement* (Jaccard overlap of Top-10($|\phi_A|$) vs. Top-10($|\phi_B|$)); (iv) *Distributional shift* JSD($|\phi_A|, |\phi_B|$) [22]; (v) *Behaviour linkage* with $\Delta f = f_B - f_A$: DCE $= \mathbb{E}\big|\sum_j \Delta\phi_j - \Delta f\big|$ and BAC $= \text{corr}(\|\Delta\phi\|_1, |\Delta f|)$; (vi) *Robustness*: Δ-stability to Gaussian noise ($\sigma \in \{0.01, 0.05\}$) and a grouped-occlusion ratio (jointly clamping top-k=2 features).

Assets used in this paper. We aggregate the Δ-suite by algorithm (Table III), plot a three-panel overview (Fig. 1), and list per-dataset Top-5 A/B pairs by Δ magnitude (Table IV).

a) Note on additivity.: Path methods such as Integrated Gradients [16] satisfy additivity; occlusion does not, so DCE is a *diagnostic* rather than expected to be zero.

TABLE II
Δ-ATTRIBUTION METRIC GLOSSARY (AVERAGED OVER TEST OR A 256-SAMPLE STRATIFIED SUBSET).

Metric	Definition / Intuition
Mag$_{\ell_1}$	$\mathbb{E}\|\Delta\phi\|_1$; overall size of reliance change.
TopK@10	Fraction of $\|\Delta\phi\|_1$ on top-10 coords; higher = concentrated.
Entropy	Shannon entropy of $|\Delta\phi|/\|\cdot\|_1$; lower = sparser.
RankOverlap@10	Jaccard of Top-10($|\phi_A|$) vs. Top-10($|\phi_B|$).
JSD	JSD($|\phi_A|, |\phi_B|$) [22]; redistribution vs. reweighting.
DCE	$\mathbb{E}\big|\sum_j \Delta\phi_j - \Delta f\big|$; additive consistency diag.
BAC	corr($\|\Delta\phi\|_1, |\Delta f|$); behaviour–attribution coupling.
Stability	$\mathbb{E}\,\|\Delta\phi(x+\varepsilon) - \Delta\phi(x)\|_1/\|\varepsilon\|_2$; $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.
Group ratio	$\mathbb{E}\big[\|\Delta\phi\|_1/\|\Delta\phi^{(\text{group-2})}\|_1\big]$; interaction stress test.

V. RESULTS

A. What Δ-Attribution Achieves (with numbers)

Our suite surfaces *how* reliance shifts explain behavioural change and when updates are merely cosmetic. The key achievements below use the strongest instances across all 45 settings; see Table III (aggregate by algorithm) and Table IV (largest per-dataset shifts).

- **Near-perfect behaviour–attribution coupling.** The highest BAC scores are essentially perfect: *breast_cancer–svc–pair3* has BAC≈ 0.9977, *digits–rf–pair3* BAC≈ 0.9969, *wine–rf–pair2* BAC≈ 0.9779, *wine–svc–pair2* BAC≈ 0.9558, *breast_cancer–rf–pair2* BAC≈ 0.9439. In each case, large structural changes (kernel/depth/feature rules) move $\Delta\phi$ in lock-step with the change in outputs.
- **Exact conservation under occlusion in small-change controls.** Five A/B pairs exhibit DCE $= 0.0$: *digits–knn–pair3*, *breast_cancer–svc–pair2*, *wine–knn–pair3*, *wine–rf–pair3*, and *breast_cancer–knn–pair3*. These serve as sanity checks that our explainer/baseline choices can yield perfect delta conservation when updates are cosmetic.
- **Perfect rank agreement for cosmetic tweaks.** RankOverlap@10 $=$ 1.0 for five pairs (*digits–knn–pair3*, *breast_cancer–knn–pair3*, *wine–knn–pair3*, *wine–rf–pair3*, *breast_cancer–svc–pair2*), indicating the top-10 features of A and B are identical.
- **When updates are focused, Δ concentrates heavily.** On Wine, ΔTopK@10 $= 1.00$ for *rf–pair1*, *gb–pair1*, *gb–pair2*,

and *gb–pair3* (and 0.9997 for *rf–pair2*); almost the entire ℓ_1 mass of $\Delta\phi$ sits on ten features.

