Soft Computing Fusion with Applications

www.scfa.reapress.com

Soft. Comput. Fusion. Appl. Vol. 3, No. 2 (2026) 87–98.

Paper Type: Original Article

Label-Efficient Active Learning on Clinical-Style Tabular Data: Uncertainty, Margin, and Committee Strategies Across Imbalance and Noise Regimes

Mahmoud Alimoradi[1,*], Awaz Ahmed Shaban[2], Isametova Madina Esdauletova[3]

[1] Department of Computer Engineering and Information Technology, La.C., Islamic Azad University, Iran; mahmoud.alimorady@gmail.com.

[2] Department of Information Technology, Technical College of Informatics–Akre, Akre University for Applied Sciences, Duhok, Iraq; Awaz.Amedy1@gmail.com.

[3] Satbayev University, 050000 Almaty, Kazakhstan; isametova69@mail.ru.

Citation:

Received: 14 January 2026 *Revised: 21 March 2026* *Accepted: 17 May 2026*	Alimoradi, M., Ahmed Shaban, A., & Madina Esdauletova, I. (2026). Label-efficient active learning on clinical-style tabular data: Uncertainty, margin, and committee strategies across imbalance and noise regimes. *Soft Computing Fusion with Applications, 3*(2), 87-98.

Abstract

This study compares uncertainty, margin, entropy and Query-by-Committee (QbC) acquisition with random sampling on the Wisconsin Diagnostic Breast Cancer dataset and two derived stress regimes. The reported design uses thirty seeds, three learners, twenty initial labels and batches of ten up to 150 labels. The retained summaries report favorable Area-Under-Learning-Curve (AULC) differences in all 36 strategy-regime-learner cells and Holm-adjusted paired Wilcoxon p-values at most 0.0039. Under the noisy regime, logistic-regression AULC is reported as 0.9587 for random and 0.9837 for uncertainty; the corresponding mean retrospective threshold budgets are 32.0 and 24.7 labels, approximately 23% lower. QbC has the largest reported mean AULC gain under imbalance for each learner. Favorable cell medians coexist with individual seeds below random sampling. AULC integration, binary-score tie handling, ANOVA units and execution provenance require original code and outputs to validate these comparisons. The conclusions concern one retrospective benchmark with controlled fixed-count label corruption, rather than clinical annotation policies or a general benefit of noisy labels.

Keywords: Active learning, Label efficiency, Class imbalance, Label noise, Query-by-committee, Clinical tabular data.

1 | Introduction

Supervised machine learning in clinical settings is rarely limited by the availability of unlabeled patient records; it is limited by the expert time required to label them. Diagnosis codes, histology readings, and adjudicated outcomes are produced by scarce specialists, so the labeling budget of a typical clinical modeling project covers only a small fraction of the candidate cohort. Active learning addresses this bottleneck by letting the

Corresponding Author: mahmoud.alimorady@gmail.com

https://doi.org/10.22105/scfa.v3i2.102

Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

model choose which instances to label: starting from a small seed set, the learner scores the unlabeled pool, queries the most promising instances, and re-trains [1]. In the three decades since uncertainty sampling and Query-by-Committee (QbC) were introduced [2], [3], a large toolbox of acquisition heuristics has accumulated, but their behavior under the data pathologies that characterize clinical tabular data—class imbalance and label noise—remains patchily documented.

WDBC contains thirty morphology descriptors and binary diagnoses [4]. The stress regimes examine class subsampling and corrupted training annotations separately; they do not represent two independent cohorts or a simultaneous imbalance-plus-noise condition. Uniform corruption locations provide a controlled intervention but cannot represent every real annotation-error mechanism [1].

The original data contain 569 records, 212 malignant and 357 benign [4]. The imbalance condition retains 89 malignant records and all benign records, redrawing the subsample per seed; the ratio is approximately 1:4. The noise condition changes 39 of 398 training labels at uniformly selected positions fixed before querying, a realized rate $39/398 = 0.09799$. Test labels remain clean. Five strategies and three learners are compared across the same initial labeling budgets.

The recorded Area-Under-Learning-Curve (AULC) summaries favor active acquisition across the reported cells. Logistic regression under noise increases from 0.9587 to 0.9837, with the retrospective mean threshold budget decreasing from 32.0 to 24.7. Nonnegative cell medians do not mean every individual run benefits. The study examines a paired simulation design; validating its numerical conclusions requires the original trajectories, acquisition logs, statistical analysis and versioned execution materials.

2 | Related Work

Uncertainty, margin and entropy scores for one binary probability vector are monotone functions of distance from probability one half [1], [2], [5]. With exactly the same fitted model, pool and tie convention, they therefore select identical ranked batches. This mathematical equivalence does not establish that separate stochastic fitting or implementation paths were identical.

QbC bases acquisition on disagreement among hypotheses [3], [6]. The tested variant is described as a five-member bagged committee of the selected learner, scored by normalized vote variance. This particular bootstrap and score choice is an implementation of the committee idea, rather than the theoretical selective-sampling construction analyzed in [6]. Historical pool-based text classification combined committee acquisition with expectation maximization [7].

