Evaluating Instance Selection for Active Learning using Label-Agnostic Instance Hardness Measures

Bárbara de Morais Coutinho[1], **Patricia S. M. Ueda**[1], **Ana Carolina Lorena**[1]

[1]Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos – SP – Brazil

barbara.coutinho.102056@ga.ita.com, uedapsm@gmail.com, aclorena@ita.br

Abstract. *Supervised machine learning often requires large amounts of labeled data. Active Learning (AL) addresses this limitation by iteratively selecting the most informative instances from an unlabeled pool, aiming to maximize predictive performance. Although traditional AL strategies are largely driven by model uncertainty, alternative approaches exploit Instance Hardness Metrics (IHMs) to prioritize challenging samples. However, many IHMs rely on true class labels for query instances, which are unavailable during the AL selection. This study investigates the impact of replacing traditional, label-dependent hardness metrics with adjusted, label-agnostic versions within an AL framework. Experiments were conducted on 18 diverse datasets using an XGBoost classifier, and performance was assessed through the Area Under the Kappa Curve. The adjusted metrics eliminate the need for prior label information for query instances, making them considerably more practical for real-world AL applications. Among the evaluated methods, the adjusted k-Disagreeing Neighbors metric consistently achieved the strongest performance across most datasets.*

1. Introduction

Supervised Machine Learning (ML) has achieved success across a wide range of application domains. However, its effectiveness often depends on the availability of large amounts of labeled data. Obtaining such labels can be expensive, time-consuming, and frequently requires the involvement of domain experts. To address this challenge, Active Learning (AL) has emerged as a learning paradigm that aims to maximize predictive performance while minimizing labeling effort by selectively querying the most informative instances for annotation [Ren et al. 2021, Settles 2009].

The central idea of AL is to identify, from a pool of unlabeled data, the instances that are expected to provide the greatest benefit to the learning process. Among the most widely adopted instance selection strategies are uncertainty-based methods, collectively known as Uncertainty Sampling. These approaches rely on the predictions of an ML model to identify samples for which the model exhibits greater uncertainty [Settles 2009]. Such instances are generally more prone to misclassification and are therefore considered informative candidates for labeling [Monarch 2021].

Although uncertainty-based approaches have demonstrated strong performance, alternative criteria for sample selection have recently attracted increasing attention. One such alternative is Instance Hardness (IH) [Nogueira et al. 2025], which focuses on identifying samples that are inherently difficult to classify [Smith et al. 2014]. IH metrics

typically characterize the difficulty of individual instances based on properties of the data itself, often exploiting local neighborhood information [Lorena et al. 2024].

IH is generally defined as the probability of an instance being misclassified by a pool of learning algorithms [Smith et al. 2014]. In practice, an instance is considered hard when different classifiers frequently misclassify it, whereas easy instances tend to be classified correctly across multiple learning models [Lorena et al. 2024]. Several less computationally demanding measures (Instance Hardness Measures - IHMs) have been proposed to estimate instance hardness from different perspectives, making them attractive candidates for guiding sample selection in AL settings.

Despite their potential for guiding sample selection, many IHMs rely on the true class label of the query instance to estimate its difficulty level. For instance, they consider whether the neighbourhood of the query instance contains elements from its class. This requirement creates a fundamental limitation in AL scenarios, where labels are unavailable during the selection process and are precisely the resource being acquired. In [Nogueira et al. 2025], the authors use predictions from ML models on query instances to obtain these labels for computing traditional IHMs, making IHM computation dependent on a particular learning algorithm.

To overcome this issue, label-free variants of hardness metrics can estimate the difficulty of query instances without prior knowledge of their labels [Ueda et al. 2024]. However, the impact of replacing traditional hardness measures with their adjusted counterparts within an AL cycle remains unexplored. Therefore, this work investigates the effectiveness of adjusted IHMs as query strategies for AL and compares their behavior with both traditional hardness-based approaches from [Nogueira et al. 2025] and uncertainty-based sampling. Experiments across 18 datasets evaluate whether label dependence can be removed without compromising selection efficiency, providing insights into the practical applicability of hardness-based AL in real-world settings.

This paper is structured as follows: Section 2 describes traditional instance-selection measures in AL. Section 3 describes both the traditional and label-agnostic IHMs. Section 4 presents the materials and methods used in this study. The experimental results are presented and discussed in Section 5. Section 6 concludes this paper.

