Trust, but Verify: Multi-Signal Diagnostics for Test-Set Label Reliability

Machine learning progress is measured against benchmark test sets treated as ground truth, yet independent audits have found average label-error rates of 3–7% across widely used benchmarks, and correcting these errors can invert the relative ranking of competing models. While an extensive toolkit exists for diagnosing label noise in training data (dataset cartography, confident learning, influence functions), no comparable toolkit exists for test sets, despite test-set errors being structurally more harmful: they directly contaminate reported numbers with no training-time self-correction mechanism to buffer them. We address this gap with three complementary mechanisms for scoring test-point label reliability, deliberately built on different signal sources so that they fail in disjoint regimes: Trajectory-Based Test Cartography (TBTC), which extends dataset-cartography confidence/variability trajectories from training data to the test set using a single model's per-epoch predictions; Ensemble Disagreement Cartography (EDC), which decomposes disagreement across an ensemble of independently trained models into aleatoric and epistemic components to flag confidently-wrong labels; and Geometric Neighborhood Consistency (GNC), which audits a label via k-nearest-neighbor consistency in embedding space using no model predictions at all. Each mechanism outputs a per-point suspicion score in $[0, 1]$.

Evaluating all three on CIFAR-100 (image), IMDB Sentiment (text), and UCI Adult Income (tabular) under synthetic noise recovery, agreement with Confident Learning as an algorithmic ground truth, and stress tests that each violate one mechanism's core assumption, we find that under controlled synthetic noise all three mechanisms achieve Average Precision competitive with the established Confident Learning baseline (within 0.01–0.03 AP across datasets), while under stress they degrade in largely disjoint regimes: on biased training data, GNC alone remains workable (AP $= 0.672$) once TBTC and EDC are corrupted by the shared model bias, whereas on a frozen encoder, TBTC and EDC are unaffected while GNC collapses. This regime-specific complementarity is the empirical case for deploying the three mechanisms jointly rather than relying on any one of them. The work extends training-data diagnostics such as dataset cartography and the area-under-margin statistic to the neglected problem of test-set auditing, and repurposes deep-ensemble uncertainty decomposition and embedding-based k-NN noise detection as label-auditing tools rather than predictive-confidence tools – arguing that test sets deserve the same auditing infrastructure as training sets, and offering a practical, low-overhead recipe (TBTC and EDC as primary detectors, GNC as an independent sanity check) for establishing an upper bound on a benchmark's label reliability before reporting results.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.