Active Learning: popular approaches and benchmark data

Amitesh Kumar* Vineet Karya* Alexandre C. B. Delbem
Eric K. Tokuda

*Equal contribution

1 Introduction

Active learning is a powerful approach for reducing the labeling effort required to train effective models, especially in domains such as computer vision and natural language processing where large amounts of unlabeled data are often available [Emam et al. 2021]. Unlike traditional supervised learning, where examples are randomly selected for annotation, active learning algorithms can intelligently choose which samples to label based on their expected informativeness. This is especially relevant for tasks like named entity recognition, where the annotation process can be highly labor-intensive [Shen et al. 2018]. Several key research areas within active learning have been explored, including the development of effective acquisition functions for selecting the most informative samples, theoretical analysis of active learning algorithms, and the study of stopping criteria for sequential data acquisition.

The methods employed in active learning research span a variety of techniques, including uncertainty sampling, where the model selects the examples it is least confident about, and diversity-based sampling, which aims to select a diverse set of examples to expand the model's understanding of the data distribution. Several benchmarks and metrics have been developed to evaluate the performance of active learning algorithms across different tasks and datasets. The choice of appropriate benchmarks and evaluation metrics is crucial for assessing the effectiveness of active learning algorithms and comparing their performance. Active learning has been applied in a wide range of domains, including image classification [Shen et al. 2018], anomaly detection, and natural language processing tasks such as named entity recognition.

In this survey, we review the latest research on active learning benchmarks and techniques across a variety of tasks and domains. When evaluating the performance of active learning algorithms, researchers have utilized a range of benchmark datasets and evaluation metrics to assess their effectiveness. As active learning is a rapidly evolving field, it is important to continuously evaluate and compare the performance of new algorithms against established benchmarks.

2 Methods

In terms of AL sampling methodologies, pool-based AL query strategies can be loosely categorized into 3 categories. First, uncertainty-based sampling strategies aim to select the unlabeled data samples with lowest confidence (largest uncertainty) for the model to be classified correctly, such as least confidence, margin/ratio of confidence, or entropy-based. Second, diversity/representative sampling strategies pick data that contains diversity information of the data pool to lessen the limitations on the supervised machine learning models from data. e.g., outlier detection, cluster-based sampling, representative/density-based sampling. Third, advanced/combined techniques merge the advantages of uncertainty-based and diversity-based criteria, and are extensively accepted in AL and its applications since they are more flexible to diverse data sources.

2.1 Explanation of Sets and Conventions Used

- **Sets and Data Pools:**
  - D_l: Labeled set containing pairs (x, y).
  - D_u: Unlabeled set consisting of samples x without labels.

– D_q: Batch of samples selected for labeling.

- **Parameters and Symbols:**
  - x, x_i, x_j: Data samples; x denotes a general sample, x_i, x_j specific instances.
  - $\text{Rep}(x)$: Representativeness measure.
  - $\text{Div}(x)$: Diversity measure.
  - $\text{Sim}(x, x^{(u)})$: Similarity between x and $x^{(u)}$.
  - $\Delta(x_i, x_j)$: Distance metric.
  - β: Weight parameter balancing uncertainty and representativeness.
  - $\ell(f, x, y)$: Loss function evaluating prediction error.
  - $\text{MMD}(D, D_l \cup D_q)$: Maximum Mean Discrepancy for BMDR.
  - $\Omega(f)$: Regularizer for SPAL.

Strategy	**Optimization Function**	**Description**
Least Confidence	$x^* = \arg\max_x (1 - p(y\|x;\theta))$	Selects samples x where the classifier has the least confidence in the most likely class y.
Margin Sampling	$x^* = \arg\min_x (p(y_1\|x;\theta) - p(y_2\|x;\theta))$	Chooses samples x with the smallest margin between the top two predicted classes y_1 and y_2.
Entropy Sampling	$x^* = \arg\max_x \left(-\sum_i p(y_i\|x;\theta) \log p(y_i\|x;\theta)\right)$	Selects samples with the highest entropy over all classes y_i, indicating maximum uncertainty.
Query-by-Committee	$x^* = \arg\max_x \left(-\sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}\right)$	Uses a committee of models to choose samples with the most disagreement among the models.
Expected Model Change	$x^* = \arg\max_x (p(y_i\|x;\theta)\|\nabla \ell(x,\theta)\|)$	Selects samples that would cause the largest change in model parameters if labeled.
Expected Error Reduction	$x^* = \arg\min_x \left(\sum_i p(y_i\|x;\theta) \cdot \text{Loss}(x, \theta^+)\right)$	Chooses samples expected to reduce classification error the most when labeled.
Variance Reduction	$x^* = \arg\min_x \sigma^2$	Selects samples that minimize model variance σ^2.
Density-Weighted Uncertainty Sampling	$x^* = \arg\max_x (\phi(x) \cdot \text{Sim}(x))$	Balances uncertainty with density, prioritizing samples that are uncertain and representative.

