CuBAS: A New Curvature-Based Adaptive Sampling Method for Supervised Classification

Anonymous authors
Paper under double-blind review

Abstract

The informativeness of a training set is often more important than its size. However, existing sampling strategies largely ignore the intrinsic geometry of labeled data. We introduce CuBAS (curvature-based adaptive sampling), a information-geometric framework for adaptive sample selection in supervised classification. The key insight of the new method is that a labeled dataset induces a statistical manifold whose local curvature provides a principled measure of sample informativeness. Specifically, we model label interactions through a q-state Potts Markov random field defined on a k-nearest-neighbor graph, and derive a closed-form curvature estimator from the observed Fisher information. Adaptive thresholding of the resulting curvature signal partitions the graph into smooth low-curvature regions – well represented by a small number of prototypes, and high-curvature regions concentrated around class interfaces – where samples are most informative for defining decision boundaries. While jointly sampling from both regimes, CuBAS constructs compact yet highly informative training subsets that preserve global representativeness as well as local discriminative structure. Extensive experiments on more than 60 benchmark datasets spanning tabular, image, and biological domains demonstrate consistent and statistically significant improvements over random and entropy-based sampling across a wide range of labeling budgets. Beyond its empirical effectiveness, CuBAS provides a computationally efficient and theoretically grounded connection between information geometry and adaptive data selection for supervised learning.

1 Introduction

The effectiveness of a classifier depends far more on the *quality* and representativeness of its training samples than on their sheer quantity. Large datasets often contain substantial redundancy, noisy observations, class imbalance, and numerous samples that contribute little to defining the underlying decision boundaries, which motivates a fundamental question: *how can we identify the most informative samples in a labeled dataset for classification?*

Traditional sampling and instance-reduction techniques, including random subsampling, prototype selection, and instance weighting, are typically heuristic and operate primarily in the input space, making limited use of the statistical relationships among neighboring samples (Wilson, 1972; Garcia et al., 2012). Graph-based learning offers a powerful alternative for capturing local and global relationships among samples (Zhu et al., 2003; Belkin et al., 2006): the informative content of a sample is not solely determined by its individual attributes, but also by its structural role within the graph topology, with vertices near class transitions typically carrying more discriminative information than vertices in homogeneous neighborhoods. Despite these advances, existing graph-based sampling strategies rarely exploit the differential geometry induced by statistical interactions on the graph, relying instead on fixed similarity measures or purely combinatorial criteria.

Geometric Machine Learning exploits the geometric structure of data, through manifolds, graphs, or other non-Euclidean domains, to design learning algorithms that are more robust, interpretable, and statistically efficient (Bronstein et al., 2017; Papillon et al., 2025; Weber, 2025). In contrast to most existing geometric learning methods, which primarily focus on representation or embedding learn-

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.

ing, we employ geometry to address a different but equally fundamental problem: identifying the most informative training samples. Information geometry offers the natural mathematical bridge: by viewing a statistical model as a Riemannian manifold endowed with the Fisher information metric, one can characterize the local geometry of probability distributions through curvature and other differential-geometric quantities (Amari, 2016; Nielsen, 2020). Curvature quantifies how rapidly the statistical model changes locally, making it a natural descriptor of sample informativeness, yet information geometry has received comparatively little attention as a mechanism for adaptive sampling in supervised learning.

Existing paradigms for selecting informative samples include active learning, which queries informative unlabeled samples according to uncertainty or margin criteria but is inherently classifier-dependent (Tong & Koller, 2002; Settles, 2010); curriculum learning, which presents samples according to a predefined, typically heuristic, notion of difficulty (Bengio et al., 2009); and instance/prototype selection, which reduces redundancy through distance- or density-based criteria in the original feature space (Wilson, 1972; Garcia et al., 2012; Liu & Motoda, 2002; Aggarwal, 2014; Olvera-López et al., 2010; de Haro-García et al., 2019). None of these explicitly account for the statistical geometry induced by class interactions.

This paper introduces a different perspective: we define informativeness as an intrinsic geometric property of a statistical manifold induced by a labeled graph. We model the graph using a q-state Potts Markov random field, whose local interactions define a probabilistic manifold parameterized by the inverse temperature β. First- and second-order Fisher information induce the metric and curvature of this manifold, enabling a local shape operator that quantifies the geometric complexity surrounding each vertex; sample selection then becomes a problem of identifying regions of high statistical curvature. Building on this formulation, we propose *Curvature-Based Adaptive Sampling* (CuBAS), a graph-based, model-agnostic sampling strategy that decomposes a labeled dataset into complementary low- and high-curvature subsets. Unlike existing sampling techniques, CuBAS derives its selection criterion directly from the geometry of the underlying statistical model, simultaneously capturing boundary awareness, redundancy reduction, and noise filtering.

Contributions. (1) We introduce a graph-based information-geometric framework that models labeled datasets as statistical manifolds induced by a q-state Potts MRF, connecting graph topology and differential geometry. (2) We derive closed-form first- and second-order Fisher information for the Potts model and show how they induce a local shape operator quantifying statistical curvature at each graph vertex. (3) We propose CuBAS, a computationally efficient curvature-driven adaptive sampling algorithm together with a data-driven threshold estimator for the curvature partition. (4) We demonstrate, on more than 60 benchmark datasets and multiple regimes (small, large, and high-dimensional), that CuBAS consistently yields higher classification accuracy than random sampling and entropy-based uncertainty sampling.

