Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 06, No. 02, July-Dec 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.62.1.14

Research Paper

When uncertainty sampling backfires: A benchmark of active learning query strategies on real image datasets

Kelvin Lleins Rojas-Córdova[1*], Monica Gabriela Leveau-Paredes[2], Jhosep Sánchez-Flores[3]
[1*,3]Escuela Profesional de Ingeniería de Sistemas e Informática, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de San Martín, Tarapoto, Peru.
[2]Universidad Nacional de San Martín, Tarapoto, Peru.

Article Info

Article History:
Received: 17 June 2026
Revised: 16 August 2026
Accepted: 09 September 2026
Published: 22 October 2026

Keywords:
Active Learning
Uncertainty Sampling
Core-Set Selection
Annotation Cost
Image Classification

ABSTRACT

Active learning promises to reduce annotation cost by letting a model choose which unlabelled examples should be labelled next, but published gains vary widely and some strategies have been reported to perform worse than random selection. This study compares six query strategies under identical conditions on three real image datasets: handwritten digits from MNIST, product photographs from Fashion-MNIST and the smaller UCI Digits collection. Starting from 20 randomly chosen labelled images, a logistic-regression classifier was retrained after each batch of queries until 1,000 images (400 for Digits) had been labelled, with five random seeds per condition. The strategies were random sampling, least-confidence, margin and entropy uncertainty sampling, k-center greedy diversity sampling, and margin sampling combined with k-means diversification of each batch. Margin sampling with batch diversification was the most label-efficient strategy on all three datasets: it reached 95% of the accuracy of a model trained on the full pool with 372, 452 and 82 labels on average, whereas random sampling needed 750 and 216 labels on MNIST and Digits and reached the target in only one of five runs on Fashion-MNIST (at 820 labels), and its area under the learning curve exceeded that of random sampling in all 15 runs. However, least-confidence and entropy sampling were worse than random sampling throughout the budget on MNIST and Fashion-MNIST, and k-center greedy sampling ended 16.3 and 9.9 percentage points below random sampling on these datasets. Inspection of the images queried by each strategy explained these failures: entropy sampling repeatedly selected visually similar ambiguous digits from a few classes, whereas k-center greedy selected atypical images, whose distance to their tenth-nearest neighbour was 1.38 and 1.82 times the pool median, and over-sampled particular classes. On the small Digits dataset all strategies outperformed random sampling. The value of active learning therefore depends strongly on the strategy and the data, and margin-based sampling with batch diversity is a robust default, whereas entropy-based and purely geometric diversity sampling should be validated before use.

Check for updates

Journal of Artificial Intelligence, Machine Learning and Neural

Corresponding Author:
Kelvin Lleins Rojas-Córdova
Escuela Profesional de Ingeniería de Sistemas e Informática, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de San Martín, Tarapoto, Peru.
Email: klrojasc@alumno.unsm.edu.pe

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Supervised image classifiers require labelled examples, and labelling is often the most expensive part of building them. In medical imaging, remote sensing, industrial inspection and document digitisation, unlabelled images are plentiful but each label may require expert time. Active learning addresses this imbalance by letting the learner choose which examples should be labelled next, in the expectation that a well-chosen subset will teach the model as much as a much larger random sample [1], [2], [3]. The most widely used family of methods, uncertainty sampling, queries the examples about which the current model is least certain [4], [5]; diversity-based methods, such as core-set selection [6], instead seek examples that cover the input space; and hybrid methods combine the two [7], [8].

The empirical evidence for these methods is, however, mixed. Several studies have reported that active learning offers little or no benefit over random sampling in realistic settings [9], [10], [11], that the best strategy depends on the labelling budget [12], and that uncertainty sampling can be attracted to outliers and unlearnable examples [13]. For logistic regression, benchmarks have found that no strategy dominates and that random sampling is a strong baseline [14], [15]. Reported gains are also sensitive to implementation details, random seeds and evaluation protocols [11], which makes it difficult for practitioners to know what to expect. This paper provides a controlled, reproducible comparison of six common query strategies on three real image datasets, and complements the quantitative results with a visual inspection of the images each strategy chooses to label. The contributions are as follows: A benchmark of random sampling, three uncertainty-sampling criteria, k-center greedy diversity sampling and a margin-plus-diversity hybrid on MNIST, Fashion-MNIST and UCI Digits, with learning curves over the full labelling budget and five random seeds per condition.

Evidence that margin sampling with batch diversification was the most label-efficient strategy on all three datasets, reaching 95% of full-pool accuracy with 50% fewer labels than random sampling on MNIST and 62% fewer on Digits, and in every run on Fashion-MNIST, where random sampling rarely reached the target. Evidence that entropy and least-confidence sampling were worse than random sampling on the two larger datasets and that k-center greedy sampling was substantially worse, together with an explanation based on the real images that these strategies select. A quantitative analysis of the atypicality and class composition of queried images, and practical recommendations for choosing a query strategy. Section 2 reviews related work. Section 3 describes the datasets, query strategies and evaluation protocol. Section 4 reports the learning curves, label savings and characteristics of the queried images. Section 5 discusses the findings and their limitations, and Section 6 concludes the paper.