2. Instance Selection for Active Learning

A fundamental component of Active Learning (AL) is the instance selection strategy for identifying the most informative unlabeled instances. The effectiveness of an AL framework largely depends on its ability to select samples that maximize model improvement while minimizing labeling effort.

Let $D_l = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$ denote a labeled dataset, where $\mathbf{x}_i$ represents an input instance and y_i its corresponding label. Let $D_u = \{\mathbf{x}_j\}_{j=1}^{m}$ denote a pool of unlabeled instances. During each AL iteration, a model is trained using D_l, and a query strategy is then employed to select one or more instances from D_u for annotation by an oracle. The newly labeled instances are subsequently incorporated into D_l, and the process is repeated until a stopping criterion is reached.

Among the numerous query strategies proposed in the literature, uncertainty-based methods are the most widely adopted. These approaches use the predictive proba-

bilities generated by a classifier to identify instances where the model is most uncertain. A commonly used strategy is Least Confident (LC), which selects the instance whose most probable class receives the lowest confidence score [Settles 2009, Lewis and Gale 1994]:

$$\mathbf{x}_{LC} = \underset{j=1}{\overset{m}{\operatorname{argmax}}}\{1 - P_\theta(y^*|\mathbf{x}_j)\} = \underset{j=1}{\overset{m}{\operatorname{argmin}}}\{P_\theta(y^*|\mathbf{x}_j)\}, \tag{1}$$

where $P_\theta(y^*|\mathbf{x}_j)$ denotes the probability assigned by the classifier, parametrized by θ, to the most likely class y^* for instance $\mathbf{x}_j$.

Although effective, LC considers only the confidence associated with the most probable class and ignores the remaining probability distribution. To address this limitation, Margin Sampling (MS) [Scheffer et al. 2001] evaluates the difference between the two most likely classes:

$$\mathbf{x}_{MS} = \underset{j=1}{\overset{m}{\operatorname{argmin}}}\{P_\theta(y_1^*|\mathbf{x}_j) - P_\theta(y_2^*|\mathbf{x}_j)\}, \tag{2}$$

where y_1^* and y_2^* correspond to the first and second most probable classes predicted for $\mathbf{x}_j$, respectively. Smaller margins indicate greater uncertainty, as the classifier assigns similar confidence to competing class labels.

A further extension, which considers all C possible classes, is entropy-based sampling, which uses the full predictive distribution. Based on Shannon's entropy, this strategy selects the instance with the highest uncertainty:

$$\mathbf{x}_H = \underset{j=1}{\overset{m}{\operatorname{argmax}}}\{-\sum_{i=1}^{C} P_\theta(y_i|\mathbf{x}_j) \log P_\theta(y_i|\mathbf{x}_j)\}. \tag{3}$$

Despite their popularity, uncertainty-based methods rely entirely on the predictive behavior of a particular model and may not always identify the most informative instances for learning. This limitation has motivated the investigation of alternative query strategies that characterize the difficulty of the data itself rather than the uncertainty of a specific classifier. Among these alternatives, IHMs have recently emerged as a promising approach for guiding sample selection in AL [Nogueira et al. 2025].

3. Instance Hardness Measures

The concept of Instance Hardness (IH) was introduced by [Smith et al. 2014] and characterizes the difficulty of an instance in terms of its probability of being misclassified by a pool of diverse learners. This perspective shifts the focus from dataset-level analyses to the properties of individual instances, acknowledging that samples within the same dataset may exhibit substantially different levels of classification difficulty. According to [Smith et al. 2014], several factors contribute to instance hardness, particularly class overlap, which is often aggravated by class imbalance, noise, or the presence of outliers.

Researchers have proposed a variety of IHMs to estimate this difficulty. Based on the taxonomy in [Lorena et al. 2024], these measures fall into five major categories. In this work, we focus on Neighborhood, Likelihood, and Tree-Based measures. Traditional IHMs generally require access to the true label of the instance being evaluated, as they use

this information directly in the hardness computation. More recently, [Ueda et al. 2024] proposed adjusted versions of several traditional IHMs that eliminate the need for the query instance label during hardness estimation. This label-agnostic formulation is particularly appealing in the context of AL, where candidate instances are selected from the unlabeled pool D_u, and their labels are unavailable at query time.

The following subsections describe the traditional IHMs and their adjusted counterparts considered in this study. All measures are computed using the labeled dataset D_l to characterize the local structure of the data. The key distinction between the two formulations lies in whether the label of the query instance itself is required during the hardness estimation process. All measures are standardized so that higher values correspond to greater instance difficulty.