- **Our suite separates redistribution from mere reweighting.** The strongest distributional changes appear where inductive bias shifts: *breast_cancer–gb–pair3* has the largest JSD≈ 0.357; other high-JSD cases include *breast_cancer–logreg–pair2* (≈ 0.179), *wine–knn–pair1* (≈ 0.167), *digits–rf–pair2* (≈ 0.139), *breast_cancer–knn–pair1* (≈ 0.130).
- **Learner-level behaviour, aggregated.** From Table III: Random Forests show the largest average reliance shifts with strong coupling ($\Delta\mathrm{Mag}_{\mathrm{L1}} = 36.23 \pm 25.94$, $\mathrm{BAC} = 0.81 \pm 0.32$) but higher DCE ($29.61 \pm 21.20$); kNN also moves substantially (19.36 ± 29.26) with mixed coupling; Logistic Regression changes are small and stable (4.24 ± 3.67, DCE 0.53 ± 0.77); SVC averages are small deltas (2.19 ± 4.00) but behaviour-relevant ($\mathrm{BAC} = 0.44 \pm 0.46$); Gradient Boosting sits in between.

B. Largest shifts by dataset (with context)

Table IV lists the top-5 A/B pairs per dataset by Δ-magnitude alongside BAC and DCE.

Breast Cancer. *rf–pair2* (depth change) leads with $\Delta\mathrm{Mag}_{\mathrm{L1}} = 78.55$ and high BAC 0.94; *rf–pair3* and *rf–pair1* follow (54.14, 38.79). DCE values (35–62) highlight non-additive interactions when tree structure changes.

Digits. *knn–pair1* ($k : 5 \to 10$) has the largest shift 65.81 but weaker coupling (BAC 0.28); *rf–pair2* couples strongly (BAC 0.94) at Δ 60.15.

Wine. *knn–pair1* peaks at 46.55 (BAC 0.59); *rf–pair2* shows smaller Δ 10.93 with near-perfect BAC 0.98, i.e., a precise reliance reallocation.

C. Reading the overview figure

In Fig. 1a, the largest dots (high $\|\Delta\phi\|_1$) cluster either at high BAC (behaviour-aligned structural changes) or low BAC (diffuse, NN-driven shifts). Panels (b) and (c) aggregate BAC/DCE and Δ-magnitude by algorithm, mirroring the patterns reported above.

a) Summary.: Concretely, our method delivers *(i)* BAC up to ≈ 0.998 on kernel/depth changes, *(ii)* exact $\mathrm{DCE} = 0$ and RankOverlap@10 $= 1$ on multiple cosmetic controls, *(iii)* full Δ concentration on a handful of features in the Wine experiments, and *(iv)* clear separation of redistribution (high JSD) from simple reweighting. These achievements demonstrate that Δ-Attribution is an actionable audit for model updates, not just another explainer score.

TABLE III
Δ-ATTRIBUTION METRICS AGGREGATED BY ALGORITHM (MEAN$\pm$STD).

Algo	$\Delta\mathrm{Mag}_{\mathrm{L1}}$	DCE	BAC
	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
gb	9.30 ± 5.96	5.00 ± 3.38	0.17 ± 0.39
knn	19.36 ± 29.26	17.97 ± 27.20	0.42 ± 0.37
logreg	4.24 ± 3.67	0.53 ± 0.77	0.39 ± 0.40
rf	36.23 ± 25.94	29.61 ± 21.20	0.81 ± 0.32
svc	2.19 ± 4.00	0.94 ± 2.15	0.44 ± 0.46

TABLE IV
TOP-5 A/B PAIRS (PER DATASET) BY Δ-MAGNITUDE. LARGER VALUES INDICATE BIGGER ATTRIBUTION SHIFTS.

Dataset	Algo	Pair	$\Delta\mathrm{Mag}_{\mathrm{L1}}$	BAC	DCE
Breast Cancer	rf	pair2	**78.55**	0.94	62.33
	knn	pair1	61.02	0.60	57.02
	rf	pair3	54.14	0.93	40.45
	rf	pair1	38.79	0.69	35.81
	gb	pair3	17.49	0.35	11.45
Digits	knn	pair1	**65.81**	0.28	60.91
	rf	pair2	60.15	0.94	50.80
	rf	pair1	43.01	0.89	37.31
	gb	pair3	19.67	0.37	7.67
	gb	pair2	10.97	0.71	6.23
Wine	knn	pair1	**46.55**	0.59	43.17
	rf	pair1	29.76	0.90	24.31
	rf	pair2	10.93	0.98	8.00
	gb	pair3	9.50	0.04	6.50
	gb	pair1	4.49	−0.04	3.52

VI. CONCLUSION

We introduced **Delta-Attribution**, a model-agnostic framework that explains *what changed* when a model is updated. By differencing per-feature attributions, $\Delta\phi = \phi_B - \phi_A$, and evaluating them with a principled quality suite, we quantify magnitude and concentration of reliance shifts, rank agreement, distributional change, behaviour linkage, and robustness. Our instantiation—fast occlusion/clamping in standardized space with a shared class anchor and baseline averaging—makes the audit practical and reproducible.