2. RELATED WORK

2.1 Uncertainty Sampling

Lewis and Gale [4] introduced uncertainty sampling for text classification, querying the documents whose predicted class probability was closest to 0.5. For multiclass problems, three criteria are standard

[2]: least confidence, which queries examples whose most probable class has the lowest probability; margin sampling [5], which queries examples with the smallest difference between the two most probable classes; and entropy sampling, which queries examples with the highest Shannon entropy [16] of the predicted class distribution. Tong and Koller [17] developed margin-based querying for support vector machines. Uncertainty sampling is simple and cheap, but it relies on the calibration of a model trained on few labels and can concentrate queries in ambiguous regions or on outliers [13].

2.2 Diversity and Hybrid Strategies

When labels are requested in batches, selecting the most uncertain examples independently tends to produce redundant batches of near-duplicates. Sener and Savarese [6] framed batch selection as a core-set problem and used the greedy 2-approximation to the k-center problem [18], choosing examples that minimise the maximum distance from any pool example to its nearest labelled example. Zhdanov [7] proposed to pre-select the most uncertain examples and then cluster them with k-means [19], querying one representative per cluster, and Ash et al. [8] combined uncertainty and diversity in the gradient embedding space of deep networks. Gal et al. [20] applied Bayesian uncertainty estimates to image data, and Ren et al. [21] surveyed deep active learning.

2.3 Critical Evaluations

Schein and Ungar [14] found that, for logistic regression, several popular strategies often failed to beat random sampling. Yang and Loog [15] benchmarked active learning for logistic regression on many datasets and found that random sampling was a strong baseline and that margin-based methods were among the more reliable. Lowell et al. [9] showed that samples acquired with one model often transfer poorly to another, Mittal et al. [10] and Munjal et al. [11] found that gains in deep active learning shrink under careful evaluation, and Hacohen et al. [12] showed that typicality-based sampling is preferable at very low budgets whereas uncertainty sampling is preferable at higher budgets. Attenberg and Provost [22] discussed the limits of active learning under class imbalance. These studies motivate a transparent comparison that reports full learning curves and inspects what each strategy actually selects.

3. MEATHODOLOGY

Figure 1 illustrates the pool-based active learning loop used in this study. A classifier is trained on the currently labelled images, a query strategy selects a batch of unlabelled images from the pool, an oracle provides their labels, and the process repeats until the labelling budget is exhausted. In this simulation the oracle is the ground-truth label of the dataset.

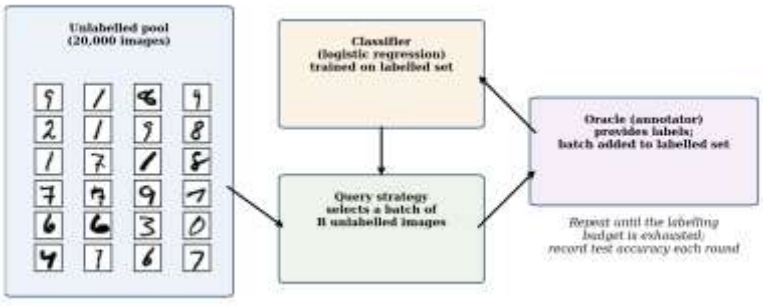


Figure 1. The Pool-Based Active Learning Loop, Illustrated with Real MNIST Images from the Unlabelled Pool

3.1 Datasets

Three real image datasets were used Table 1 and Figure 2. MNIST [23] contains scanned handwritten digits, and Fashion-MNIST [24] contains product photographs of ten categories of clothing and

accessories; both consist of 28 × 28 grey-scale images. For each, a fixed random pool of 20,000 images was drawn from the official training set and a fixed test set of 5,000 images from the official test set. The UCI optical recognition of handwritten digits dataset [25], distributed with scikit-learn [26], contains 1,797 8 × 8 images; 497 images were held out for testing and the remaining 1,300 formed the pool. Pixel intensities were scaled to [0, 1]. As a reference, a logistic-regression classifier trained on the entire pool reached test accuracies of 91.9%, 84.3% and 96.8%, respectively; the label-efficiency target was set to 95% of these values.

Table 1. Datasets and Reference Accuracy

Dataset	Image Size	Classes	Pool Size	Test Size	Initial Labels / Batch / Budget	Full-Pool Accuracy (%)	Target: 95% of Full-Pool (%)
MNIST [23]	28 × 28	10	20,000	5,000	20 / 20 / 1,000	91.92	87.32
Fashion-MNIST [24]	28 × 28	10	20,000	5,000	20 / 20 / 1,000	84.26	80.05
Digits [25]	8 × 8	10	1,300	497	20 / 10 / 400	96.78	91.94

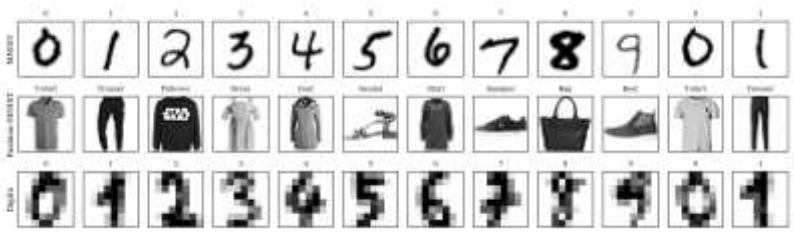


Figure 2. Examples of the Real Images in the Three Datasets (Inverted Grey Scale). Top: MNIST Handwritten Digits; Middle: Fashion-MNIST Product Photographs; Bottom: UCI Digits at 8 × 8 Resolution

3.2 Classifier

A multinomial logistic-regression classifier [27] with L2 regularisation (C = 1) on raw pixel intensities was used throughout, retrained from scratch after every batch. Logistic regression was chosen because it is fast enough to retrain hundreds of times per run, produces probability estimates for uncertainty sampling, and has been the subject of previous active learning benchmarks [14], [15], which allows direct comparison. When a class was not yet represented in the labelled set, its predicted probability was set to zero.