3.1. Neighborhood Category

k-Disagreeing Neighbors (kDN) quantifies the proportion of an instance's nearest neighbors that belong to classes different from its own. Under this assumption, an instance is considered difficult to classify when it is surrounded by samples from competing classes, indicating a high degree of local class overlap. Formally, kDN is [Smith et al. 2014]:

$$\text{kDN}(\mathbf{x}_i, y_i) = \frac{|\{\mathbf{x}_j \mid \mathbf{x}_j \in kNN(\mathbf{x}_i) \wedge y_j \neq y_i\}|}{k}, \tag{4}$$

where $\text{kNN}(\mathbf{x}_i)$ denotes the set of the k nearest neighbors of $\mathbf{x}_i$. kDN values close to 1 indicate that the instance is located in a neighborhood predominantly composed of examples from other classes, suggesting a high level of classification difficulty. In contrast, values near 0 indicate that most neighboring instances belong to the same class, characterizing an easier classification scenario.

The adjusted kDN measure proposed by [Ueda et al. 2024] eliminates the need for the query instance label by characterizing the class distribution within its local neighborhood. For all neighborhood-based measures (kDN and kDN_{adj}), the target neighborhood size was set to $k = 5$. During the initial AL iterations, when the labeled set size $|D_l|$ was smaller than k, the neighborhood size was dynamically adjusted as $k_{\text{adj}} = \min(k, |D_l|)$. Instead of counting the proportion of neighboring instances that disagree with the query label, kDN_{adj} computes the entropy of the class proportions observed among the k nearest neighbors in the labeled dataset D_l:

$$\text{kDN}_{\text{adj}}(\mathbf{x}) = -\sum_{i=1}^{C} p(y_j = c_i) \log p(y_j = c_i), \text{ for } \mathbf{x}_j \in \text{kNN}(\mathbf{x}), \tag{5}$$

where $p(y_j = c_i)$ denotes the proportion of neighboring instances belonging to class c_i. By measuring the entropy of the neighborhood, kDN_{adj} captures the degree of class heterogeneity surrounding the query instance. High entropy values indicate that the neighborhood contains a balanced mixture of different classes, suggesting that the instance lies near a decision boundary and is therefore more difficult to classify. Conversely, low entropy values indicate a neighborhood dominated by a single class, characterizing an easier classification scenario.

3.2. Class Likelihood Category

The **Class Likelihood Difference (CLD)** measure estimates instance hardness by comparing how strongly an instance is associated with its true class relative to competing classes. Under this perspective, an instance is considered difficult when it deviates from the characteristic patterns of its own class while simultaneously exhibiting similarities to other classes [Smith et al. 2014, Lorena et al. 2024].

$$\text{CLD}(\mathbf{x}_i, y_i) = \frac{1 - (p(\mathbf{x}_i|y_i)p(y_i) - \max_{y_j \neq y_i}[p(\mathbf{x}_i|y_j)p(y_j)])}{2}, \tag{6}$$

where $p(\mathbf{x}_i|y_j)$ denotes the likelihood of instance $\mathbf{x}_i$ given class y_j, assuming conditional independence among features, and $p(y_i)$ represents the prior probability of class y_i, which is set to $\frac{1}{C}$, where C is the number of classes. Intuitively, CLD measures the separation between the likelihood of the true class and the highest likelihood among competing classes. High CLD values indicate that the instance is almost as likely (or even more likely) to belong to a rival class than to its true class, suggesting a high degree of classification difficulty. Conversely, low CLD values indicate a clear preference for the true class, characterizing an easier classification scenario.

In scenarios where the true label of the query instance is unavailable, the adjusted CLD measure replaces the true class likelihood with the highest class likelihood [Ueda et al. 2024]. The resulting formulation is given by:

$$\text{CLD}_{\text{adj}}(\mathbf{x}) = \frac{1 - \left(\max_{y_i}[p(\mathbf{x}|y_i)p(y_i)] - \max_{y_j \neq y_i}[p(\mathbf{x}|y_j)p(y_j)]\right)}{2}. \tag{7}$$

Rather than comparing the likelihood of the true class against the most competitive alternative, CLD_{adj} relies on the two classes with the highest posterior likelihoods. Consequently, high values indicate that the instance can be plausibly associated with multiple classes, suggesting greater classification difficulty, whereas low values reflect a clear preference for a single class and, therefore, an easier classification scenario.