Density- and hierarchy-based acquisition offers a complementary emphasis on input coverage. Dasgupta and Hsu analyze hierarchical sampling under their specified model [8]. Such results motivate considering representation and class prevalence, but do not imply that uncertain boundary regions are always majority dominated. Noisy-oracle behavior is also sensitive to the acquisition rule and error mechanism [1].

The learners are regularized logistic regression, random forest [9] and histogram gradient boosting, drawn from scikit-learn [10]. The classical gradient-boosting source [11] describes the methodological family, while the concrete histogram estimator has library-specific implementation and defaults. The comparison holds these configurations fixed; it does not rank all members of their model families.

3 | Research Gap and Questions

The study asks how acquisition rules interact with learner choice and two separately constructed stress regimes. It also records cell-level and per-seed downside outcomes. It does not provide a general bound on active-learning risk or establish that previous comparisons were uniformly clean, balanced or anecdotal.

The research questions concern retrospective label efficiency and variation among the specified learners and regimes.

RQ1 asks how the recorded AULC and retrospective budgets to test AUC 0.95 differ from random sampling, and how frequently individual seeds show a deficit.

RQ2 asks whether the reported mean strategy ordering changes with learner and regime, including the approximate 1:4 imbalance condition.

Two directional hypotheses guide interpretation. H1 anticipates benefits from committee and margin sampling under imbalance; H2 anticipates degradation of uncertainty sampling relative to random under noise. No dated registration record is supplied, so these are stated hypotheses without a preregistration claim. Comparisons use the reported Holm-corrected paired AULC tests, with their direction retained regardless of the anticipated outcome.

4 | Materials and Methods

4.1 | Problem Setting

In the reported pool-based design, labels are encoded as 0 for malignant and 1 for benign. A labeled set and unlabeled pool are disjoint subsets of training records. Each round fits the current learner, scores the pool and queries ten labels. Budgets are 20,30,…,150, giving fourteen evaluations and thirteen acquisition rounds. Strategies within a seed share the split and initial labeled set; random-stream coupling, model fits and tie resolution require explicit implementation logs to establish exact pairing.

4.2 | Acquisition Strategies

Random sampling selects a uniform batch. Uncertainty maximizes $1 - \max(p0,p1)$, margin minimizes $|p1 - p0|$, and entropy maximizes binary predictive entropy with the endpoint convention $0 \log 0 = 0$ [1], [2], [5]. For a common fitted binary model these are rank-equivalent, so any batch difference must be explained by ties, numerical evaluation, different fits or random-stream handling. The record states a minimum score-rank correlation 0.9991 over 6,660 score pairs, 288 selection mismatches and fallback loops in all 90 environments. These counts require acquisition logs: near-unit rank correlation is not a substitute for exact batch equality, and floating-point arithmetic alone has not been shown to explain the reported learner-specific differences.

QbC trains five learner instances on bootstrap resamples of the current labeled set [3], [6]. If v is the fraction voting for class 1, the acquisition score is $4v(1-v)$, in [0,1]. Single-class resamples are described as redrawn up to ten times, followed by a full-set fit if needed; the record reports zero fallbacks. The tie-breaking rule and dedicated random streams must preserve intended pairing while allowing strategy-specific query histories.

4.3 | Learners

The recorded learners are LogisticRegression with L2 regularization and maximum 1,000 iterations, RandomForest with 30 trees, and HistGradientBoosting with 50 iterations, each using random_state 42 [10]. These smaller ensembles define the evaluated configurations. The histogram booster is described as using minimum leaf size 20, which ordinarily prevents a two-leaf split at budgets 20 and 30 in this unweighted design and produces constant scores with AUC 0.5. The recorded floor is a configuration property, rather than a fundamental active-learning limitation; all remaining defaults and fitted behavior need the environment and logs.

4.4 | Data Regimes

The original condition uses 569 WDBC records [4]. The imbalance condition retains 89 malignant and 357 benign rows, totaling 446, with a malignant-to-benign ratio 89:357, close to 1:4. The noise condition flips a fixed count 39 of 398 training labels at uniformly drawn locations, before queries begin. This is fixed-count sampling without replacement, not independent Bernoulli corruption at exactly 10%. Test labels remain clean, and the stress conditions are derived transformations of one cohort.