Table 1: Uncertainty-Based Sampling Strategies

Strategy	**Optimization Function**	**Description**
Cluster-based Sampling (KCenter)	$x^* = \arg\min_{C:\|C\|<b} \max_i \min_{j\in C} \Delta(x_i, x_j)$	Selects a subset C that minimizes maximum distance $\Delta(x_i, x_j)$ to ensure diversity.
Outlier Detection	$x^* = \arg\max_x \text{OD}(x)$	Chooses samples with high outlier scores $\text{OD}(x)$ to avoid redundancy.
Representative Sampling (Density-Based)	$x^* = \arg\max_x \left(\frac{1}{U} \sum_{u=1}^{U} \text{Sim}(x, x^{(u)})\right)$	Selects representative samples maximizing average similarity with all samples.
Hierarchical Clustering	$x^* = \arg\max_x (\text{Rep}(x) \cdot \text{Div}(x))$	Combines representativeness $\text{Rep}(x)$ and diversity $\text{Div}(x)$ for balanced coverage.

Table 2: Diversity-Based Sampling Strategies

Strategy	Optimization Function	Description
Graph Density	$x^* = \arg\min_x (\beta \cdot \text{Rep}(x) + (1-\beta) \cdot \text{Unc}(x))$	Balances representativeness $\text{Rep}(x)$ and uncertainty $\text{Unc}(x)$ using β.
AAL	$x^* = \arg\max_x \left(f(x)^\beta \cdot d(x)^{1-\beta}\right)$	Adapts based on uncertainty $f(x)$ and diversity $d(x)$, with β controlling trade-off.
BMDR	$\min_{D_q, f} \left(\sum_{x,y \in D_l} \ell(f,x,y) + \sum_{x \in D_q} \ell(f,x,\hat{y})\right) + \lambda\|\|f\|\|^2 + \beta\text{MMD}(D, D_l \cup D_q)$	Minimizes loss by selecting batches that maximize both informativeness and representativeness.
Self-Paced AL(SPAL)	$\min_{f,w,v} (\ell(f,w,v) + \lambda g(v) + \mu h(D_l \cup D_q, D_u/D_q) + \gamma\Omega(f))$	Prioritizes samples based on informativeness, representativeness, and ease of learning.

Table 3: Hybrid Sampling Strategies

3 Datasets

To rigorously assess the efficacy of active learning methods across multiple tasks, different studies employ different datasets spanning across distinct domains. These datasets exhibit varying levels of complexity, particularly with respect to their attribute dimensions, sample sizes, and class distributions, making them well-suited for benchmarking diverse active learning techniques. Below, we provide a detailed description of the datasets utilized in each task category.

3.1 Tabular Classification Datasets

For the tabular classification task, [Zhan et al. 2021] utilize a selection of both real and synthetic datasets, primarily obtained from well-known repositories such as UCI and KEEL. These datasets vary in feature dimensionality (d), the number of samples (n), and the number of class categories (K), thus offering a diverse range of challenges for active learning models. Table 4 summarizes the key properties of these tabular classification datasets.

3.2 Entity Matching Datasets

For the entity matching task, [Meduri et al. 2020] leverage several real-world datasets that capture the complexity of matching entities across disparate sources. These datasets differ significantly in the number of attributes, total entity pairs, post-blocking pairs, and class skew. Such variation poses substantial challenges for maintaining balance in the training data during active learning. Table 5 provides a comprehensive summary of these entity matching datasets.

3.3 Anomaly Detection Datasets

For the anomaly detection task, [Luo et al.] employ a suite of well-established datasets known for their varying dimensionalities and proportions of anomalies. These datasets are particularly useful for evaluating how well active learning techniques can detect outliers in complex, high-dimensional feature spaces. Table 6 summarizes the key characteristics of these datasets.

4 Experiments

This section presents the experimental setups and results from benchmarking various active learning methods across different tasks. The experiments were conducted to evaluate the performance of different active learning methods on diverse datasets.