3.3. Tree-Based Category

Measures in this category rely on a Decision Tree (DT) [Quinlan 1986] induced from the labeled dataset D_l. Traditional formulations assume that the query instance is part of D_l and therefore has a known class label, whereas adjusted variants relax this assumption to accommodate unlabeled instances.

The **Disjunct Class Percentage (DCP)** measure evaluates the purity of the leaf node in which an instance is classified by a pruned DT. This leaf node, commonly referred to as a *disjunct*, represents the final partition reached by the instance during the tree traversal process. DCP is defined as:

$$\text{DCP}(\mathbf{x}_i, y_i) = \frac{|\{\mathbf{x}_j \in \text{Disjunct}(\mathbf{x}_i) \mid y_j = y_i\}|}{|\text{Disjunct}(\mathbf{x}_i)|}, \tag{8}$$

where $\text{Disjunct}(\mathbf{x}_i)$ denotes the leaf node containing $\mathbf{x}_i$. This measure quantifies the proportion of instances within the disjunct that share the same label as the query instance [Smith et al. 2014]. Consequently, high DCP values indicate that the instance belongs to

a relatively pure leaf node dominated by its own class, suggesting a lower classification difficulty. Conversely, low DCP values indicate that the disjunct contains a substantial number of instances from competing classes, characterizing a more difficult classification scenario. Since all hardness measures considered in this study are standardized so that larger values correspond to greater difficulty, the DCP score is inverted accordingly.

In the adjusted DCP formulation, the true label of the query instance is unknown, making it impossible to determine the proportion of instances within the disjunct that belong to the same class as the query instance [Ueda et al. 2024]. Instead, the class heterogeneity of the disjunct is quantified through Shannon's entropy, yielding the measure:

$$\text{DCP}_{\text{adj}}(\mathbf{x}) = -\sum_{i=1}^{C} p(y_j = c_i) \log p(y_j = c_i), \text{ for } \mathbf{x}_j \in \text{Disjunct}(\mathbf{x}). \tag{9}$$

The **Disjunct Size (DS)** measure evaluates instance hardness based on the size of the leaf node (*disjunct*) to which an instance is assigned in an unpruned DT. Unlike DCP, which focuses on the class purity of the disjunct, DS is motivated by the observation that decision trees grown without pruning often produce small disjuncts as a result of repeatedly partitioning the data until little or no information gain remains [Smith et al. 2014]. These small regions frequently correspond to difficult or atypical instances. DS is:

$$\text{DS}(\mathbf{x}_i, y_i) = \frac{|\text{Disjunct}(\mathbf{x}_i)| - 1}{\max_k(|\text{Disjunct}(\mathbf{x}_k)|) - 1}, \tag{10}$$

where $\text{Disjunct}(\mathbf{x}_i)$ denotes the leaf node containing $\mathbf{x}_i$. The underlying assumption of DS is that difficult instances require a longer sequence of decision boundaries to be correctly classified, thereby isolating them into smaller disjuncts. Consequently, instances belonging to small leaf nodes are considered harder to classify than those located in larger and more representative regions of the feature space. Since all measures considered in this study are standardized so that larger values indicate greater hardness, the DS score is inverted accordingly.

The adjusted DS measure is computed analogously to its traditional counterpart. However, because the query instance is unlabeled, it is excluded from the training set used to induce the unpruned DT. The hardness score is then calculated with Equation 10.

The depth at which an instance is classified within a DT can provide valuable information about its classification difficulty [Smith et al. 2014]. The underlying intuition is that easy instances tend to be separated from other classes after only a few decision splits and are therefore assigned to leaf nodes located near the root of the tree. In contrast, more difficult instances typically require a longer sequence of partitioning decisions before reaching a leaf node, placing them deeper in the tree. Based on this rationale, **Tree Depth (TD)** uses the depth of the leaf node associated with an instance to indicate its hardness. In this study, both pruned and unpruned decision trees were considered, following the formulation presented below:

$$\text{TD}(\mathbf{x}_i, y_i) = \frac{\text{depth}_{\text{DT}}(\mathbf{x}_i)}{\max(\text{depth}_{\text{DT}}(\mathbf{x}_j \in D_l))}. \tag{11}$$

The adjusted variants of TD are computed analogously to their traditional counterparts. The only difference is that the training set excludes the query instance, eliminating

the need for its true label. Consequently, the hardness of an unlabeled instance is estimated from the depth of the leaf node to which it is assigned, relative to the maximum depth observed in the DT.