Section 2 reviews the Potts MRF and its parameter estimation and summarizes the differential-geometric background. Section 3 develops the information-geometric curvature framework and the CuBAS algorithm. Section 4 reports the experimental evaluation, and Section 5 concludes. Full derivations, additional tables, and extended results are provided in the Appendix.

2 Theoretical Background

The isotropic Potts Markov random field. The Potts model is a discrete MRF in which each random variable assumes one of q states, originally introduced for interacting spin systems and since adopted across statistics, biology, and vision as a model of contextual dependence among discrete variables (Ge et al., 1996; Montroll, 1941; Ouchi et al., 2003; Storath et al., 2015). By the Hammersley–Clifford theorem (Hammersley & Clifford, 1971), the model admits an equivalent local conditional formulation, which we adopt here as it avoids the intractable partition function. Let η_i denote the neighborhood of node i; the local conditional distribution of the isotropic pairwise q-state Potts model is

$$p\left(x_i = m \mid \eta_i, \beta\right) = \frac{\exp(\beta\, U_i(m))}{\sum_{\ell=1}^{q} \exp(\beta\, U_i(\ell))}, \tag{1}$$

where $U_i(m)$ is the number of neighbors of node i labeled m, and β is the inverse temperature controlling the strength of contextual dependence: larger β promotes homogeneous label agreement, while $\beta = 0$ yields statistical independence.

Parameter estimation. Maximum likelihood estimation of β is intractable due to the partition function; we instead use maximum pseudo-likelihood (MPL) (Besag, 1977), which replaces the joint distribution by a product of local conditional terms and is consistent and asymptotically normal under mild conditions (Jensen & Künsh, 1994). The MPL estimating equation is solved numerically via the secant method (Burden & Faires, 2015), which converges superlinearly without requiring explicit derivatives; the full derivation and complexity analysis are given in Appendix A.1. This estimation step is $O(T_\beta\, n\, q)$, with $T_\beta < 20$ iterations typically sufficient, and is therefore negligible relative to graph construction.

Differential geometry basics. We interpret the Potts parameter space as a statistical manifold endowed with a Riemannian structure induced by Fisher information (Amari, 2016; Nielsen, 2020). For a manifold M parameterized by local coordinates θ via $\mathbf{X}(\theta)$, the *first fundamental form* (metric tensor) $g_{ij}(\theta) = \langle \partial\mathbf{X}/\partial\theta_i, \partial\mathbf{X}/\partial\theta_j \rangle$ encodes intrinsic distances, while the *second fundamental form* $b_{ij}(\theta) = \langle \partial^2\mathbf{X}/\partial\theta_i\partial\theta_j, \mathbf{N} \rangle$, with $\mathbf{N}$ the unit normal field, captures how the manifold bends within its embedding space. Combining both yields the *shape operator*

$$\mathbf{S}(\theta) = -\mathbf{B}(\theta)\,\mathbf{G}(\theta)^{-1}, \tag{2}$$

whose eigenvalues are the principal curvatures, whose determinant is the Gaussian curvature, and whose trace (halved) is the mean curvature (do Carmo, 2017). In information geometry, the role of $\mathbf{G}$ is naturally played by the Fisher Information Matrix. Because the isotropic Potts model is parameterized solely by β, the induced statistical manifold is one-dimensional and the shape operator reduces to a *scalar*, which we use directly as a local curvature score in Section 3. A geometric illustration of the shape operator is given in Figure 1.

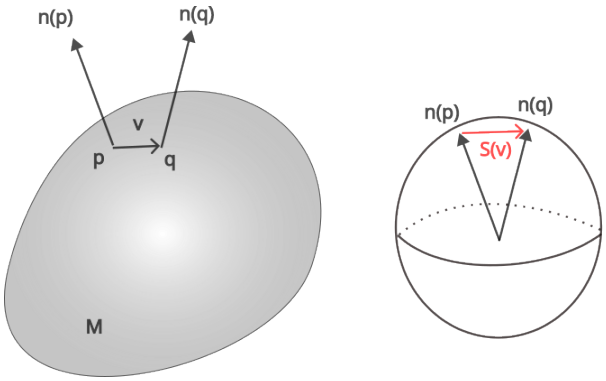

Figure 1: Geometric interpretation of the shape operator. As one moves along a tangent direction $\mathbf{v}$, the shape operator measures the directional variation of the unit normal vector, providing a quantitative description of the local bending of the manifold.

3 CuBAS: Curvature-Based Adaptive Sampling

CuBAS interprets the labeled dataset as a statistical manifold whose curvature, at each vertex of a k-NN graph, is estimated from the Fisher information of the local Potts conditional distribution. Regions of high curvature correspond to parameter configurations where small perturbations induce large changes in the likelihood, increased sensitivity, reduced redundancy, and greater discriminative potential, typically near class boundaries. Flat regions indicate statistical homogeneity, where samples are largely interchangeable.