Dataset	Type	d	n	K	Source
Breast Cancer	Real	10	478	2	UCI
Appendicitis	Real	7	106	2	KEEL3
Heart	Real	13	270	2	UCI
Haberman	Real	3	306	2	UCI
Diabetes	Real	8	768	2	UCI
Statlog (Australian)	Real	14	690	2	UCI
Sonar	Real	60	108	2	UCI
Ionosphere	Real	34	351	2	UCI
Statlog (German)	Real	20	1,000	2	UCI
MUSK (Clean1)	Real	168	475	2	UCI
Molecular Biology (Splice)	Real	61	1,000	2	UCI
Iris	Real	4	150	3	UCI
Wine	Real	13	178	3	UCI
Thyroid	Real	5	215	4	UCI
Statlog (Vehicle)	Real	18	946	4	UCI
EX8a	Synthetic	2	863	2	ML Course
EX8b	Synthetic	2	206	2	ML Course
Gaussian Cloud Balance	Synthetic	2	1,000	2	-
Gaussian Cloud Unbalance	Synthetic	2	1,000	2	-
XOR (Checkerboard2×2)	Synthetic	2	1,600	2	-
Parkinson	Real	22	195	2	UCI
Seeds	Real	7	210	3	UCI
Glass	Real	9	214	7	UCI
Wdbc	Real	30	569	2	UCI
Statlog (Vehicle)	Real	18	946	4	UCI
Tic-Tac-Toe	Real	9	958	2	UCI
Phishing Websites	Real	30	2456	2	UCI
D31	Synthetic	1	3100	31	-
Spambase	Real	57	4601	2	UCI
Banana	Synthetic	2	5300	2	-
Phoneme	Real	5	5404	2	ELENA Project
Texture	Real	40	5500	11	UCI
Ringnorm	Synthetic	21	7400	2	Leo Breiman
Twonorm	Synthetic	50	7400	2	Leo Breiman

Table 4: Summary of tabular classification datasets with their key characteristics.

Dataset Name	Number of Attributes	Total Pairs	Post-Blocking Pairs	Class Skew
Abt-Buy	3	1.18M	8682	0.12
Amazon-GoogleProducts	4	4.39M	14294	0.09
DBLP-ACM	4	6M	11194	0.198
DBLP-Scholar	4	168M	49042	0.19
Cora	9	0.97M	114525	0.124
Walmart-Amazon	7	56.37M	13843	0.083
Amazon-BestBuy	3	21.29M	395	0.147
BeerAdvocate-RateBeer	4	13.03M	450	0.151
BuyBuyBaby-BabiesRUs	7	54.5M	400	0.27

Table 5: Summary of entity matching datasets with their key characteristics.

Data	n	d	# Anomaly	% Anomaly
ALOI	49534	27	1508	3.04
InternetAds	1966	1555	368	18.72
fault	1941	27	673	34.67
letter	1600	32	100	6.25
magic.gamma	19020	10	6688	35.16
mammography	11183	6	260	2.32
satellite	6435	36	2036	31.64
Waveform	3443	21	100	2.90
yeast	1484	8	507	34.16

Table 6: Summary of anomaly detection datasets with their key characteristics.

4.1 Tabular Classification

4.1.1 Datasets

[Zhan et al. 2021] included 35 varied datasets from the UCI Machine Learning Repository, spanning both actual and synthetic datasets. These datasets include well-known examples such as Breast Cancer, Sonar, Iris, and MUSK (Clean1). The datasets were selected to illustrate a wide range of difficulties, including varying feature dimensions, sample sizes, and class distributions. Specifically, some datasets like Iris and Seeds have lower feature dimensions and fewer class categories, while others like MUSK (Clean1) and Phoneme depict high-dimensional and more complex scenarios. Such diversity in the datasets is critical for testing the durability of active learning strategies across a broad spectrum of classification tasks.

4.1.2 Experimental Setup

They utilized a support vector machine (SVM) with a radial basis function (RBF) kernel as the base classifier. Each dataset was separated into a training set (60%) and a test set (40%). The training procedure was iterative, including active learning approaches where only a part of the training data was initially labeled. At each iteration, a batch of the most informative samples from the unlabeled pool was picked for labeling. The major assessment metric was the Area Under the Budget Curve (AUBC), which captures model performance at different stages of the query budget, thereby providing a comprehensive perspective of the classifier's performance as additional samples are labeled. To establish robustness, each experiment was performed numerous times (100 trials for datasets with fewer than 2000 samples, and 10 trials for bigger datasets) with random splits of training and test sets. The findings were averaged to minimize the impact of variance in performance owing to random initialization.