4. Materials and Methods

This study adopts a Pool-Based Active Learning (PAL) strategy, following the experimental protocol proposed by [Nogueira et al. 2025]. The process begins with a small labeled set D_l, which is iteratively expanded by querying informative instances from the unlabeled pool D_u. Initially, D_l contains C randomly selected instances, one from each class. A classification model is trained using this initial set, and its predictive performance is evaluated on a held-out test set. At each iteration, a query strategy is employed to select one instance from D_u for annotation by the oracle. The newly labeled instance is then incorporated into D_l, the classifier is retrained, and its performance is re-evaluated. This procedure is repeated until all instances in D_u have been queried and transferred to D_l.

To investigate the effectiveness of different query criteria, the following instance selection strategies are considered: (i) **Random Sampling**: an instance is selected uniformly at random from D_u; (ii) **Uncertainty Sampling**: instances are selected according to the uncertainty-based criteria described in Section 2; (iii) **Traditional Hardness Sampling**: following [Nogueira et al. 2025], the classifier trained on D_l first predicts the labels of the instances in D_u. These predicted labels are then used to compute the traditional IHMs, and the instance with the highest estimated hardness is selected for querying; (iv) **Label-Agnostic Hardness Sampling**: the adjusted IHMs are directly computed for the instances in D_u, without requiring their labels. We then select the instance with the highest estimated hardness for querying.

We conducted the experiments using datasets from the *UCI Machine Learning Repository*. This study considered 18 datasets. These datasets vary considerably in dimensionality, number of instances, and application domains, and are summarized in Table 1. We partitioned all datasets using 5-fold stratified cross-validation. At each iteration, one fold was held out for testing, while the remaining four folds formed the training data from which the labeled set D_l and the unlabeled pool D_u were derived. Initially, D_l consisted of C randomly selected instances, one representative from each class, and D_u comprised all remaining training instances. The processes are repeated five times, resulting in 25 runs per dataset.

The experiments use XGBoost as the base classifier because of its strong performance and robustness on tabular datasets [Shwartz-Ziv and Armon 2022]. We implemented all models in Python (version 3.14.3) using the official XGBoost package (version 3.2.0) and scikit-learn (version 1.8.0) with default hyperparameter settings. Predictive performance was evaluated on the test folds using Cohen's Kappa coefficient, which measures the agreement between predicted and true labels while accounting for chance agreement [Cohen 1960]. We chose Cohen's Kappa over standard accuracy or F1-score because it provides a more robust evaluation in multi-class and class-imbalanced scenarios. Learning curves were generated by tracking the evolution of the Kappa coefficient as the size of the labeled set D_l increased throughout the AL process. To summarize the overall effectiveness of each sampling strategy, the Area Under the Learning Curve (AUC) was computed from the Kappa values obtained across the AL iterations. This mea-

Table 1. Characteristics of the datasets used in the experiments.

Dataset	Dimensionality	Number of Instances	Number of Classes
lymphography	18	148	4
iris	4	150	3
hepatitis	19	155	2
wine	13	178	3
parkinsons	22	195	2
sonar	60	208	2
seeds	7	210	3
glass	9	214	6
spect-heart	22	267	2
statlog-heart	13	270	2
heart-disease-cleveland	13	303	2
haberman	3	306	2
vertebra-column-2c	6	310	2
vertebra-column-3c	6	310	3
ecoli	7	336	8
ionosphere	34	351	2
libras_move	90	360	15
voting	16	435	2

sure provides a single scalar value that captures the quality of the learning process over time, enabling a straightforward ranking and statistical comparison of the evaluated methods. Since each experiment was repeated multiple times, the reported results correspond to the average AUC across all runs.

5. Results and Discussion

This section presents and discusses the experimental results. First, we compare hardness sampling strategies using traditional and adjusted IHMs to assess the impact of removing dependence on class labels. Subsequently, we compare the hardness sampling approach against baseline instance selection methods commonly adopted in the literature.

5.1. Comparing hardness sampling approaches

Figure 1 presents a comparison between the AUC values achieved by the XGBoost classifier when hardness sampling is guided by traditional and adjusted IHMs. The x-axis represents the AUC obtained using the traditional metric, whereas the y-axis corresponds to the AUC achieved using its adjusted counterpart. Each point represents a dataset.