4.5 | Evaluation Metrics

Test ROC AUC is evaluated at every budget using a single learner; QbC is described as affecting acquisition only. For fourteen equally spaced budgets, normalized trapezoidal AULC is the weighted expression below, with A_j the test AUC at budget 20+10j. Endpoint weights are one half, so it is not the simple arithmetic mean of fourteen AUCs. The threshold budget is the first observed crossing of test AUC 0.95, interpolated between adjacent evaluations and censored at 150 if absent. This is a retrospective descriptive estimate, not an operational stopping rule that could access unseen test labels. Reported bootstrap intervals describe seed resampling conditional on this dataset.

$$\mathrm{AULC} = \frac{10}{130}\left[\frac{A_0}{2} + \sum_{j=1,\ldots,12} A_j + \frac{A_{13}}{2}\right],$$

4.6 | Statistical Analysis Plan

Family A contains 36 paired Wilcoxon comparisons of AULC, active strategy versus random within each regime and learner; FamilyB contains 36 comparisons of learner-averaged test AUC at budgets 50,100 and 150 [12]. Holm correction is reported separately within each family [13]. Rank-biserial correlations describe paired differences. The ANOVA record additionally requires its observational unit and repeated-seed dependence to be specified; ordinary independent-error inference is not justified merely by fitting a two-factor model. Backfire is recorded at both cell-median and per-seed levels.

5 | Experimental Setup

The reported environment construction subsamples the imbalance condition before a stratified 30% test split, fixes any training-label corruption before querying, and draws the initial labeled set and pool from training only. StandardScaler is described as fitted to all training features once, including unlabeled pool features, and applied across rounds. This is a transductive use of unlabeled training covariates rather than test leakage or labeled-only preprocessing. Code assertions and split indices are needed to verify that test rows and labels did not influence fitting, querying, configuration selection or stopping.

The recorded design contains 90 seed-regime environments, each running five strategies and three learners from a common initial state. Fourteen evaluated budgets imply $90\times5\times3\times14 = 18{,}900$ evaluation fits. Thirteen QbC rounds with five members and three learners imply $90\times3\times13\times5 = 17{,}550$ committee fits. The source reports base seed 42, PCG64 seed sequences [42,seed,regimeindex], minimum observed minority count 4, zero split-assertion failures and mean corruption rate 0.0980. These execution diagnostics are reported rather than independently reproduced.

The execution record reports 324s on a workstation CPU with two workers, an environment phase 322s and mean 7.16s per environment, with figure generation 1.4s. It lists Python 3.12.14, NumPy 2.1.3 [14], SciPy 1.14.1 [15], scikit-learn 1.5.2 [10], statsmodels 0.14.5 [16] and Matplotlib 3.9.2 [17]. Hardware details, wall-time accounting and a versioned runnable artifact are required before those performance and reproduction claims can be verified.

Table 1. Datasets, regimes, and experimental protocol.

Component	Specification
Dataset	UCI WDBC via sklearn.datasets.load_breast_cancer (real): 569 × 30; 212 malignant / 357 benign [4,11].
Regime: Original	Real data unmodified.
Regime: Imbalance 1:4	Malignant subsampled to 89 rows (all 357 benign kept); subsample redrawn each seed; derived variant.
Regime: Label noise 10%	39 of 398 train labels flipped (uniform, fixed per seed before querying; test labels clean); realized rate 0.0980; derived variant.
Split	Stratified 30% test, drawn and frozen before any querying; pool = train-split rows only.
Initial labeled set	20 stratified rows (≥ 3 per class; minimum observed minority count 4), shared by all strategies (paired design).
Query protocol	Batch mode, top-10 scoring instances per round; budgets 20–150 step 10 (14 points).
Strategies	Random; uncertainty; margin; entropy; QbC (5-member bagging committee, vote disagreement $4v(1-v)$).
Learners	LogisticRegression(max_iter=1000) on standardized features; RandomForest(n_estimators=30); HistGradientBoosting(max_iter=50); model random_state = 42 fixed.
Replications	30 seeds; base seed 42 (PCG64 sequence [42, seed, regime index]).
Metrics	Test AUC per budget; AULC (normalized trapezoidal area over budgets 20–150); labels to AUC 0.95 (interpolated, censored at 150).
Statistics	Paired Wilcoxon vs random: family A (AULC; 36 tests) and family B (learner-averaged AUC at budgets 50/100/150; 36 tests), each Holm-corrected; two-way ANOVA with partial η^2; bootstrap 95% CIs (2,000 resamples).

The two stress regimes are derived transformations of the real WDBC data, documented in Section 4.4.

6 | Results

6.1 | Learning Curves and Overall Performance

Fig. 1 displays the recorded original-regime logistic-regression curves. Uncertainty, margin and entropy coincide in this panel, whereas their equivalence need not carry through inconsistent fitting or tie handling in another implementation. Random reaches reported mean AUC 0.9926 at 150 labels, uncertainty 0.9937 and QbC 0.9939; all start from a shared 0.985 at 20 labels. *Table 2* retains the full AULC summaries. Their ordering describes these configurations on WDBC, conditional on the metric integration and acquisition code.