Fisher information of the Potts model. The first-order Fisher information $\mathcal{I}(\beta) = \mathbb{E}[(\partial_\beta \log p(X;\beta))^2]$ defines the metric structure of the manifold, while the second-order Fisher information $\mathcal{II}(\beta) = -\mathbb{E}[\partial_\beta^2 \log p(X;\beta)]$ captures its curvature (Lehmann & Casella, 2011). Following the observed-Fisher-information approximation (Cao & Spall, 2009), these expectations are estimated empirically at $\hat{\beta}$ (the MPL estimate) and decompose naturally into per-node contributions $\Phi_i(\beta)$ and $\Psi_i(\beta)$. Substituting the Potts conditional (Eq. 1) and simplifying (full derivation in

Appendix A.2) yields

$$\Phi_i(\beta) = \left[\frac{\sum_{\ell=1}^{q} \big(U_i(x_i) - U_i(\ell)\big) \exp\{\beta U_i(\ell)\}}{\sum_{\ell=1}^{q} \exp\{\beta U_i(\ell)\}} \right]^2 \tag{3}$$

$$\Psi_i(\beta) = \frac{\sum_{\ell=1}^{q} \sum_{k=1}^{q} \hat{U}_i(\ell,k) \exp\{\beta(U_i(\ell) + U_i(k))\}}{\sum_{\ell=1}^{q} \sum_{k=1}^{q} \exp\{\beta(U_i(\ell) + U_i(k))\}} = \mathrm{Var}_{\hat{\beta}}\big[U_i(\cdot)\big] \tag{4}$$

where

$$\hat{U}_i(\ell,k) = U_i(\ell)^2 - U_i(\ell)U_i(k). \tag{5}$$

Note that $\Phi_i(\beta)$ measures the sensitivity of the local energy landscape to perturbations of β, and $\Psi_i(\beta)$ is the conditional variance of the local energy function under the Gibbs distribution at $\hat{\beta}$, a curvature-sensitive quantity complementary to Φ_i. Both admit compact tensorial forms amenable to vectorized computation (Appendix A.2).

Local shape operator and curvature score. Since the Potts statistical manifold is one-dimensional, the shape operator of Eq. 2 reduces to the scalar

$$S_i(\beta) = -\frac{\Psi_i(\beta)}{\Phi_i(\beta) + \lambda}, \tag{6}$$

with $\lambda = 10^{-4}$ a regularization constant preventing division by zero. High $|S_i(\beta)|$ indicates a neighborhood where the local statistical structure changes rapidly (high information concentration); near-zero values indicate flat, redundant regions. After computing $S_i(\beta)$ for all nodes, scores are normalized to $[0,1]$ to obtain the curvature field $K = \{K_i\}_{i=1}^n$.

Algorithm. CuBAS constructs the k-NN graph, estimates $\hat{\beta}$ via MPL, computes the curvature field K, partitions the graph into low- and high-curvature sets $\mathcal{L} = \{i : K_i < T\}$ and $\mathcal{H} = \{i : K_i \geq T\}$ using an adaptive threshold T (Section 3.1), and samples a fraction p of nodes from each set to form the training subset. The procedure is summarized in Algorithm 1.

Algorithm 1 CuBAS: Curvature-Based Adaptive Data Decomposition

Require: Labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$; distance $d(\cdot,\cdot)$; number of neighbors k; training fraction p

Ensure: Training set $\mathcal{D}_{train}$, test set $\mathcal{D}_{test}$

1: Construct k-NN graph $G = (\mathcal{V}, \mathcal{E})$ from $\mathcal{D}$ using $d(\cdot,\cdot)$.
2: Estimate $\hat{\beta}$ via maximum pseudo-likelihood (MPL) on G.
3: **for** each node $i \in \mathcal{V}$ **do**
4: Compute $\Phi_i(\hat{\beta})$, $\Psi_i(\hat{\beta})$ (Eq. 4) and $S_i(\hat{\beta}) \leftarrow -\frac{\Psi_i}{(\Phi_i + \lambda)}$
5: **end for**
6: Normalize $\{S_i\}$ to obtain $K \in [0,1]^n$.
7: Estimate adaptive threshold T^* (Algorithm 2); set $\mathcal{L} = \{i : K_i < T^*\}$, $\mathcal{H} = \{i : K_i \geq T^*\}$.
8: $\mathcal{N}_L \leftarrow \lfloor p|\mathcal{L}| \rfloor$, $\mathcal{N}_H \leftarrow \lfloor p|\mathcal{H}| \rfloor$.
9: Sample $\mathcal{N}_L$ nodes from $\mathcal{L}$ and $\mathcal{N}_H$ nodes from $\mathcal{H}$ without replacement to form $\mathcal{X}_L, \mathcal{X}_H$.
10: **return** $\mathcal{D}_{train} = [\mathcal{X}_L, \mathcal{X}_H]$, $\mathcal{D}_{test} = [\mathcal{L} \setminus \mathcal{X}_L,\ \mathcal{H} \setminus \mathcal{X}_H]$

This mechanism admits a natural analogy with low-pass/high-pass filtering: $\mathcal{L}$-nodes yield a smoothed approximation of the graph that preserves cluster structure, while $\mathcal{H}$-nodes selectively highlight class interfaces and decision boundaries, the most discriminative interactions in the data. A visualization of the identified high-curvature nodes on four representative datasets is given Figure 2, confirming that they concentrate near decision boundaries.