As shown in Figure 1, the adjusted kDN metric yields results that closely match those obtained with the traditional kDN. This observation is corroborated by the strong correlation between the two versions (shown in the top left of the plots), suggesting that the proposed adjustment has little effect on kDN's ability to identify informative samples for AL. In contrast, the remaining metrics exhibit greater dispersion, with their traditional versions achieving higher AUC values across more datasets.

To further compare the adjusted IHMs, we applied the Friedman test for multiple comparisons [Demšar 2006] to the AUC values across the 18 datasets. The test revealed statistically significant differences among the evaluated metrics at the 95% confidence level. Figure 2 presents the corresponding critical difference diagram obtained using the

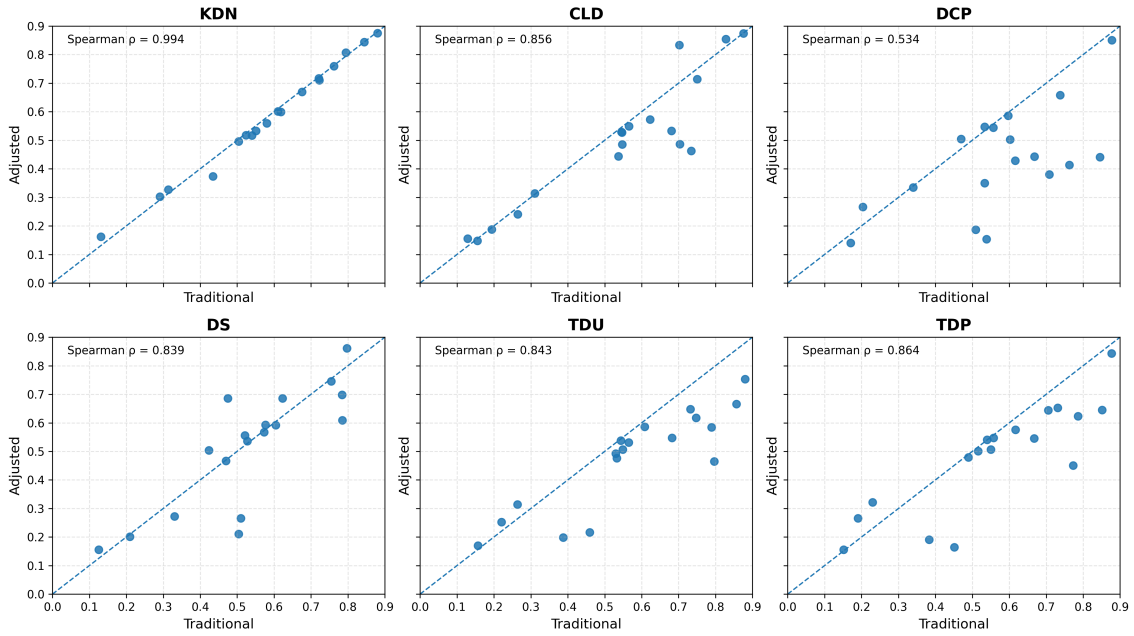


Figure 1. Comparison of traditional (x-axis) versus adjusted (y-axis) hardness samplings concerning the final AUC measured.

Nemenyi post-hoc test. The results indicate that the adjusted kDN achieved the best average rank and significantly outperformed most of the remaining adjusted IHMs. The only metric that did not differ significantly from kDN was DS. Given its superior ranking and its strong agreement with the traditional kDN, the subsequent analyses focus exclusively on the kDN measure.

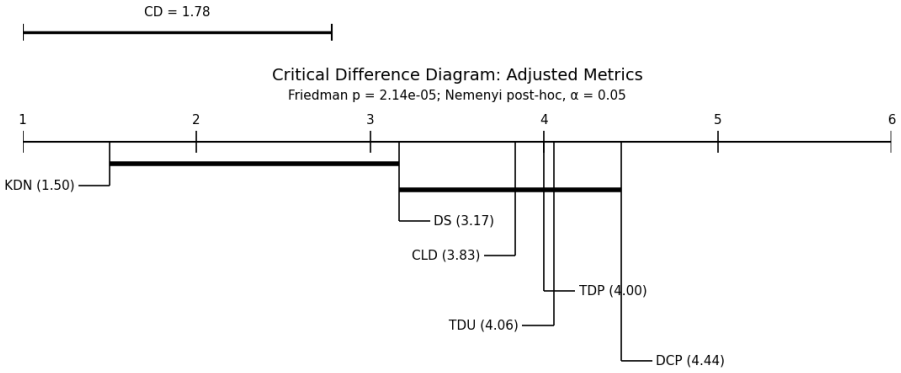


Figure 2. Critical Diagram plot for the Friedman test with Nemenyi post-test comparing the AUCs achieved with the adjusted IHMs.