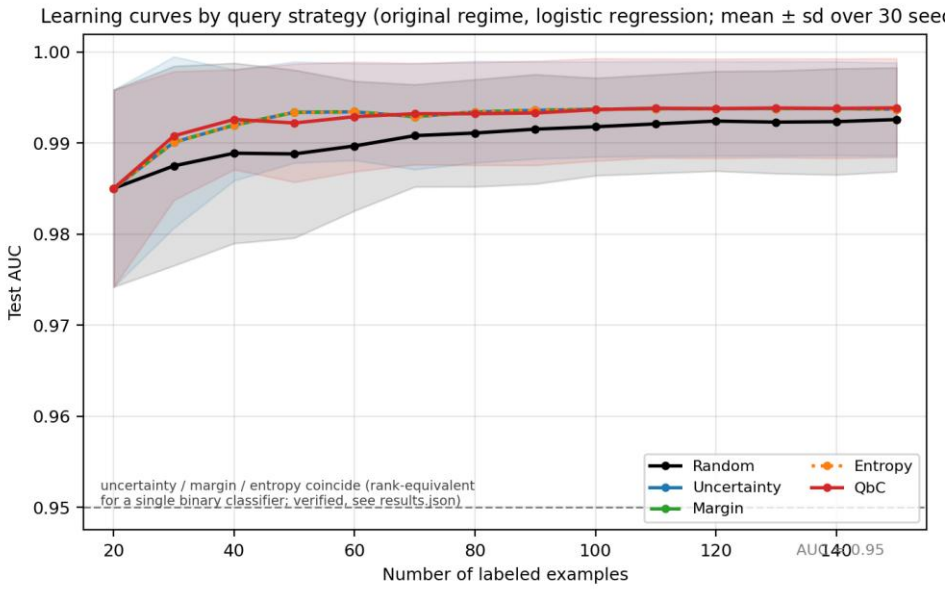


Fig. 1. Recorded test AUC learning curves for the original regime with logistic regression, summarized over 30 seeds.

Three single-model curves coincide in this panel. Original title/threshold-label placement requires the source plotting file for correction; the scientific bitmap is retained unchanged.

The recorded histogram-booster scores have AUC 0.5 at budgets 20 and 30, consistent with the described minimum leaf size. This floor depresses its AULC and influences subsequent acquisition histories; sharing the floor does not make its effect on strategy comparison negligible. Cross-learner AULC differences therefore reflect both learner configuration and acquisition.

Table 2. Reported AULC summaries by regime, strategy and learner, as mean±standard deviation over 30 seeds.

Regime	Strategy	Logistic Regression	Random Forest	Gradient Boosting
Original	Random	0.9906 ± 0.0061	0.9830 ± 0.0076	0.9243 ± 0.0082
	Uncertainty	0.9928 ± 0.0056	0.9884 ± 0.0060	0.9283 ± 0.0060
	Margin	0.9928 ± 0.0056	0.9881 ± 0.0063	0.9284 ± 0.0058
	Entropy	0.9928 ± 0.0056	0.9883 ± 0.0067	0.9277 ± 0.0073
	QbC	0.9928 ± 0.0055	0.9888 ± 0.0060	0.9274 ± 0.0059
Imbalance 1:4	Random	0.9904 ± 0.0100	0.9818 ± 0.0117	0.9173 ± 0.0160
	Uncertainty	0.9932 ± 0.0082	0.9877 ± 0.0100	0.9241 ± 0.0118
	Margin	0.9932 ± 0.0082	0.9869 ± 0.0113	0.9241 ± 0.0114
	Entropy	0.9932 ± 0.0082	0.9871 ± 0.0109	0.9234 ± 0.0117
	QbC	0.9936 ± 0.0076	0.9878 ± 0.0094	0.9251 ± 0.0107
Label noise 10%	Random	0.9587 ± 0.0204	0.9744 ± 0.0101	0.9103 ± 0.0088
	Uncertainty	0.9837 ± 0.0095	0.9849 ± 0.0070	0.9203 ± 0.0106
	Margin	0.9837 ± 0.0095	0.9845 ± 0.0084	0.9221 ± 0.0084
	Entropy	0.9837 ± 0.0095	0.9841 ± 0.0089	0.9222 ± 0.0085
	QbC	0.9817 ± 0.0116	0.9836 ± 0.0079	0.9218 ± 0.0081

The intended metric is normalized trapezoidal area across budgets 20-150; its implementation must be reconciled with the source description as a mean.

6.2 | Regime Effects

Fig. 2 maps AULC by strategy and regime for each learner. The regime effect dwarfs the strategy effect. Moving from the original to the noisy regime costs logistic regression 0.032 AULC under random sampling (0.9906 → 0.9587), versus 0.009 for random forest and 0.014 for gradient boosting; the imbalance regime is milder (0.0002, 0.0012, and 0.0071 respectively under random sampling). The interaction is visible in the same figure: the strategy contrast widens in the noisy regime—precisely where the baseline weakens, the active strategies recover the most (logistic-regression AULC 0.9837 under uncertainty versus 0.9587 under random),

narrowing the regime-induced gap to -0.0069 relative to the original-regime random baseline. This is the empirical basis for the interaction term in the ANOVA of Section 7.