3.1 Adaptive Information Threshold Estimation

A critical design choice in our framework is the threshold T that partitions the normalized curvature scores $K = \{K_i\}_{i=1}^n$, $K_i \in [0,1]$, into low-curvature and high-curvature populations. A fixed quantile threshold, while simple, fails to account for the heterogeneous shape of the curvature distribution across datasets: some labeled graphs exhibit a clearly bimodal curvature profile, with a dense low-curvature mode corresponding to homogeneous clusters and a separated high-curvature

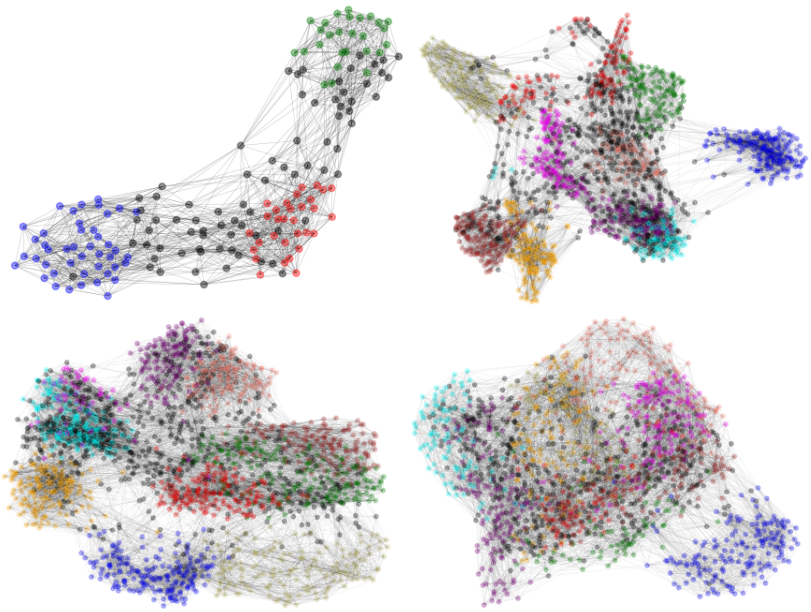

Figure 2: Visualization of high-curvature nodes ($\mathcal{H}$-nodes) identified by CuBAS in the *digits*, *mfeat_karhunen*, *semeion*, and *wine* datasets. High-curvature samples are highlighted in black and concentrate around class boundaries, where inter-class interactions are most prominent.

mode concentrated near decision boundaries, while others exhibit a unimodal, right-skewed profile in which only a small outlier tail is genuinely informative. We therefore propose an *adaptive* threshold estimator that selects between two complementary criteria based on a quantitative bimodality diagnostic, rather than committing a priori to either regime.

3.1.1 Otsu's Criterion for Bimodal Curvature Distributions

When the curvature distribution is bimodal, the threshold separating the two populations can be estimated by maximizing the between-class variance of the induced partition, following the classical criterion of Otsu (1979). Given a histogram of K with B bins and bin centers $\{\kappa_b\}_{b=1}^{B}$, a candidate threshold T induces a low-curvature class $L(T) = \{i : K_i < T\}$ and a high-curvature class $H(T) = \{i : K_i \geq T\}$ with population fractions $w_L(T)$ and $w_H(T) = 1 - w_L(T)$, and class means $\mu_L(T)$ and $\mu_H(T)$. The optimal threshold maximizes the inter-class variance.

3.1.2 Bimodality Validation via Ashman's D Coefficient

Because Otsu's criterion always returns *some* threshold regardless of whether the underlying distribution is genuinely bimodal, we validate its applicability using Ashman's D statistic (Ashman et al., 1994), a standard measure of separation between two Gaussian components. We fit a two-component Gaussian mixture model (GMM) to K via expectation-maximization with multiple random restarts, and compute

$$D = \frac{\sqrt{2}\,|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \tag{7}$$

The coefficient D quantifies the separation between the two estimated modes relative to their dispersion: $D \geq 2$ is the classical threshold in the mixture-modeling literature for *well-separated, unimodally-resolvable* components, below which the two-component fit cannot be reliably distinguished from a single mode (Ashman et al., 1994). In our implementation we adopt a slightly more conservative threshold, $D_{\min} = 2.6$, to reduce the risk of spuriously invoking the bimodal regime on curvature distributions with heavy but unimodal tails. Additionally, since a two-component GMM