3.3 Query Strategies

Let p(y | x) denote the class probabilities predicted by the current model for an unlabelled image x, and let p [1](x) ≥ p [2](x) be the two largest probabilities. Six strategies were compared. Random sampling selects B images uniformly at random from the pool. Least confidence selects the B images with the largest value of [1] Margin sampling [5] selects the B images with the smallest difference between the two most probable classes, that is, the largest value of [2] Entropy sampling [2], [16] selects the B images with the highest predictive entropy:

[3] k-Center greedy sampling [6], [18] ignores the model and selects, one at a time, the pool image whose Euclidean distance to its nearest labelled (or already selected) image is largest, thereby greedily minimising the covering radius of the labelled set:

[4] Where U and L denote the unlabelled and labelled sets. Finally, Margin + diversity [7] pre-selects the 10B images with the smallest margin, clusters them into B groups with k-means [19] and queries the image closest to each cluster centre, so that each batch contains uncertain but mutually dissimilar images. Table 2 summarises the strategies.

Table 2. Query Strategies Compared in This Study

Strategy	Type	Selection Rule	Uses Model?	Batch-Aware?
Random	Baseline	Uniform random sample of B pool images	No	No
Least confidence	Uncertainty	Largest 1 − p(1)(x)	Yes	No
Margin	Uncertainty	Smallest p(1)(x) − p(2)(x)	Yes	No
Entropy	Uncertainty	Largest predictive entropy	Yes	No
k-Center greedy	Diversity	Greedy farthest-point selection in pixel space	No	Yes
Margin + diversity	Hybrid	k-means over the 10B smallest-margin images; one image per cluster	Yes	Yes

3.4 Protocol and Evaluation

Each run started from 20 images drawn at random from the pool; the same initial set was used for all strategies within a seed, so that differences arise only from the query strategy. Batches of B = 20 images (10 for Digits) were queried until the labelled set contained 1,000 images (400 for Digits). After each batch, the classifier was retrained and evaluated on the fixed test set. Each condition was repeated with five seeds. Label efficiency was summarised by the normalised area under the learning curve (AULC), the mean test accuracy over the budget computed with the trapezoidal rule, and by the number of labels required to first reach 95% of full-pool accuracy:

[5] Strategies were compared with the Friedman test [28] on AULC over seeds, and pairwise with the number of seeds in which one strategy exceeded another [29]; with five seeds, the minimum attainable two-sided p-value of an exact paired test is 0.0625, so pairwise results are reported as win counts rather than significance tests. To characterise the queried images, the atypicality of each pool image was measured as the Euclidean distance to its tenth-nearest neighbour in the pool, normalised by the pool median, and averaged over the first 200 queried images of each run; the class composition of these queries was summarised by the share of the most frequent class. All code was written in Python with scikit-learn [26].

4. RESULTS AND DISCUSSION

4.1 Learning Curves

Figure 3 shows mean learning curves and Table 3 lists accuracy at three points of the labelling budget. The strategies differed significantly on all three datasets (Friedman test on AULC: $\chi^2 = 25.0$, $p = 1.4 \times 10{-}4$ for MNIST; $\chi^2 = 24.5$, $p = 1.7 \times 10{-}4$ for Fashion-MNIST; $\chi^2 = 23.1$, $p = 3.3 \times 10{-}4$ for Digits). On MNIST and Fashion-MNIST, only the two margin-based strategies outperformed random sampling. Margin + diversity led from the first batches onwards; with 100 labels it reached 80.1% on MNIST compared with

73.3% for random sampling, and at the end of the budget it reached 89.7% compared with 87.7%. Least-confidence and entropy sampling fell behind random sampling immediately after the initial set and remained behind for the whole budget, by up to 7.8 and 11.4 percentage points on MNIST and up to 6.5 and 13.5 points on Fashion-MNIST. k-Center greedy sampling improved very slowly and finished 16.3 points below random sampling on MNIST (71.4% versus 87.7%) and 9.9 points below on Fashion-MNIST.