4.1.3 Active Learning Methods

They focused primarily on uncertainty sampling as the active learning technique, where the model selects data points for labeling based on the lowest categorization confidence. Variants such as least confident sampling, margin sampling, and entropy-based sampling were applied. Additionally, approaches like k-Center, which assures diversity by selecting samples that are maximally far from previously labeled points, were applied to offset the limits of uncertainty sampling in circumstances when the selected samples lack diversity. Furthermore, Query-by-Committee (QBC), which maintains a collection of classifiers and picks samples based on the disagreement among them, was employed for more complex datasets to balance uncertainty and diversity.

4.1.4 Evaluation Metrics

The key assessment tool utilized was the Area Under the Budget Curve (AUBC), which provides a holistic perspective of model performance across varied labeling budgets. A higher AUBC value suggests that a model performs better with fewer labeled examples, therefore indicating a more efficient

active learning technique. The performance of the models was examined not only at full budget but also across intermediate budget phases to highlight the label efficiency of different active learning methodologies. Additionally, in some cases, accuracy and F1-score were also tracked to verify that increases in AUBC were not obtained at the price of overall predictive performance.

4.1.5 Key Results

The trials demonstrated that uncertainty sampling and k-Center performed better on simpler datasets such as Iris, where the feature space is narrower and decision boundaries are more defined. However, these approaches struggled with more complex and high-dimensional datasets like MUSK (Clean1) and Ringworm. In contrast, Query-by-Committee regularly outperformed other techniques on complex datasets by properly balancing uncertainty and variety. It was observed that diversity-based methods like k-Center were particularly effective during the early stages of labeling when the labeled set is small, while uncertainty-based methods became more useful as the decision boundary became more defined with an increased number of labeled examples. Advanced methods like self-paced active learning (SPAL) also exhibited high performance in scenarios with uneven class distributions by dynamically modifying query criteria depending on both informativeness and ease of classification.

Dataset	**RS**	**BSO**	**Avg**	**Best**	**Worst**
Appendicitis	0.836	0.881	0.844	EER	DWUS
Sonar	0.617	0.830	0.755	LAL	HintSVM
Iris	0.883	0.910	0.927	BMDR	US
Wine	0.858	0.946	0.900	BMDR	HintSVM
Parkinsons	0.949	0.942	0.901	EER	DWUS
EkbB	0.863	-	0.912	SPAL	DWUS
Seeds	0.862	0.922	0.912	BMDR	QS
Glass	0.696	0.708	0.708	EER	DWUS
Thyroid	0.696	0.745	0.748	EER	QS
Heart	0.691	0.771	0.721	InfDivQ	DWUS
Haberman	0.617	0.620	0.637	BMDR	QS
Ionosphere	0.901	0.951	0.927	LAL	HintSVM
Clean1	0.758	0.923	0.859	SPAL	DWUS
Breast	0.954	0.957	0.965	QS	DWUS
Wbc	0.925	0.951	0.943	LAL	DWUS
R15	0.970	0.958	0.965	QBC	QUIRE
Australian	0.865	0.921	0.875	BMDR	QUIRE
Vehicle	0.567	0.598	0.486	BMDR	SPAL
Tic-tac-toe	0.870	0.831	0.724	QBC	QUIRE
German	0.686	0.730	0.741	QBC	DWUS
Splice	0.860	0.821	0.791	QS	DWUS
Cloudub1	0.942	0.963	0.929	QBC	HintSVM
Cloudub4	0.943	0.941	0.926	QBC	EER
Checkerboard	0.951	0.969	0.964	LAL	VR
Phishing	0.954	-	0.956	KCenter	Graph
D31	0.862	0.922	0.899	KCenter	QS
Spambase	0.920	-	0.919	QBC	DWUS
Banana	0.550	-	0.847	QBC	HintSVM
Phoneme	0.683	-	0.838	Hier	QS
Texture	0.666	-	0.918	Hier	DWUS
Ringnorm	0.666	-	0.919	Hier	DWUS
Twonorm	0.976	-	0.975	KCenter	DWUS

Table 7: AL performance by AUBC(acc) from the dataset aspect. Random Sampling (RS) performance, BSO results, average (Avg) performance of each dataset across 17 AL methods, and the best (Best) and worst (Worst) performing AL methods. Symbol "–" indicates that BSO results are not present.