Algorithm 2 Adaptive Threshold Estimation

Require: Normalized curvatures $K \in [0,1]^n$, number of bins $B = 256$, bimodality threshold $D_{\min}$
Ensure: Threshold T^*, selection method $\text{tag} \in \{\text{OTSU}, \text{IQR}, \text{DEGENERATE}\}$
if $\max(K) - \min(K) < \epsilon$ **then**
  return $Q_{0.95}(K)$, DEGENERATE
end if
Compute T_{Otsu} using cumulative-sum histogram evaluation
Fit two-component GMM to K; compute D, μ_1, μ_2
if T_{Otsu} is defined **and** $D \geq D_{\min}$ **then**
  if $\mu_2 < 0.1$ **then**
    $T^* \leftarrow T_{\text{IQR}}$ via Eq. equation 8 ▷ spurious bimodal fit; revert to fallback
    return T^*, DEGENERATE
  else
    return T_{Otsu}, OTSU
  end if
else
  $T^* \leftarrow T_{\text{IQR}}$ via Eq. equation 8
  return T^*, IQR
end if

can in principle converge to two modes that are both concentrated near zero curvature (i.e., $\mu_2 < 0.1$, indicating the "high-curvature" component is itself an artifact of noise in a predominantly homogeneous graph), we treat this configuration as a degenerate case and revert to the fallback criterion even when $D \geq D_{\min}$.

We note that Ashman's D statistic, as originally formulated, implicitly assumes comparable dispersion between the two mixture components. When the two curvature modes are markedly heteroscedastic ($\sigma_1 \neq \sigma_2$), the asymptotic χ^2 behavior underlying related likelihood-ratio bimodality tests breaks down for small or moderate sample sizes, and a parametric bootstrap is recommended instead (Lo, 2008). In our setting, the curvature populations induced by $\mathcal{L}$ and $\mathcal{H}$ are not assumed homoscedastic; we therefore treat $D \geq D_{\min}$ as a practical screening criterion rather than a formal hypothesis test, consistent with its standard usage in the mixture-modeling literature (Ashman et al., 1994; Lo, 2008).

3.1.3 TUKEY-FENCE FALLBACK FOR UNIMODAL DISTRIBUTIONS

When the bimodality test fails ($D < D_{\min}$) or the Otsu criterion itself is degenerate, the curvature distribution is treated as effectively unimodal, and only the upper outlier tail is considered genuinely high-information. We adopt a quantile-based threshold informed by the interquartile range (IQR),

$$T_{\text{IQR}} = \begin{cases} Q_3(K), & \text{if } \text{IQR}(K) > 0, \\ Q_{0.95}(K), & \text{if } \text{IQR}(K) = 0, \end{cases} \tag{8}$$

where $Q_3(K)$ is the third quartile and $\text{IQR}(K) = Q_3(K) - Q_1(K)$. The degenerate case $\text{IQR}(K) = 0$ arises when the curvature distribution is heavily concentrated (e.g., more than 75% of nodes share an identical curvature value, typically $K_i = 0$ after normalization), in which case the 75th-percentile threshold collapses and the 95th-percentile is used instead to preserve a non-trivial high-curvature subset.

The complete adaptive threshold estimator is summarized as a decision cascade in Algorithm 2. The procedure first attempts the bimodal (Otsu) criterion and accepts it only if both (i) the inter-class variance is non-degenerate and (ii) Ashman's D statistic confirms genuine separation between the two curvature modes with the high-curvature mode being non-trivial ($\mu_2 \geq 0.1$). Otherwise, the method falls back to the Tukey-fence quantile rule of Eq. equation 8. This cascade allows T to adapt not only to the dataset at hand, but implicitly to the local geometry and class overlap structure encoded in the Potts-derived curvature field, without requiring any dataset-specific tuning.

Once T^* is determined, the high- and low-information partitions are computed replacing a fixed-quantile rule with a threshold that responds to the empirical shape of the curvature field on a per-dataset basis.

3.2 Computational Complexity

CuBAS decomposes into k-NN graph construction ($O(nm \log n)$ for low-to-moderate ambient dimension m, via a k-d/ball tree), the label-count matrix ($O(nk)$), MPL estimation of β ($O(T_\beta\, nc)$), Fisher information and curvature scores ($O(nc)$, fully vectorized), adaptive threshold estimation ($O(n + B)$), and graph partitioning/sampling ($O(n)$). The overall time complexity is

$$O(nm \log n + nk + T_\beta\, nc)\,, \tag{9}$$

which reduces to $O(nm \log n)$ when k, c (number of classes), and T_β are treated as constants, the typical regime in supervised classification benchmarks. Crucially, once the k-NN graph is built and $\hat{\beta}$ is estimated, the curvature decomposition itself, CuBAS's distinguishing computational contribution relative to any k-NN-based method, adds only $O(nc)$ linear overhead. This compares favorably with entropy-based uncertainty sampling, which additionally requires training and repeatedly querying an auxiliary classifier. A full per-stage breakdown is given in Appendix A.4 (Table 2).

4 Computational Experiments and Results

We evaluate CuBAS against random sampling and entropy-based uncertainty sampling (Roy & McCallum, 2001; Settles, 2009; Ren et al., 2021), one of the most widely adopted active-learning baselines, addressing three questions: (i) does curvature-aware sampling select more informative training subsets than geometry-agnostic and uncertainty-based strategies; (ii) how does its benefit scale with training-set size, feature dimensionality, and number of classes; and (iii) does the adaptive threshold estimator preserve robustness across datasets with heterogeneous curvature distributions?