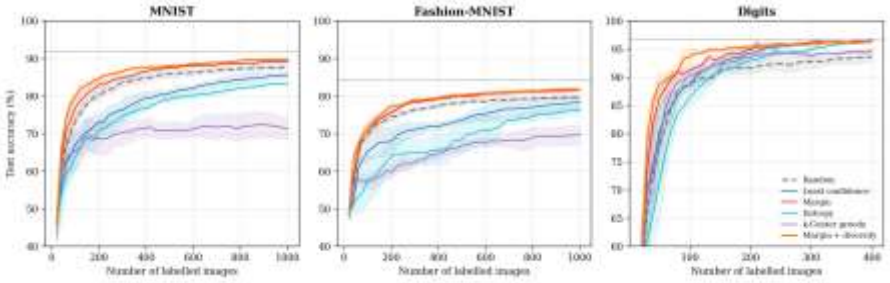


Figure 3. Mean Test Accuracy (± 1 SD over Five Seeds) as a Function of the Number of Labelled Images. The Dotted Line Marks the Accuracy of Logistic Regression Trained on the Full Pool; the Dashed Grey Curve is Random Sampling

Table 3. Test Accuracy (%) at Three Labelling Budgets (Mean ± SD over Five Seeds). the Best Value in Each Row is in Bold

Dataset	Labels	Random	Least Confidence	Margin	Entropy	K-Center Greedy	Margin + Diversity
MNIST	100	73.3 ± 1.9	67.2 ± 3.2	77.0 ± 2.7	63.8 ± 3.5	65.5 ± 5.3	80.1 ± 1.2
	300	83.7 ± 1.0	76.3 ± 2.7	85.0 ± 0.6	74.7 ± 1.8	70.2 ± 3.6	86.5 ± 0.4
	1,000	87.7 ± 0.5	85.6 ± 1.0	89.1 ± 0.2	83.2 ± 0.3	71.4 ± 2.8	89.7 ± 0.5
Fashion-MNIST	100	69.8 ± 2.0	65.4 ± 2.5	70.9 ± 2.1	57.9 ± 6.4	57.2 ± 1.8	71.2 ± 1.6
	300	76.4 ± 0.5	71.1 ± 1.6	78.4 ± 0.3	65.1 ± 3.6	62.4 ± 1.1	79.0 ± 0.5
	1,000	79.7 ± 0.3	78.4 ± 1.9	81.8 ± 0.5	76.3 ± 1.2	69.8 ± 2.4	81.8 ± 0.1
Digits	60	81.4 ± 6.6	83.3 ± 2.7	88.1 ± 2.1	76.8 ± 4.7	84.3 ± 2.1	89.4 ± 1.5
	150	91.0 ± 0.7	92.4 ± 0.6	93.4 ± 0.5	91.5 ± 0.8	92.2 ± 0.8	94.8 ± 0.2
	400	93.6 ± 0.8	96.4 ± 0.2	96.5 ± 0.2	96.6 ± 0.2	94.8 ± 0.2	96.3 ± 0.2

The small Digits dataset behaved differently. All five non-random strategies outperformed random sampling at the end of the budget, and least-confidence, margin and entropy sampling all reached about 96.5%, close to the full-pool accuracy of 96.8%, whereas random sampling reached 93.6%. Entropy sampling was nevertheless the worst strategy in the early phase (76.8% at 60 labels versus 81.4% for random sampling), and Margin + diversity again gave the best accuracy for most of the budget.

4.2 Label Savings

Table 4 summarizes label efficiency. Margin + diversity had the highest AULC on all three datasets and exceeded random sampling in all 15 runs, and plain margin sampling did so in all 15 runs as well. Measured by the labels required to reach 95% of full-pool accuracy, Margin + diversity needed on average 372 labels on MNIST, 452 on Fashion-MNIST and 82 on Digits, and reached the target in every run. Random sampling needed 750 labels on MNIST (reaching the target in four of five runs within the budget) and 216 labels on Digits, and reached the target in only one of five runs on Fashion-MNIST. The hybrid strategy therefore roughly halved the labelling effort on MNIST and reduced it by 62% on Digits. Least-confidence,

entropy and k-center greedy sampling failed to reach the target within 1,000 labels in almost every run on the two larger datasets.

Table 4. Label Efficiency: Normalised Area under the Learning Curve (AULC, %) and Labels Needed to First Reach 95% of Full-Pool Accuracy (Mean over the Runs That Reached it; Number of Runs in Brackets)

Strategy	MNIST AULC	MNIST Labels to Target	Fashion-MNIST AULC	Fashion-MNIST Labels to Target	Digits AULC	Digits Labels to Target
Random	83.2 ± 0.8	750 (4/5)	76.4 ± 0.5	820 (1/5)	89.3 ± 1.4	216 (5/5)
Least confidence	78.5 ± 1.6	– (0/5)	72.6 ± 1.5	660 (1/5)	91.5 ± 0.9	138 (5/5)
Margin	85.0 ± 0.5	472 (5/5)	78.0 ± 0.2	516 (5/5)	92.9 ± 0.5	108 (5/5)
Entropy	76.3 ± 1.0	– (0/5)	68.1 ± 2.6	– (0/5)	89.5 ± 0.9	166 (5/5)
k-Center greedy	70.1 ± 3.0	– (0/5)	65.0 ± 1.1	– (0/5)	90.8 ± 0.2	130 (5/5)
Margin + diversity	86.3 ± 0.3	372 (5/5)	78.6 ± 0.2	452 (5/5)	93.7 ± 0.5	82 (5/5)

4.3 What Do the Strategies Select?

Figures 4 and 5 show the first 16 images queried after the initial set by each strategy in the first seed, with their true labels. The images make the quantitative results intelligible. On MNIST, entropy sampling repeatedly selected the same kind of image: 14 of its first 16 queries were ambiguous 2s and 7s written in similar, unusual styles. Least-confidence sampling showed the same tendency to a lesser degree. Margin sampling also selected ambiguous digits but from more classes, and Margin + diversity produced batches that were both uncertain and visibly varied in class and style. k-Center greedy sampling, in contrast, selected images that were unusual in pixel space: very thick, heavily inked or unusually large digits, which lie far from other images precisely because they are atypical and are therefore poor representatives of their classes.