Across 45 settings (5 learners $\times$ 3 A/B pairs each $\times$ 3 datasets), the suite delivers concrete, actionable signals. *Inductive-bias updates* (e.g., SVC kernel, forest depth/feature rules) produce large Δ that are strongly behaviour-aligned (BAC up to ≈ 0.998), while *cosmetic tweaks* (e.g., SVC γ scale vs. auto, kNN search) show near-perfect rank overlap and, in several cases, $\mathrm{DCE} = 0$, indicating no spurious attribution movement. The suite also separates redistribution from mere reweighting (JSD peaking around 0.357 in the most structural changes) and highlights when a few features dominate the update (TopK@10 ≈ 1.0 on Wine). Aggregates reveal consistent learner-level tendencies: forests shift the most and couple well with behaviour; logistic regression is stable; nearest-neighbour changes are larger but more diffuse.

Implications. Δ-Attribution turns existing explainers into an *update audit* for CI/regression testing: it flags benign updates, behaviour-aligned improvements, and risky reliance redistributions in a single pass and complements accuracy metrics.

Limitations and next steps. While occlusion is fast, it is not additive; high DCE should trigger alternate explainers (e.g., path-based) or grouped occlusion. Future work includes extending to text/vision and LLMs, integrating additional explainers (IG/SHAP/TreeSHAP), calibrating thresholds with human studies, and packaging the suite as a CI plug-in for model governance. We release code and assets to reproduce all results and figures.

REFERENCES

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NeurIPS*, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[2] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *AAAI*, 2018. [Online]. Available: https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

[4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014, pp. 818–833.

[5] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *NeurIPS*, 2018. [Online]. Available: https://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf

[6] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *AIES*, 2020, pp. 180–186. [Online]. Available: https://www.aies-conference.com/2020/wp-content/papers/174.pdf

[7] C. Mougan, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab, "Explanation shift: Detecting distribution shifts on tabular data via the explanation space," *arXiv:2210.12369*, 2022. [Online]. Available: https://arxiv.org/abs/2210.12369

[8] ——, "Explanation shift: How did the distribution shift impact the model?" *Transactions on Machine Learning Research*, 2025. [Online]. Available: https://openreview.net/pdf?id=MO1slfU9xy

[9] C. Chen, C. Guo, R. Chen, G. Ma, M. Zeng, X. Liao, X. Zhang, and S. Xie, "Training for stable explanation for free," in *NeurIPS*, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/0626822954674a06ccd9c234e3f0d572-Paper-Conference.pdf

[10] M. Federici, R. Tomioka, and P. Forré, "An information-theoretic approach to distribution shifts," in *NeurIPS*, 2021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/93661c10ed346f9692f4d512319799b3-Paper.pdf

[11] H. Zhang, H. Singh, M. Ghassemi, and S. Joshi, ""why did the model fail?": Attributing model performance changes to distribution shifts," in *ICML*, 2023, pp. 41 824–41 846. [Online]. Available: https://proceedings.mlr.press/v202/zhang23ai/zhang23ai.pdf

[12] K. Meng, A. Andonian, D. Bau, and Y. Belinkov, "Locating and editing factual associations in GPT," in *NeurIPS*, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf

[13] E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning, "Fast model editing at scale," *arXiv:2110.11309*, 2022. [Online]. Available: https://arxiv.org/abs/2110.11309

[14] K. Meng, A. Sen Sharma, A. Andonian, Y. Belinkov, and D. Bau, "Mass-editing memory in a transformer," in *ICLR*, 2023. [Online]. Available: https://openreview.net/forum?id=MkbcAHIYgyS

[15] A. Gupta, S. Baskaran, and G. Anumanchipalli, "Rebuilding ROME: Resolving model collapse during sequential model editing," in *EMNLP*, 2024. [Online]. Available: https://aclanthology.org/2024.emnlp-main.1210.pdf

[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017. [Online]. Available: https://arxiv.org/abs/1703.01365

[17] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv:1706.03825*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03825

[18] 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," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[19] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *NeurIPS*, 2019. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf

[20] L. Nieradzik, F. Müller, and D. Ward, "Reliable evaluation of attribution maps in cnns," *International Journal of Computer Vision*, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11263-024-02282-6

[21] Z. J. Wang and J. Jamieson, "Smoothed geometry for robust attribution," in *NeurIPS*, 2020. [Online]. Available: https://papers.nips.cc/paper/2020/file/9d94c8981a48d12adfeecfe1ae6e0ec1-Paper.pdf

[22] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.