4.2 Entity Matching

4.2.1 Datasets

[Meduri et al. 2020] included numerous benchmark datasets from the product and publication domains, including Abt-Buy, Amazon-GoogleProducts, DBLP-ACM, and Cora. These datasets are widely utilized for entity matching tasks and represent considerable challenges due to the prevalence of large-scale class imbalance and the need for efficient entity resolution. For instance, datasets such as Abt-Buy and Amazon-GoogleProducts include millions of pairings after blocking, underlining the necessity for active learning to decrease labeling efforts. The datasets are illustrative of real-world circumstances where the goal is to distinguish distinct mentions of the same real-world entity.

4.2.2 Experimental Setup

The tests were conducted using a range of classifiers across multiple learning paradigms, including linear classifiers (e.g., support vector machines), non-linear models (e.g., feed-forward neural networks), tree-based classifiers (e.g., random forests), and rule-based models. Active learning approaches were applied to boost the efficiency of the entity matching process by lowering the labeling burden. Specifically, Query-by-Committee (QBC) and margin-based selection procedures were investigated. The initial seed set comprised of 30 labeled instances, and in each active learning iteration, 10 new examples were queried and labeled. The models were tested on different metrics such as F1-score, to quantify classification performance, and latency, which measured the time taken for example selection and model training.

4.2.3 Active Learning Methods

Active learning methods such as Query-by-Committee (QBC) and margin-based selection were utilized in the context of entity matching. QBC, a widely-used method in active learning, picks examples for classification based on the largest disagreement among a committee of classifiers. This method was particularly helpful in ensuring that diverse and informative samples were selected for labeling, thus increasing model performance on highly imbalanced datasets. Margin-based selection, another famous strategy, was applied to discover samples that sit closest to the decision boundary, therefore balancing both informativeness and computing efficiency. These methods were further refined by using blocking strategies to limit the amount of candidate pairs that need to be assessed.

4.2.4 Evaluation Metrics

The key assessment metric utilized for entity matching was the F1-score, which balances precision and recall, making it a viable measure for both matching and non-matching entity pairs. Given the vast scale of the datasets, latency was also a key statistic, showing the computing efficiency of each active learning method. Latency was quantified as the time spent for both picking instances for labeling and training the model, providing insights into the scalability of the suggested methods. The use of progressive F1-scores allowed the authors to analyze how the performance improved as more cases were categorized.

4.2.5 Key Results

The results suggested that the application of active learning approaches considerably improved the performance of entity matching models. Random forests paired with Query-by-Committee achieved near-perfect F1-scores on large-scale datasets such as DBLP-ACM, illustrating the effectiveness of ensemble-based techniques in extremely imbalanced circumstances. However, this strategy incurred significant latency because to the computing expense associated with committee setup and rating. In contrast, margin-based selection, particularly when applied to linear classifiers, offers a more balanced trade-off between performance and efficiency, making it ideal for scenarios where computing resources are constrained. The results underline the necessity of selecting the proper active learning technique depending on dataset features and computing constraints.

4.3 Anomaly Detection

4.3.1 Datasets

For anomaly detection, datasets from the ADBench suite were used, including SWaT, WADI, and SMD. These datasets span numerous domains such as critical infrastructure, industrial control systems, and system monitoring, and offer varying dimensionalities and percentages of abnormalities. This variation in dataset features provides major hurdles, making them a suitable baseline for evaluating the efficacy of active learning strategies in anomaly detection tasks [Luo et al.].

4.3.2 Experimental Setup

The tests utilized two state-of-the-art anomaly detection models: DevNet, which employs deep neural networks for learning deviations from typical patterns, and XGBOD, an ensemble method combining XGBoost and k-Nearest Neighbors. Active learning methodologies, including Bayesian Active Learning by Disagreement (BALD) and BADGE Sampling, were incorporated with these models. The iterative procedure required querying more data samples depending on a preset budget, with models being modified after each batch of labeled data. Performance was largely evaluated using AUC-ROC and AUC-PR to measure the effectiveness of the anomaly detection models under different active learning strategies.

4.3.3 Active Learning Methods

Hybrid active learning methods such as Loss Prediction, which mixes uncertainty and diversity-based sampling, were utilized to improve anomaly identification. BALD, an uncertainty-based strategy that picks samples maximizing information gain, and BADGE, a diversity-based method that clusters uncertain samples, were also applied to iteratively select data points for labeling. These strategies seek to enhance label efficiency and increase model performance on highly imbalanced anomaly detection datasets.