Figure 4. The First 16 MNIST Images Queried After the Initial Labelled Set by Each Strategy (Seed 0), with Their True Labels. Entropy Sampling Concentrates on Similar Ambiguous 2s and 7s, Whereas K-Center Greedy Sampling Selects Atypical, Heavily Inked Digits

On Fashion-MNIST Figure 5, uncertainty-based strategies concentrated on the notoriously confusable upper-body garments, namely shirts, pullovers and coats, which dominated entropy sampling's queries. k-Center greedy sampling again selected extreme images: dark, high-contrast sandals, boots and bags with unusual shapes or strong patterns. Random sampling, by construction, reproduced the class distribution of the pool.

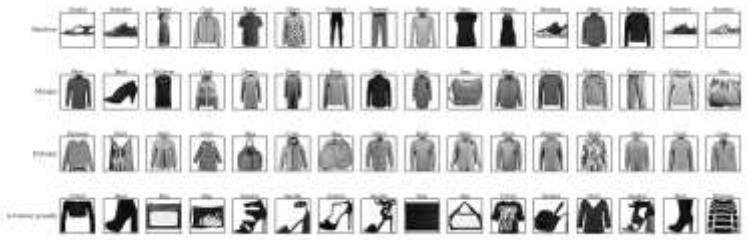


Figure 5. The First 16 Fashion-MNIST Images Queried by Four Strategies (Seed 0). Uncertainty-Based Strategies Focus on Confusable Tops (Shirt, Pullover, Coat), Whereas K-Center Greedy Sampling Selects Dark, Atypical Items

Table *5* and Figure 6 quantify these impressions over all seeds and the first 200 queries. Images selected by k-center greedy sampling had tenth-nearest-neighbour distances 1.38 times the pool median on MNIST and 1.82 times on Fashion-MNIST, far above those of random sampling (0.96 and 1.04), and its queries were the most imbalanced, with the most frequent class accounting for 20% and 31% of queries compared with about 14% and 13% under random sampling. Uncertainty-based strategies selected moderately atypical images (1.06–1.13 times the median) and were somewhat more imbalanced than random sampling on Fashion-MNIST (17–20% for the most frequent class). Margin + diversity combined uncertainty with near-typical images (1.02 on both large datasets) and balanced classes on MNIST (14%), which is consistent with its superior accuracy.

Table 5. Characteristics of the First 200 Queried Images (Mean over Five Seeds): Atypicality (Distance to the Tenth-Nearest Pool Neighbour Divided by the Pool Median) and Share of the Most Frequent Class

Strategy	**MNIST Atypicality**	**MNIST Largest Class Share**	**Fashion-MNIST Atypicality**	**Fashion-MNIST Largest Class Share**	**Digits Atypicality**	**Digits Largest Class Share**
Random	0.96	0.14	1.04	0.13	1.01	0.14
Least confidence	1.06	0.15	1.13	0.17	1.17	0.17
Margin	1.06	0.14	1.08	0.18	1.16	0.17
Entropy	1.06	0.15	1.09	0.20	1.16	0.17
k-Center greedy	1.38	0.20	1.82	0.31	1.19	0.18
Margin + diversity	1.02	0.14	1.02	0.16	1.11	0.18

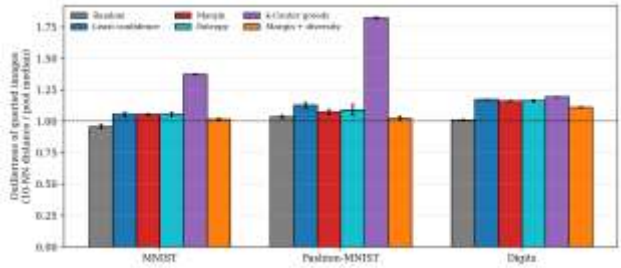


Figure 6. Atypicality of the First 200 Images Queried by Each Strategy (Distance to the Tenth-Nearest Neighbour in the Pool, Relative to the Pool Median; Mean ± SD over Five Seeds). Values above 1 Indicate That a Strategy Prefers Atypical Images

The Digits dataset explains why all strategies helped there. Its pool contains only 1,300 small images with little variation in style, so the most atypical images (atypicality 1.11–1.19 for all non-random strategies) are much less extreme than on the larger datasets, and uncertain examples are genuinely informative rather than merely ambiguous.

4.4 Effect of Batch Size

If the failure of entropy sampling were caused mainly by redundancy within a batch, it should diminish with smaller batches. To test this, the random, entropy, margin and Margin + diversity strategies were rerun on MNIST with batch sizes of 5 and 50 images (three seeds, budget of 520 labels) and compared with the batch size of 20 used above Table 6 and Figure 7. Entropy sampling remained clearly worse than random sampling at every batch size (AULC 71.5%, 72.2% and 72.4% versus 79.3%, 79.1% and 79.5%), even when only five images were queried per round. Its failure is therefore not primarily a consequence of redundant batches but of the kind of image that entropy prefers. Batch size mattered more for plain margin sampling, whose advantage over random sampling shrank from 2.9 to 1.4 percentage points of AULC as the batch grew from 5 to 50 images, whereas Margin + diversity was insensitive to batch size (83.0–83.6%). Diversification therefore becomes more valuable as batches grow, as expected [6], [7].