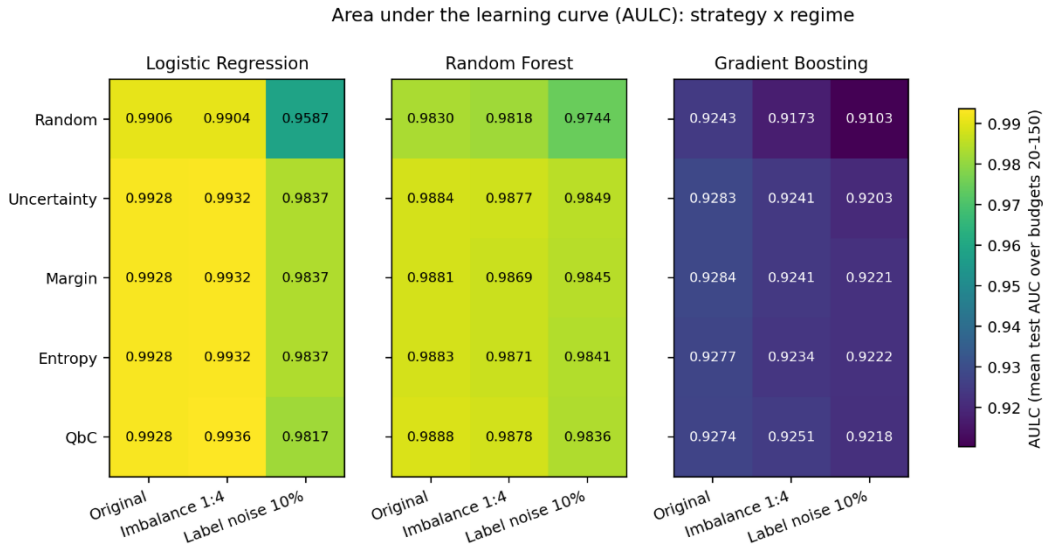


Fig. 2. AULC by strategy (rows) and regime (columns) for each learner, with numerical annotations on a common color scale.

The label-noise regime depresses every learner under random sampling; the active strategies recover a substantial share of the loss, most visibly for logistic regression.

6.3 | Label-Budget Efficiency

Fig. 3 and *Table 3* summarize retrospective threshold budgets. All recorded cells have reach rate 1.00 by 150 labels. Original-regime pooled means are 27.1 for random and 26.8 for uncertainty, with medians 20. Under imbalance the means are 31.2 and 28.7; under noise, random 35.4 and margin 30.6. Logistic regression under noise changes from 32.0 to 24.7, approximately 23% fewer labels using the rounded means. Gradient-boosting noise medians are 47.7 for random and 39.8 for margin. Pooling three learners yields 90 trajectories per cell but does not create 90 independent datasets.

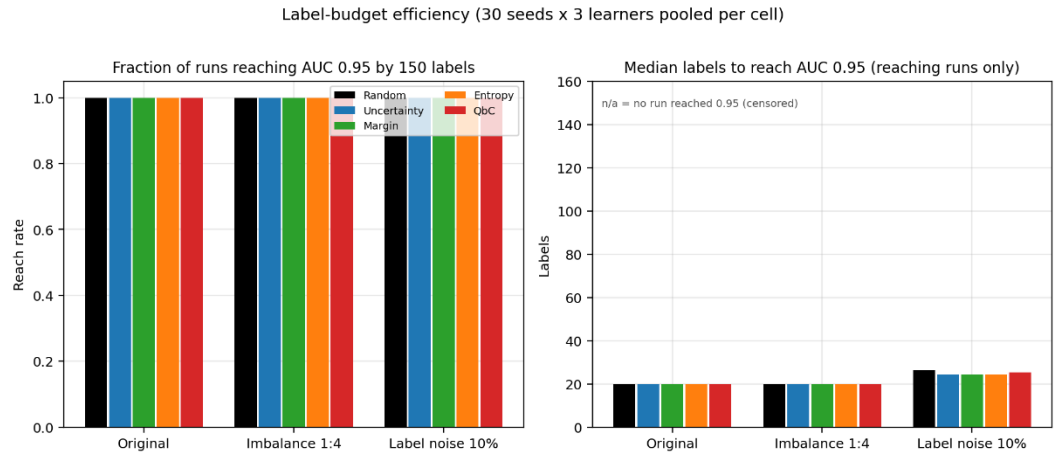


Fig. 3. Label-budget efficiency, pooled over the three learners (30 seeds × 3 learners per cell). Left: fraction of runs reaching AUC 0.95 within 150 labels (all cells reach 1.00). Right: median labels to AUC 0.95 among reaching runs. The medians saturate at the initial budget of 20 in the easier regimes (see Table 3 for means).

Table 3. Labels required to reach test AUC 0.95, pooled over learners (30 seeds × 3 learners = 90 runs per cell). Reach rate is the fraction of runs attaining the threshold within 150 labels (censoring bound).