A similar statistical analysis conducted on the traditional IHMs also identified kDN as the most effective metric for maximizing AUC in active learning. However, in that case, the differences between kDN and most of the remaining traditional IHMs were not statistically significant, except for DS. These findings further reinforce the robustness of kDN as a hardness measure and support its adoption for the remainder of the experimental evaluation. Previous work by [Nogueira et al. 2025] also identified kDN as the most effective metric for hardness sampling. This study extends this finding by showing

that the adjusted kDN achieves performance comparable to traditional kDN within the AL framework. These findings are particularly relevant because the adjusted metrics do not rely on class labels, making them better suited to realistic AL scenarios in which only a small fraction of the data is initially labeled. Moreover, unlike traditional hardness measures, the adjusted metrics do not require training a predictive model to estimate labels before computing hardness values, thereby reducing computational cost while maintaining competitive performance.

5.2. Comparison to baselines

Table 2 presents the average AUC results obtained for each dataset using kDN_{adj}, random sampling and uncertainty sampling strategies. The best AUC value for each dataset is highlighted in bold. The last row reports the average rank each method achieved across all datasets. kDN_{adj} obtained the best average rank, followed by Random Sampling. The strong performance of Random Sampling in active learning settings was also reported by [Nogueira et al. 2025]. Nevertheless, the Friedman test did not detect statistically significant differences among the evaluated strategies.

Table 2. Datasets and baseline performance results

Dataset	Random	LS	MS	Entropy	kDN_{adj}
lymphography	0.5971	0.6064	**0.6148**	0.6007	0.5991
iris	**0.8841**	0.7264	0.7077	0.7434	0.8070
hepatitis	0.2952	0.2346	0.2403	0.2345	**0.3025**
wine	0.8411	0.8305	0.8264	0.8104	**0.8443**
parkinsons	0.5827	0.5515	0.5494	0.5474	**0.6010**
sonar	**0.4913**	0.1990	0.1956	0.2050	0.3736
seeds	**0.7955**	0.7043	0.7158	0.7204	0.7595
glass	0.5029	0.5384	0.5311	**0.5387**	0.5164
spect-heart	**0.3667**	0.3057	0.3147	0.3105	0.3271
statlog-heart	0.5361	0.5347	**0.5462**	0.5431	0.5333
heart-disease-cleveland	0.5453	**0.5725**	0.5613	0.5661	0.5599
haberman	**0.1671**	0.1244	0.1550	0.1330	0.1623
vertebra-column-2c	**0.5572**	0.1584	0.1576	0.1543	0.4964
vertebra-column-3c	0.6686	0.6574	**0.6824**	0.6821	0.6691
ecoli	0.6875	0.7274	0.7294	**0.7304**	0.7174
ionosphere	0.7205	0.7408	**0.7500**	0.7443	0.7099
libras_move	0.5030	**0.5520**	0.5396	0.5500	0.5181
voting	0.8679	0.8781	0.8739	0.8755	**0.8799**
Ranking	2.94	3.28	3.00	3.06	**2.72**

For some datasets, AUC drops substantially when uncertainty-based selection methods are compared with Random Sampling. The *sonar* dataset provides a notable example of this behavior. To better understand this phenomenon, Figure 3 shows how predictive performance evolves as additional instances are selected during the AL process for all evaluated sampling strategies. The learning curves indicate that uncertainty-based methods tend to perform poorly during the initial stages of the process. A plausible explanation is that, at this stage, only a very small number of labeled instances is available, making uncertainty estimates unreliable. The AL protocol adopted in this work is particularly conservative, starting with only one labeled instance per class. Consequently, the

learning task is highly challenging in the early iterations, especially for strategies that rely on an underlying classifier's predictions to guide sample selection.