Table 6. Effect of Batch Size on MNIST: AULC up to 520 Labels and Accuracy at 520 Labels (%, Mean ± SD over Three Seeds)

Strategy	AULC, B = 5	AULC, B = 20	AULC, B = 50	Accuracy at 520, B = 5	Accuracy at 520, B = 20	Accuracy at 520, B = 50
Random	79.3 ± 0.3	79.1 ± 0.8	79.5 ± 0.7	85.6 ± 0.9	85.7 ± 1.4	86.0 ± 0.4
Entropy	71.5 ± 1.4	72.2 ± 1.5	72.4 ± 0.5	79.2 ± 1.3	79.6 ± 0.6	79.6 ± 1.4
Margin	82.2 ± 0.4	82.3 ± 0.7	80.9 ± 0.2	87.2 ± 0.2	87.4 ± 0.2	87.0 ± 0.5
Margin + diversity	83.0 ± 0.6	83.6 ± 0.5	83.0 ± 0.9	88.1 ± 0.1	88.3 ± 0.3	88.4 ± 0.6

Results for B = 20 are taken from the first three seeds of the main experiment, truncated at 520 labels.

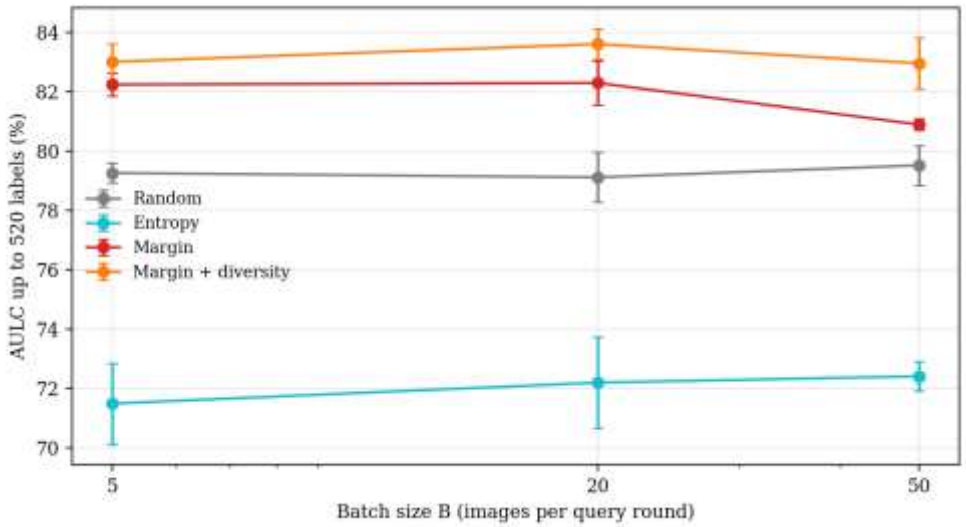


Figure 7. Area under the Learning Curve up to 520 Labels on MNIST as a Function of Batch Size (Mean ± SD over Three Seeds). Entropy Sampling Remains Below Random Sampling Even with Batches of Five Images

4.5 Discussion

4.5.1 Principal Findings

On three real image datasets, a hybrid of margin sampling and batch diversification was consistently the most label-efficient strategy, reaching 95% of full-pool accuracy with roughly half the labels needed by random sampling. At the same time, two of the most widely used strategies, entropy and least-confidence sampling, performed worse than random sampling on MNIST and Fashion-MNIST throughout the budget, and the core-set-inspired k-center greedy strategy performed far worse. These findings agree with the observations of Schein and Ungar [14] and Yang and Loog [15] that random sampling is a strong baseline for logistic regression and that margin-based methods are among the most reliable, and with the concerns about outlier-seeking behaviour raised by Karamcheti et al [13]. They also show that the value of active learning is not a property of the method alone: on the smaller, less variable Digits dataset, all strategies helped.

4.5.2 Why Some Strategies Backfire

The image galleries and the batch-size experiment point to two distinct failure mechanisms. The first concerns the kind of uncertainty that is rewarded. Entropy sampling scores examples by the spread of the whole predicted distribution and therefore favours images that are confusable with several classes at once, such as the atypically written 2s and 7s in Figure 4, whereas margin sampling targets images near the boundary between the two most likely classes. Because entropy sampling remained worse than random sampling even when only five images were queried per round (Section 4.4), its failure cannot be attributed mainly to redundancy within a batch; rather, the images it prefers are hard, ambiguous examples that carry little information about the typical appearance of each class, in line with the outlier-seeking behaviour described by Karamcheti et al [13]. Redundancy does matter for plain margin sampling at large batch sizes, which is why explicit diversification [7] was most useful with batches of 50. The second mechanism is atypicality in the input space. The k-center objective rewards covering the input space, and in raw pixel space the points farthest from the labelled set are unusual images: heavily inked digits and high-contrast garments. These images are poor examples of their classes, their selection skews the class distribution, and they leave the dense, typical regions that dominate the test set underrepresented. In the deep learning setting, k-center selection is usually applied in a learned feature space [6], where distances are more semantically meaningful; the present results show that it should not be applied naively in pixel space. The early-phase results also agree with the finding of Hacohen et al. [12] that typical examples are most valuable when very few labels are available.