Regime	Strategy	Reach Rate	Median Labels	Mean Labels
Original	Random	1.00	20.0	27.1
	Uncertainty	1.00	20.0	26.8
	Margin	1.00	20.0	26.9
	Entropy	1.00	20.0	27.0
	QbC	1.00	20.0	27.0
Imbalance 1:4	Random	1.00	20.0	31.2
	Uncertainty	1.00	20.0	28.7
	Margin	1.00	20.0	28.8
	Entropy	1.00	20.0	29.1
	QbC	1.00	20.0	29.2
Label noise 10%	Random	1.00	26.4	35.4
	Uncertainty	1.00	24.4	31.1
	Margin	1.00	24.3	30.6
	Entropy	1.00	24.4	30.9
	QbC	1.00	25.4	32.0

The median saturates at the initial budget (20) wherever most curves already start above threshold.

6.4 | Committee Behavior

The example QbC batch contains 8 benign and 2 malignant instances from a pool 378. Recorded logistic-regression committee disagreement is 0.946 for selected instances versus 0.113 for the remaining pool initially, an approximate 8.4-fold ratio, and 0.017 versus 0.001 late. Noise-regime initial values are 0.960 and 0.253. In early histogram-booster rounds constant committee votes can make disagreement independent of the instance, effectively leaving tie-breaking to choose queries. These diagnostics describe the acquisition scores; they do not establish minority coverage or annotation quality.

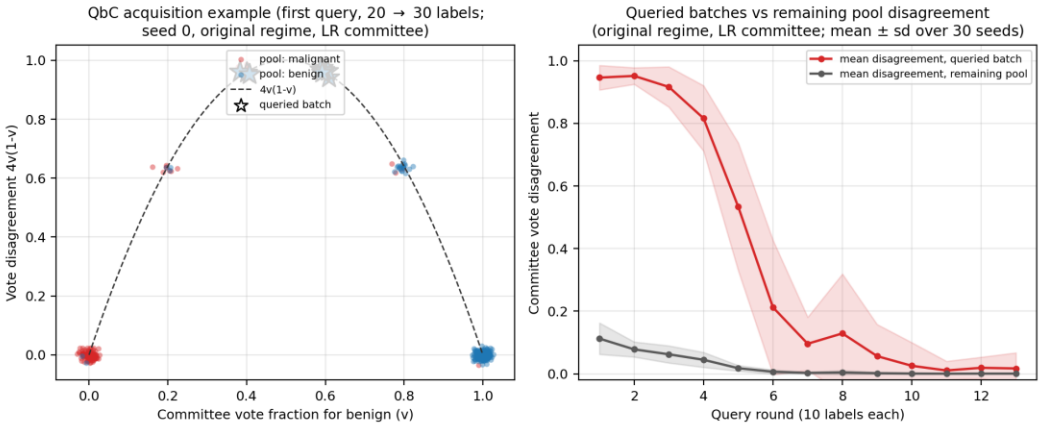


Fig. 4. Committee diagnostics for QbC. Left: committee vote fraction v versus vote disagreement 4v(1−v) across the pool at the first query (seed 0, original regime, logistic-regression committee; 20 labeled, pool 378); stars mark the 10 queried instances. Right: mean committee disagreement of the queried batch versus the remaining pool across the 13 query rounds (original regime, logistic regression; mean ± sd over 30 seeds).

6.5 | The Label-Noise Regime in Detail

For logistic regression under noise, reported QbC AULC 0.9817 is below the single-model value 0.9837. For gradient boosting, QbC 0.9218 exceeds uncertainty 0.9203 but is below margin 0.9221 and entropy 0.9222. For random forest, uncertainty 0.9849 exceeds QbC 0.9836. Thus the committee does not dominate all single-

model strategies in the noisy regime. Differences among exact-rank-equivalent scores require their fitting, tie and random-stream histories.

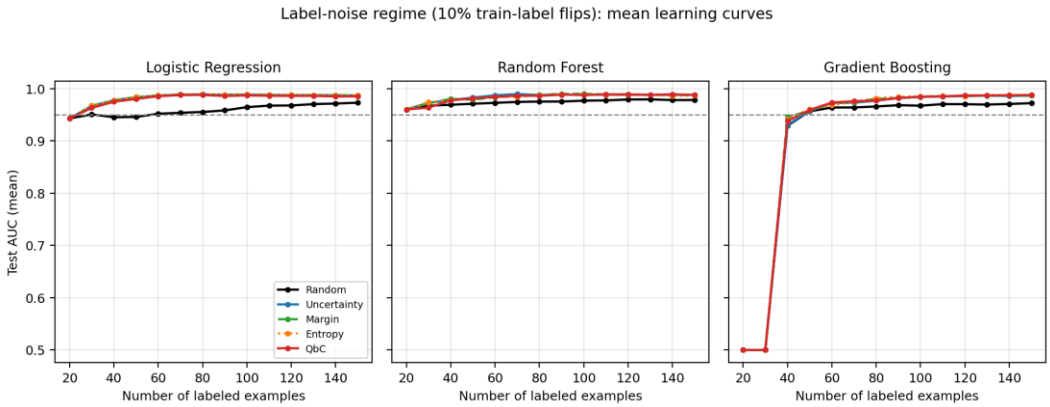


Fig. 5. Mean learning curves in the label-noise regime (10% of training labels flipped) for the three learners (30 seeds).