Datasets and protocol. We use more than 60 datasets from OpenML (Vanschoren et al., 2013; Bischl et al., 2021), grouped into small ($n < 1100$, 20 datasets), large ($n > 2000$, 32 datasets), and high-dimensional ($m > 1000$, 16 datasets) benchmarks, spanning tabular, image, genomic, spectroscopic, and text domains, with n ranging from 60 to nearly 15,000, m from 4 to over 22,000, and c from 2 to 40 classes (full dataset listings in Appendix A.5). For each dataset, we evaluate 15 training-set proportions from 10% to 80% (5% increments), with 3 sampling strategies (random, entropy, CuBAS with $k = 15$), each repeated over 100 independent train/test partitions. A k-NN classifier ($k = 5$) is trained and evaluated on each partition using balanced accuracy, averaged over the 100 repetitions and 15 training sizes, i.e., $1{,}500$ runs per dataset and strategy.

Table 1: Balanced accuracy of a k-NN classifier under random sampling, entropy-based uncertainty sampling, and CuBAS, aggregated over the 15 training-set sizes (10%–80%) and 100 repetitions per size, grouped by dataset regime. Per-dataset results for all 68 datasets are given in Appendix A.6 (Tables 6–8).

		Random		**Entropy**		**CuBAS**	
Benchmark	**# datasets**	Avg	Median	Avg	Median	Avg	Median
Small ($n < 1100$)	20	0.7202	0.7238	0.7416	0.7412	**0.8556**	**0.8944**
Large ($n > 2000$)	32	0.7273	0.7991	0.7623	0.8358	**0.8455**	**0.9397**
High-dimensional ($m > 1000$)	16	0.7339	0.7365	0.7573	0.7612	**0.8451**	**0.9404**

Overall results. Table 1 summarizes results across the three benchmark groups. CuBAS achieves the highest balanced accuracy on *every single one* of the 68 datasets, with absolute average improvements over entropy sampling ranging from $+8.3$ to $+11.4$ percentage points, and over random sampling from $+11.1$ to $+13.5$ points, depending on the benchmark group. Gains are largest on datasets with well-defined geometric decision boundaries or high class overlap (e.g., `breast_cancer`: 0.9996 vs. 0.9752/0.9469; `credit-g`: 0.7940 vs. 0.5854/0.5960) and remain positive even in adversarial, high-imbalance settings such as `arrhythmia` (0.1724 vs. 0.1439/0.1227). In the high-dimensional, low-sample-size regime, where Euclidean distance concentration is known to challenge

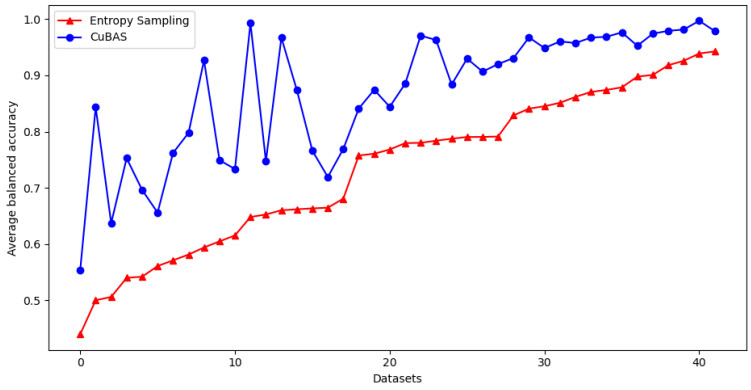

Figure 3: Balanced accuracy of entropy-based sampling vs. CuBAS at the smallest training-set size (10%), averaged over 100 independent train/test partitions per dataset.

geometry-based methods (Donoho, 2000), CuBAS still dominates on all 16 datasets, suggesting that its graph-based, non-parametric curvature estimator provides a form of implicit dimensionality reduction. Per-dataset tables and discussion are provided in Appendix A.6.

Low-budget regime. Figure 3 compares CuBAS against entropy sampling using only 10% of the available samples for training, across all 68 datasets: the setting of greatest practical relevance, since the quality of a small labeled subset has the largest impact on downstream generalization. CuBAS outperforms entropy sampling on every dataset without exception, with a mean absolute gain of $+13.5$ percentage points ($+20.2\%$ relative) and a median gain of $+10.7$ points (Wilcoxon signed-rank test, $p < 10^{-10}$). The minimum observed gain across all 68 datasets is $+3.6$ points, and CuBAS also exhibits lower variance across datasets (std. 0.114 vs. 0.137), indicating more predictable performance at low labeling budgets.

Data efficiency across budgets. Across the full range of training sizes (10–80%), the accuracy curve of CuBAS lies above both baselines at virtually every training size and on every dataset examined (accuracy-vs-budget curves for eight representative datasets are shown in Figures 4–6), and it reaches, at smaller training fractions, the level of balanced accuracy that competing methods require substantially more data to attain, a property we call *data efficiency*. This has direct practical implications in domains where labeled data is scarce or costly (medical imaging, genomic profiling, industrial quality control), where matching or exceeding random-sampling performance with a fraction of the labels translates into reduced annotation cost without sacrificing accuracy.