4.5.3 Recommendations for Practice

Table 7. Practical Recommendations Derived from the Results

Situation	Recommendation	Evidence in This Study
Default choice for batch active learning	Use margin sampling with batch diversification (cluster the most uncertain candidates and query one per cluster)	Best AULC on all three datasets; beat random sampling in 15 of 15 runs
Entropy or least-confidence sampling planned	Validate against random sampling on a pilot budget before committing annotation resources	Worse than random on MNIST and Fashion-MNIST throughout the budget
Diversity-only (core-set) selection	Avoid in raw input space; use a meaningful feature space and check the atypicality of queries	Ended 9.9–16.3 points below random on the large datasets; queried atypical images
Monitoring an ongoing campaign	Inspect the queried images and their class composition	Redundant and atypical queries were visible immediately in the galleries
Small, homogeneous datasets	Most strategies help; margin-based sampling remains a safe choice	All strategies beat random sampling on Digits

Two further points apply to any deployment. First, the benefit of active learning should be judged against a random-sampling baseline under the same budget, because a strategy that underperforms random sampling wastes rather than saves annotation effort [9], [11]. Second, inspecting the queried images, as done in Figure 4 and Figure 5, is a cheap and effective way to detect redundancy or outlier-seeking behaviour early in a labelling campaign.

4.5.4 Limitations

The study used logistic regression on raw pixels; deep networks and learned feature spaces can change the behaviour of both uncertainty and diversity strategies, and the conclusions about k-center greedy sampling in particular apply to pixel space only. Three datasets of small grey-scale images were used; colour, high-resolution and imbalanced datasets may behave differently. The oracle was simulated with ground-truth labels, so annotation noise and varying annotation cost per image were not modelled; real ambiguous images, which uncertainty sampling favours, may be harder and more error-prone for human annotators [30]. Five seeds give stable estimates of mean behaviour but limited statistical power for pairwise comparisons, which were therefore reported as win counts. Finally, the batch-size analysis was limited to MNIST and three seeds.

5. CONCLUSION

This paper compared six active learning query strategies on three real image datasets with a common protocol and inspected the images each strategy chose to label. Margin sampling combined with batch diversification was the most label-efficient strategy on every dataset, reaching 95% of full-pool accuracy with 372 labels on MNIST, 452 on Fashion-MNIST and 82 on Digits, and outperforming random sampling in all runs. Entropy and least-confidence sampling were worse than random sampling on the two larger datasets, even with small batches, because they favoured hard, ambiguous images that poorly represent their classes, and k-center greedy selection in pixel space was substantially worse because it selected atypical images and skewed the class distribution. On the smaller, more homogeneous Digits

dataset all strategies were beneficial. Active learning can therefore halve labelling effort, but only with an appropriate strategy; margin-based sampling with batch diversity is a robust default, and any strategy should be validated against random sampling and monitored by inspecting the images it selects. Future work will extend the benchmark to deep networks with learned representations, colour and high-resolution images, imbalanced data and noisy human annotation.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Kelvin Lleins Rojas-Córdova	✓	✓	✓		✓				✓		✓			
Monica Gabriela Leveau-Paredes		✓		✓		✓		✓		✓		✓		
Jhosep Sánchez-Flores			✓				✓				✓		✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Acknowledgments

The authors thank the creators of the MNIST, Fashion-MNIST and UCI Digits datasets for making them publicly available, the developers of scikit-learn, and the researchers whose work is cited in this article.

Funding Information

This research has received no external funding.

Conflict of Interest Statement

No conflict of interest.

Informed Consent

Not Applicable.

Ethical Approval

Not Applicable.

Data Availability

This study draws on the publicly available MNIST, Fashion-MNIST and UCI Digits datasets. No new datasets were created. The analysis code is available from the corresponding author on reasonable request.

REFERENCES

[1] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," Journal of Artificial Intelligence Research, vol. 4, pp. 129–145, 1996. doi.org/10.1613/jair.295

[2] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. minds.wisc.edu/handle/1793/60660

[3] B. Settles, Active Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 6, no. 1. San Rafael, CA: Morgan & Claypool, 2012. doi.org/10.2200/S00429ED1V01Y201207AIM018

[4] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in Proc. 17th Annu. Int. ACM SIGIR Conf., 1994, pp. 3–12. doi.org/10.1007/978-1-4471-2099-5_1

[5] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in Advances in Intelligent Data Analysis, Lecture Notes in Computer Science, vol. 2189. Berlin: Springer, 2001, pp. 309–318. doi.org/10.1007/3-540-44816-0_31

[6] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in Proc. 6th Int. Conf. Learning Representations, 2018. doi.org/10.48550/arXiv.1708.00489

[7] F. Zhdanov, "Diverse mini-batch active learning," arXiv preprint arXiv:1901.05954, 2019. doi.org/10.48550/arXiv.1901.05954

[8] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in Proc. 8th Int. Conf. Learning Representations, 2020. doi.org/10.48550/arXiv.1906.03671