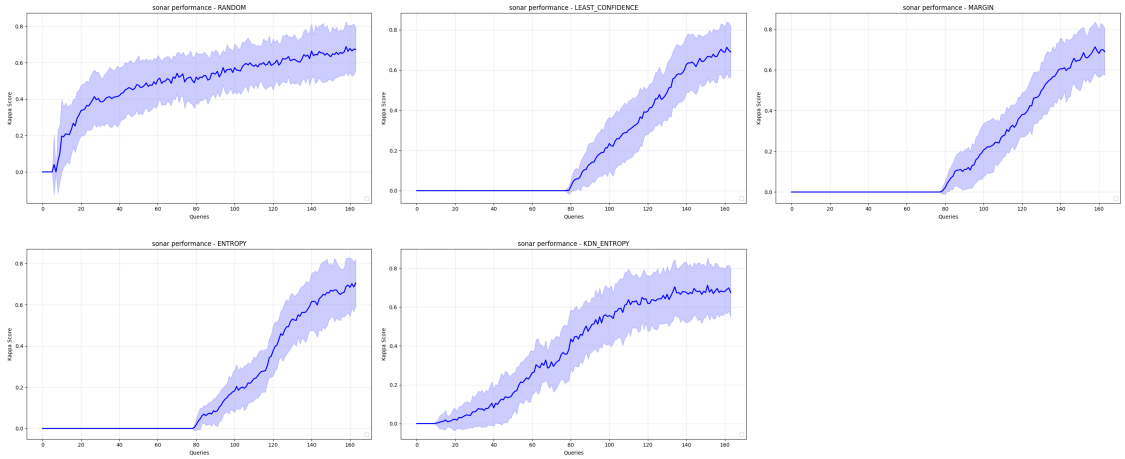


Figure 3. Graph of the Kappa score in relation to the number of instances for the sonar dataset. The shaded areas represent the standard deviation across 5-fold cross-validation.

These results highlight a well-known limitation of uncertainty-based sampling known as the *cold start* problem. Because the process begins with an extremely small labeled set (D_l containing only one instance per class), the classifier's probability estimates are initially uncalibrated and highly unreliable. Consequently, uncertainty-based strategies query uninformative or noisy instances during early iterations, leading to poor initial performance (as observed in the *sonar* dataset).

While the experimental evaluation covers 18 diverse datasets and 25 repetitions per dataset, some limitations should be noted. First, we used XGBoost as the sole base classifier. Although XGBoost is a state-of-the-art model for tabular data, the extent to which these findings generalize to other model families (such as Neural Networks or Random Forests) requires further empirical validation. Second, the choice of initial seed and the conservative initialization protocol (1 sample per class) strongly influence early model stability, which may impact the relative ranking of model-dependent sampling strategies.

6. Conclusion and Future Work

This paper introduced a hardness sampling strategy for active learning based on adjusted IHMs that can be computed without labels for query instances. The experimental results indicate that the adjusted kDN measure is a promising label-agnostic criterion for guiding instance selection, achieving competitive overall ranking compared to traditional AL strategies without requiring prior label knowledge for query instances.

As future work, we plan to extend the experimental evaluation to a larger and more diverse collection of datasets. In addition, we intend to investigate alternative initialization strategies that replace the current one-instance-per-class approach with a more representative subset of the data. Such an initialization procedure may provide a stronger foundation for the learning process, potentially improving the effectiveness and stability of AL strategies during the early stages of training. Furthermore, while this study evaluated performance using XGBoost due to its recognized effectiveness on tabular data,

future work will extend this framework to additional classifier architectures to evaluate model independence across different inductive biases.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, and by the São Paulo Research Foundation (FAPESP), Brasil. Process Number #2021/06870-3.

References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12.

Lorena, A. C., Paiva, P. Y., and Prudêncio, R. B. (2024). Trusting my predictions: on the value of instance-level analysis. *ACM Computing Surveys*, 56(7):1–28.

Monarch, R. M. (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Nogueira, G. S. C. d., P. dos Santos, D., CP L F. de Carvalho, A., and PF Garcia, L. (2025). Hardness sampling: Exploring instance hardness in pool-based active learning. In *Brazilian Conference on Intelligent Systems*, pages 3–17. Springer.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

Scheffer, T., Decomain, C., and Wrobel, S. (2001). Active hidden markov models for information extraction. In *Proceedings of the International Conference on Advances in Intelligent Data Analysis*, IDA '01, pages 309–318, Berlin, Heidelberg. Springer-Verlag.

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

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information fusion*, 81:84–90.

Smith, M. R., Martinez, T., and Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Machine learning*, 95(2):225–256.

Ueda, P. S., Rivolli, A., and Lorena, A. C. (2024). An instance level analysis of classification difficulty for unlabeled data. In *Brazilian Conference on Intelligent Systems*, pages 141–155. Springer.