4.3.4 Evaluation Metrics

The performance of the anomaly detection algorithms was largely tested using AUC-ROC and AUC-PR. AUC-ROC examines the model's ability to differentiate between true positives and false positives, while AUC-PR focuses on the balance between precision and recall, making it particularly useful for highly imbalanced datasets. These metrics enabled a complete evaluation of model performance in detecting anomalies under different active learning conditions.

Strategies	**fault**	**internetads**	**ALOI**	**letter**	**magic**	**mammo**	**satellite**	**wave**	**yeast**
RandomSampling	0.6919	0.7644	0.5112	0.5890	0.8581	0.8967	0.8471	0.8522	0.6924
AdversarialBIM	0.6467	0.8108	0.5681	0.5406	0.8029	0.9126	0.8255	0.5129	0.6906
AdversarialDF	0.6645	0.8780	0.5851	0.6297	0.8367	0.9013	0.8513	0.6850	0.6225
BALDDropout	0.7294	0.7942	0.5581	0.7453	0.8569	0.9276	0.8538	0.8721	**0.6989**
BadgeSampling	0.7149	0.8043	0.5530	0.8130	0.8596	0.9295	0.8383	0.8880	0.6745
EntropySampling	0.6692	0.7609	0.5031	0.8130	0.8390	0.9216	0.8419	0.9126	0.6774
EntropyDropout	0.6692	0.7609	0.5031	0.8130	0.8390	0.9216	0.8419	0.9126	0.6774
KCenterGreedy	0.7434	0.9232	0.5324	0.7332	0.8584	0.9277	0.8497	0.9860	0.6624
KCenterPCA	0.7542	0.9214	0.5703	0.8070	0.8602	0.9147	0.8497	0.5254	0.6586
KMeansSampling	0.6678	0.8407	0.5347	0.6290	0.8422	0.9290	0.8568	0.9170	0.6823
MarginSampling	0.6692	0.7609	0.5185	0.7561	0.8410	0.9195	0.8390	0.9126	0.6774
MarginDropout	0.6692	0.7609	0.5185	0.8410	0.9195	0.8390	0.9126	0.6774	
LossPrediction	**0.8215**	**0.9614**	**0.5728**	0.9062	0.9040	0.9212	**0.9494**	**0.9748**	0.6774
MeanSTD	0.7053	0.7912	0.5565	0.7607	0.8770	0.9275	0.8559	0.8883	0.6911
VarRatio	0.6692	0.7609	0.5031	0.8130	0.9195	0.8410	0.9126	0.6774	
WAAL	0.7850	0.9643	0.5728	**0.9804**	0.9040	**0.9212**	0.9494	0.9444	0.6610

Table 8: AUC-ROC of different query strategies, and baseline (random sampling) on all datasets. Base Model: DevNet. Budget Ratio: 0.5.

4.3.5 Key Results

The results suggested that hybrid tactics like Loss Prediction outperformed classic uncertainty-based methods in detecting abnormalities across all datasets. Notably, BALD and BADGE Sampling displayed increased detection performance in the early phases of the learning process by efficiently picking useful samples. However, the experiments demonstrated declining results as more data was tagged, showing that after a certain point, extra labeling had a limited impact on performance. These findings underscore the necessity of properly managing labeling budgets and selecting appropriate query strategies for maximizing model efficiency and effectiveness in anomaly detection tasks.

References

Zeyad Ali Sami Emam, Hong-Min Chu, Ping-Yeh Chiang, Wojciech Czaja, Richard Leapman, Micah Goldblum, and Tom Goldstein. Active learning at the imagenet scale, 2021. URL https://arxiv.org/abs/2111.12880.

Haoyan Luo, Xiaofan Gui, Wei Cao, and Jiang Bian. Activead: Enhancing anomaly detection in tabular data through active learning strategies.

Venkata Vamsikrishna Meduri, Lucian Popa, Prithviraj Sen, and Mohamed Sarwat. A comprehensive benchmark framework for active learning methods in entity matching. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, pages 1133–1147, 2020.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition, 2018. URL https://arxiv.org/abs/1707.05928.

Xueying Zhan, Huan Liu, Qing Li, and Antoni B Chan. A comparative survey: Benchmarking for pool-based active learning. In *IJCAI*, pages 4679–4686, 2021.