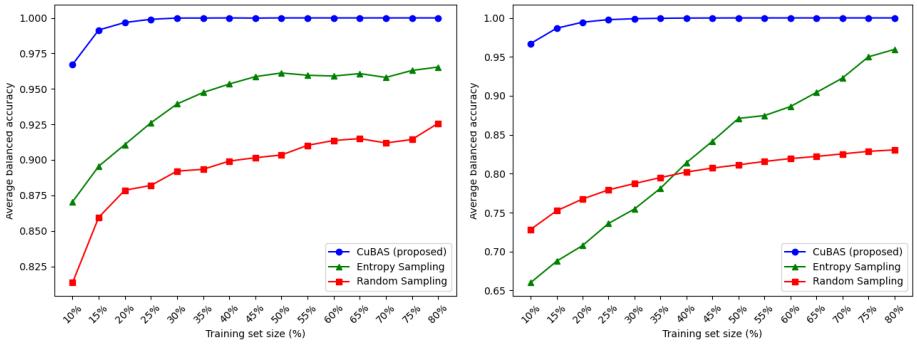

Figure 4: Average balanced accuracies obtained by random sampling, entropy-based sampling, and CuBAS for datasets AP_Breast_Uterus (left) and eeg_eye_state (right), as a function of training-set size.

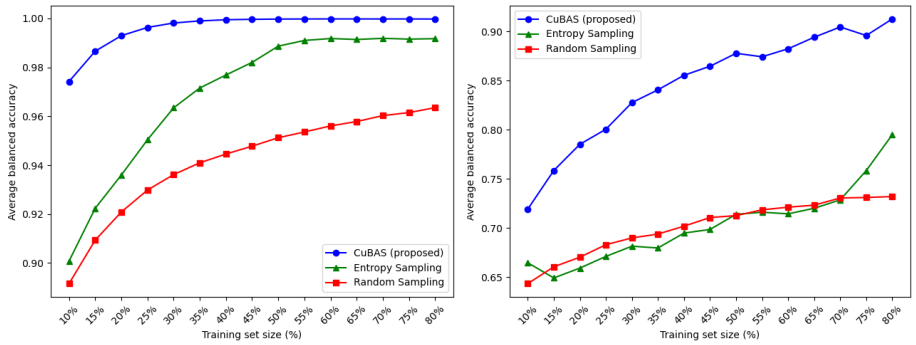

Figure 5: Average balanced accuracies obtained by random sampling, entropy-based sampling, and CuBAS for datasets har (left) and Indian_pines (right), as a function of training-set size.

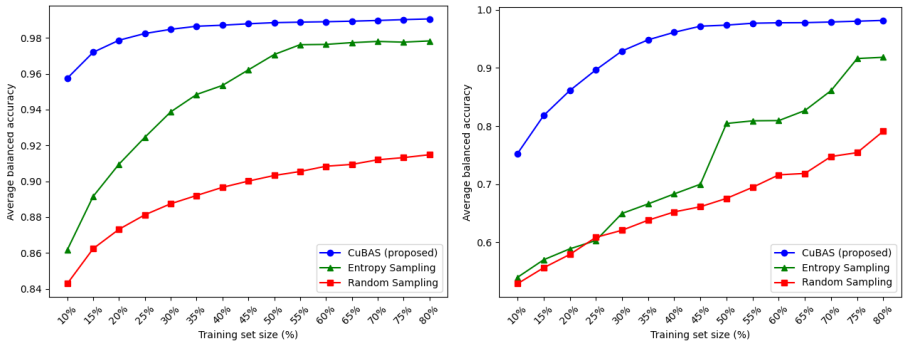

Figure 6: Average balanced accuracies obtained by random sampling, entropy-based sampling, and CuBAS for datasets MNIST (left) and nursery (right), as a function of training-set size.

5 Conclusion

We introduced CuBAS, an information-geometric framework that treats sample selection for supervised classification as a curvature-estimation problem on a statistical manifold induced by a q-state Potts Markov random field. The central contribution is a closed-form, per-vertex curvature score derived from the ratio of second- to first-order observed Fisher information on a labeled k-NN graph, requiring no eigendecompositions, kernel density estimates, or auxiliary pre-trained classifiers. This score naturally separates smooth, redundant cluster interiors from decision-boundary regions that are disproportionately informative for classification, and, combined with a data-driven adaptive threshold estimator (Otsu's criterion validated via Ashman's D, with a Tukey-fence fallback), yields training subsets that are simultaneously compact, representative, and discriminative, without dataset-specific tuning. On a benchmark of more than 60 OpenML datasets spanning tabular, image, genomic, and spectroscopic domains, CuBAS outperformed random and entropy-based uncertainty sampling on every dataset, with a cost that is linear in the number of edges.

Several directions remain open for future work: extending the framework to inhomogeneous or anisotropic MRFs (e.g., edge-weighted or class-dependent couplings); a curvature-aware *active* learning variant that refines the partition as labels are acquired sequentially; curvature regularization (e.g., Laplacian smoothing) for the ultra-small-sample, large-p regime; integration with learned embeddings to extend CuBAS to unstructured data such as images and text; a formal sample-complexity analysis; extensions to heterogeneous, multi-modal graphs; and using CuBAS-selected subsets as geometry-aware warm-up sets for neural network initialization. These directions, together with further discussion, are elaborated in Appendix A.7. In summary, CuBAS offers a principled, scalable, and theoretically grounded alternative to heuristic data selection for supervised learning, bridging information geometry, Markov random field theory, and adaptive sampling.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing. We have not used generative AI tools to generate synthetic data sets, interpret results, implement methods, or tasks regarding the support for mathematical proofs. Lastly, we take responsibility for the final content of this work.