[9] D. Lowell, Z. C. Lipton, and B. C. Wallace, "Practical obstacles to deploying active learning," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing, 2019, pp. 21–30. doi.org/10.18653/v1/D19-1003

[10] S. Mittal, M. Tatarchenko, Ö. Çiçek, and T. Brox, "Parting with illusions about deep active learning," arXiv preprint arXiv:1912.05361, 2019. doi.org/10.48550/arXiv.1912.05361

[11] P. Munjal, N. Hayat, M. Hayat, J. Sourati, and S. Khan, "Towards robust and reproducible active learning using neural networks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2022, pp. 223–232. doi.org/10.1109/CVPR52688.2022.00032

[12] G. Hacohen, A. Dekel, and D. Weinshall, "Active learning on a budget: Opposite strategies suit high and low budgets," in Proc. 39th Int. Conf. Machine Learning, PMLR, vol. 162, 2022, pp. 8175–8195. proceedings.mlr.press/v162/hacohen22a.html

[13] S. Karamcheti, R. Krishna, L. Fei-Fei, and C. D. Manning, "Mind your outliers! Investigating the negative impact of outliers on active learning for visual question answering," in Proc. 59th Annu. Meeting of the Association for Computational Linguistics, 2021, pp. 7265–7281. doi.org/10.18653/v1/2021.acl-long.564

[14] A. I. Schein and L. H. Ungar, "Active learning for logistic regression: An evaluation," Machine Learning, vol. 68, no. 3, pp. 235–265, 2007. doi.org/10.1007/s10994-007-5019-5

[15] Y. Yang and M. Loog, "A benchmark and comparison of active learning for logistic regression," Pattern Recognition, vol. 83, pp. 401–415, 2018. doi.org/10.1016/j.patcog.2018.06.004

[16] C. E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948. doi.org/10.1002/j.1538-7305.1948.tb01338.x

[17] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," Journal of Machine Learning Research, vol. 2, pp. 45–66, 2001. jmlr.org/papers/v2/tong01a.html

[18] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," Theoretical Computer Science, vol. 38, pp. 293–306, 1985. doi.org/10.1016/0304-3975(85)90224-5

[19] S. P. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137, 1982. doi.org/10.1109/TIT.1982.1056489

[20] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in Proc. 34th Int. Conf. Machine Learning, PMLR, vol. 70, 2017, pp. 1183–1192. proceedings.mlr.press/v70/gal17a.html

[21] P. Ren et al., "A survey of deep active learning," ACM Computing Surveys, vol. 54, no. 9, p. 180, 2021. doi.org/10.1145/3472291

[22] J. Attenberg and F. Provost, "Why label when you can search? Alternatives to active learning for applying human resources to build classification models under extreme class imbalance," in Proc. 16th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2010, pp. 423–432. doi.org/10.1145/1835804.1835859

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998. doi.org/10.1109/5.726791

[24] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017. doi.org/10.48550/arXiv.1708.07747
[25] E. Alpaydin and C. Kaynak, "Optical recognition of handwritten digits," UCI Machine Learning Repository, 1998. archive.ics.uci.edu/dataset/80/optical+recognition+of+handwritten+digits
[26] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. jmlr.org/papers/v12/pedregosa11a.html
[27] D. R. Cox, "The regression analysis of binary sequences," Journal of the Royal Statistical Society: Series B (Methodological), vol. 20, no. 2, pp. 215–242, 1958. doi.org/10.1111/j.2517-6161.1958.tb00292.x
[28] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," Journal of the American Statistical Association, vol. 32, no. 200, pp. 675–701, 1937. doi.org/10.1080/01621459.1937.10503522
[29] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006. jmlr.org/papers/v7/demsar06a.html
[30] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," in Proc. NeurIPS Datasets and Benchmarks Track, 2021. doi.org/10.48550/arXiv.2103.14749

How to Cite: Kelvin Lleins Rojas-Córdova, Monica Gabriela Leveau-Paredes, Jhosep Sánchez-Flores. (2026). When uncertainty sampling backfires: A benchmark of active learning query strategies on real image datasets. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 6(2), 1-14. https://doi.org/10.55529/jaimlnn.62.1.14

BIOGRAPHIES OF AUTHORS

	**Kelvin Lleins Rojas-Córdova,** is a researcher in the Escuela Profesional de Ingeniería de Sistemas e Informática, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de San Martín, Tarapoto, Peru. His work is in computer vision and image segmentation, with an interest in how image classifiers behave outside the conditions they were trained on. He is concerned with the robustness of visual models to corruption and distribution shift and with the design of controlled benchmarking protocols. Email: klrojasc@alumno.unsm.edu.pe
	**Monica Gabriela Leveau-Paredes,** is a researcher at Universidad Nacional de San Martín, Tarapoto, Peru. Her interests lie in applied machine learning and information systems, with particular attention to the data side of model development: how examples are selected, annotated and curated, and what that costs in practice. She works on empirical studies of learning algorithms on real image data within her faculty research group.
	**Jhosep Sánchez-Flores,** is a researcher in the Escuela Profesional de Ingeniería de Sistemas e Informática, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de San Martín, Tarapoto, Peru. His interests lie in applied machine learning and systems engineering, with attention to the empirical evaluation of learning algorithms on real data. He works on reproducible experimental comparison of classifiers and on the practical cost of model selection and tuning in resource-constrained settings.