The gradient-boosting learner is at AUC 0.5 for budgets 20 and 30 (Section 4.3). The dashed line marks AUC 0.95.

6.6 | Backfire Analysis

No reported cell has a negative median AULC delta, but some individual seeds are worse than random. The record gives 26.7% such seeds for uncertainty and margin with gradient boosting in the original regime, with conditional deficit−0.0018. The deepest reported conditional deficit is−0.0093 for uncertainty with gradient boosting under noise, occurring in 10.0% of seeds. These are dataset-conditional outcomes rather than a universal downside bound.

7 | Statistical Analysis

Table 4 retains Family A statistics:36 reported Holm-significant comparisons, largest adjusted p0.0039 and smallest 6.7×10^{-8}, with rank-biserial range 0.65-1.00. FamilyB is reported as 36 significant comparisons, largest adjusted p0.0035 and rank-biserial range 0.60-1.00. The noise/budget 100 examples give median AUC deltas 0.0155 for uncertainty and 0.0160 for QbC, W 0 and adjusted p 6.7×10^{-8}. Raw paired trajectories, zero handling, sidedness and the correction code are required to validate these recorded inferential summaries.

Table 4. Family A—paired Wilcoxon signed-rank tests of AULC (strategy vs random sampling, paired by seed; n = 30 pairs, zero differences excluded), Holm-corrected within the family of 36 tests.

Regime	Strategy	W (LR)	p (LR)	W (RF)	p (RF)	W (GB)	p (GB)
Original	Uncertainty	30	8.3×10^{-5}	0	6.7×10^{-8}	75	3.9×10^{-3}
	Margin	30	8.3×10^{-5}	18	1.2×10^{-5}	72	3.9×10^{-3}
	Entropy	30	8.3×10^{-5}	18	1.2×10^{-5}	79	3.9×10^{-3}
	QbC	47	5.5×10^{-4}	7	1.0×10^{-6}	81	3.9×10^{-3}
Imbalance 1:4	Uncertainty	72	3.9×10^{-3}	58	1.5×10^{-3}	52	9.2×10^{-4}
	Margin	72	3.9×10^{-3}	64	2.1×10^{-3}	42	3.5×10^{-4}
	Entropy	72	3.9×10^{-3}	59	1.5×10^{-3}	58	1.5×10^{-3}
	QbC	37	1.1×10^{-3}	36	1.8×10^{-4}	44	4.3×10^{-4}
Label noise 10%	Uncertainty	0	6.7×10^{-8}	1	1.2×10^{-7}	46	5.2×10^{-4}
	Margin	0	6.7×10^{-8}	0	6.7×10^{-8}	37	1.9×10^{-4}
	Entropy	0	6.7×10^{-8}	15	7.1×10^{-6}	24	3.3×10^{-5}
	QbC	4	3.9×10^{-7}	21	2.0×10^{-5}	17	1.0×10^{-5}

All tests favor the active strategy (positive median deltas; backfire analysis in Section 6.6). LR = logistic regression, RF = random forest, GB = gradient boosting.

The recorded ANOVA summary gives regime F(2,435)=53.60 with partial$\eta^2$0.198, strategy F(4,435)=18.23 with partial$\eta^2$0.144, and interaction F(8,435)=3.76 with partial$\eta^2$0.065. These effect sizes agree with F times effect degrees of freedom divided by that quantity plus 435; they are partial associations, not percentages of total variance causally explained. A 450-observation design matches one learner across 30 seeds, three regimes and five strategies; pooling all learners would instead contain 1,350 observations unless an aggregation is specified. The reported collapsed-strategy summary F(2,261)=28.83 is similarly conditional on its unit and model. Original per-seed outputs and analysis code must resolve the model labeling and repeated-measures inference. Reported noise intervals are[0.9803,0.9869] for logistic-regression uncertainty and[0.9513,0.9655] for random. The mean ordering is compatible with H1 and opposite to H2, but no preregistration or causal interpretation is established.

8 | Discussion

The recorded original-regime logistic-regression baseline is near the upper end of the AUC scale, so remaining absolute improvements are small. Random sampling reaches reported AUC 0.987 at 30 labels, while several cells have AULC gains only in the thousandths. Some of these cells also have adverse individual seeds. This co-occurrence is descriptive: the study does not identify a causal relationship between the ceiling, variance of paired differences and the probability of underperforming random.

The largest recorded strategy premium occurs under the fixed-count noise condition, particularly for logistic regression. A plausible interpretation is that acquisition can provide more informative labeled examples for that learner. It is not established that adaptively selected and random batches experience the same conditional corruption rate: although locations were drawn uniformly before querying, past corrupted labels influence the evolving model and its future selections. Query-level corruption counts and controlled alternatives are needed to identify the mechanism.