REFERENCES

C.C. Aggarwal. *Data Classification: Algorithms and Applications*. Chapman and Hall/CRC, 2014. doi: https://doi.org/10.1201/b17320.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

Kathryn M. Ashman, Christine M. Bird, and Stephen E. Zepf. Detecting bimodality in astronomical datasets. *The Astronomical Journal*, 108:2348–2361, 1994. doi: 10.1086/117248.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399–2434, 2006. URL http://jmlr.org/papers/v7/belkin06a.html.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. doi: 10.1145/1553374.1553380.

Julian Besag. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 64 (3):616–618, 12 1977.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G. Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites. In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/1708.03731.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017. doi: 10.1109/MSP.2017.2693418.

Richard L. Burden and J. Douglas Faires. *Numerical Analysis*. Cengage Learning, Boston, MA, 10th edition, 2015. ISBN 9781305253667.

Xumeng Cao and James C. Spall. Preliminary results in comparing the expected and observed fisher information for maximum likelihood estimates. In *2009 43rd Annual Conference on Information Sciences and Systems*, pp. 436–441, 2009.

Aida de Haro-García, Gonzalo Cerruela-García, and Nicolás García-Pedrajas. Instance selection based on boosting for instance-based learners. *Pattern Recognition*, 96:106959, 2019. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2019.07.004.

Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces*. Dover Publications Inc., 2nd edition, 2017.

David L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS Conference on Math Challenges of the 21st Century*, pp. 1–32. American Mathematical Society, 2000. URL https://pdfs.semanticscholar.org/63c6/8278418b69f60b4814fae8dd15b1b1854295.pdf.

Salvador Garcia, Joaquin Derrac, Jose Cano, and Francisco Herrera. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):417–435, 2012. doi: 10.1109/TPAMI.2011.142.

Mo-Lin Ge, Liangzhong Hu, and Yiwen Wang. Knot theory, partition function and fractals. *Journal of Knot Theory and Its Ramifications*, 05(01):37–54, 1996.

J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices (preprint). unpublished, 1971. URL www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf.

J. L. Jensen and H. R. Künsh. On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics*, 46:475–486, 1994.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer, 2 edition, 2011.

H. Liu and H. Motoda. On issues of instance selection. *Data Mining and Knowledge Discovery*, 6: 115–130, 2002. URL https://doi.org/10.1023/A:1014056429969.

Yungtai Lo. A likelihood ratio test of a homoscedastic normal mixture against a heteroscedastic normal mixture. *Statistics and Computing*, 18(2):233–240, 2008. doi: 10.1007/s11222-008-9052-4.

Elliott W. Montroll. Statistical Mechanics of Nearest Neighbor Systems. *Journal of Chemical Physics*, 9(9):706–721, September 1941.

Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020. ISSN 1099-4300. doi: 10.3390/e22101100. URL https://www.mdpi.com/1099-4300/22/10/1100.

J. Arturo Olvera-López, J. Ariel Carrasco-Ochoa, J. Francisco Martínez-Trinidad, and Josef Kittler. A review of instance selection methods. *Artificial Intelligence Review*, 34:133–143, 2010. URL https://doi.org/10.1007/s10462-010-9165-y.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Noriyuki Bob Ouchi, James A. Glazier, Jean-Paul Rieu, Arpita Upadhyaya, and Yasuji Sawada. Improving the realism of the cellular potts model in simulations of biological cells. *Physica A: Statistical Mechanics and its Applications*, 329(3):451–458, 2003.

Mathilde Papillon, Sophia Sanborn, Johan Mathe, Louisa Cornelis, Abby Bertics, Domas Buracas, Hansen J. Lillemark, Christian Shewmake, Fatih Dinc, Xavier Pennec, and Nina Miolane. Beyond Euclid: An illustrated guide to modern machine learning with geometric, topological, and algebraic structures. *Machine Learning: Science and Technology*, 6, 2025. doi: 10.1088/2632-2153/adf375.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9): 1–40, 2021.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 441–448, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Burr Settles. Active learning literature survey. *University of Wisconsin–Madison Technical Report*, 2010.

Martin Storath, Andreas Weinmann, Jürgen Frikel, and Michael Unser. Joint image reconstruction and segmentation using the potts model. *Inverse Problems*, 31(2):025003, jan 2015.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, March 2002. ISSN 1532-4435. doi: 10.1162/153244302760185243.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations Newsletter*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198.

Melanie Weber. Geometric machine learning. *AI Magazine*, 46(1):e12210, 2025. doi: https://doi.org/10.1002/aaai.12210.

Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972. doi: 10.1109/TSMC.1972.4309137.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 912–919. AAAI Press, 2003. ISBN 1577351894.