Under imbalance, QbC has the largest reported mean AULC for each learner. This ordering does not establish that it preferentially acquires minority examples or is clinically the safest default. Under the studied noise process uncertainty outperforms random in the reported summaries, but this cannot be generalized to difficulty-correlated, class-conditional or adversarial annotation error [1].

The small-sample boosting floor and saturated threshold medians limit interpretation. A common initial floor affects subsequent query histories as well as the integral. The interpolated test threshold is retrospective and must not be used to choose a deployed stopping budget without independent validation. These facts matter when comparing learners as well as strategies.

8.1 | Implications

A pilot labeling study can test whether an acquisition rule is worth its scoring cost, but a deployable choice needs representative validation data, annotation-cost accounting and a realistic error model. The reported WDBC observations motivate comparison rather than a universal preference for active sampling whenever labels are unreliable. Binary uncertainty, margin and entropy should select the same ranks for a shared fitted model and tie rule; multiclass models and distinct committees need not share that equivalence.

8.2 | Limitations

The study uses one dataset and two separate derived stress conditions. The fixed-count uniform corruption rate is approximately 9.8%, not a model of independent repeated rater errors. Batch size 10, committee size 5, learner configurations and the retrospective 0.95 threshold are fixed. AULC is a trapezoidal time average, not an annotation utility function. Overlapping patient records across seeds, correlated strategies and repeated budgets limit inferential independence. No external clinical cohort, real annotation costs or rater behavior is evaluated.

8.3 | Reproducibility and Data Availability

WDBC is publicly available [4], and the imbalance and fixed-count corruption transformations are specified in the manuscript. The execution record names code/experiment.py, results/results.json and dataset_information.md, but no retrievable release or those files accompany the supplied manuscript. Their availability and exact reproduction of the reported values therefore remain unverified. Validation requires the versioned script, environment, seed scheme, splits, noise indices, queried indices, per-budget predictions and raw statistics, together with any archived registration record.

9 | Conclusion

The recorded summaries favor active acquisition over random at the cell level, with the largest logistic-regression gain under fixed-count noise. Individual seeds can nevertheless perform worse than random, and QbC does not dominate every noisy-regime comparison. The results concern the supplied single-dataset configurations and require original execution evidence to validate AULC, exact score-equivalence handling and inferential units. Extending the study to external cohorts, simultaneous stress regimes and realistic annotation mechanisms would be necessary before using it to select clinical labeling policies.

Declarations

The study uses the public WDBC dataset [4] and derived class-subsampling and label-corruption conditions. No new primary data collection is described.

The computational protocol describes no patient contact or new human-participant intervention. Institutional requirements and any secondary-data approval statement must be confirmed by the author.

The source declares no funding and no competing interests. The author must confirm these declarations before publication.

The recorded seed scheme is documented in Section 11. Availability and numerical reproducibility of scripts, results and figures require an accessible versioned release.

The source declares AI assistance in text and code preparation. The author must identify the tools and tasks and confirm the independent validation of the scientific content; no execution or verification of the reported experiments follows from the disclosure itself.

References

[1] Settles, B. (2009). *Active learning literature survey*. https://burrsettles.com/pub/settles.activelearning.pdf

[2] Lewis, D. D. (1995). A sequential algorithm for training text classifiers: corrigendum and additional data. *ACM sigir forum* (Vol. 29, pp. 13–19). Association for Computing Machinery. https://doi.org/10.1007/978-1-4471-2099-5_1

[3] Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. *Proceedings of the fifth annual workshop on computational learning theory* (pp. 287–294). Association for Computing Machinery. https://doi.org/10.1145/130385.130417

[4] Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). *Breast cancer wisconsin (diagnostic).* https://doi.org/10.24432/C5DW2B

[5] Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of machine learning research*, *2*(11), 45–66. https://jmlr.org/papers/v2/tong01a.html

[6] Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine learning*, *28*(2), 133–168. https://doi.org/10.1023/A:1007330508534

[7] McCallum, A. K., & Nigam, K. (1998). Employing em and pool-based active learning for text classification. *ICML* (Vol. 98, pp. 350–358). UMass Amheret. https://people.cs.umass.edu/~mccallum/papers/active-em-icml98.pdf

[8] Dasgupta, S., & Hsu, D. (2008). Hierarchical sampling for active learning. *Proceedings of the 25th international conference on machine learning* (pp. 208–215). Association for Computing Machinery. https://doi.org/10.1145/1390156.1390183

[9] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[10] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, *29*(5), 1189–1232. https://www.jstor.org/stable/2699986

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: Machine learning in Python. *The journal of machine learning research*, *12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[12] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

[13] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733

[14] Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., & Cournapeau, D. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[15] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., & Cournapeau, D. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[16] Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with python. *Scipy*, *7*(1), 92–96. https://doi.org/10.25080/Majora-92bf1922-011

[17] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55