Data-Efficient Automatic Model Selection in Unsupervised Anomaly Detection

Gautham Krishna Gudur, Raaghul R, Adithya K, Shrihari Vasudevan
Global AI Accelerator, Ericsson
gautham.krishna.gudur@ericsson.com

Abstract

Anomaly Detection is a widely used technique in machine learning that identifies context-specific outliers. Most real-world anomaly detection applications are unsupervised, owing to the bottleneck of obtaining labeled data for a given context. In this paper, we solve two important problems pertaining to unsupervised anomaly detection. First, we identify only the most informative subsets of data points to obtain ground truths from the domain expert (oracle); second, we perform efficient model selection using a Bayesian Inference framework and recommend the top-k models to be fine-tuned prior to deployment. To this end, we exploit multiple existing and novel acquisition functions, and successfully demonstrate the effectiveness of the proposed framework using a weighted Ranking Score (η) to accurately rank the top-k models. Our empirical results show a significant reduction in data points acquired (with at least 60% reduction) while not compromising on the efficiency of the top-k models chosen, with both uniform and non-uniform priors over models.

1 Introduction

Anomaly Detection (AD) is the process of identifying unexpected or unforeseen events in data sets of various kinds. Anomalies could also be considered unlikely events with respect to a deterministic threshold, assuming the existence of a distribution of various events. In the past few years, machine learning has led to major breakthroughs in various areas related to automation and digitization tasks, and anomaly detection plays an instrumental role in such tasks.

There have been multiple anomaly detection frameworks conventionally with a rich literature of supervised, unsupervised, and semi-supervised algorithms [1, 19]. In general, an anomaly is a domain/business-specific definition, which evolves over time depending on changes in data distributions, geographical constraints, business contexts, and many more. This makes anomaly detection in any industry a cumbersome task and requires extensive human expertise and domain knowledge to be inculcated in the current frameworks. Currently, domain experts rely on their expertise to decide what is or what is not an anomaly. Moreover, most real-world AD systems across a myriad of use cases like identifying anomalies in resource utilization in a telecommunication network, identifying operational network issues using Quality of Experience (QoE), Quality of Service (QoS), and other factors, identifying abnormal medical conditions amongst patients, extremities in climate change, and many more, have to rely on unsupervised anomaly detection for various reasons like labeling costs from humans and other sources, data privacy issues, and so on.

In addition to detecting anomalies, *choosing the necessary algorithms for the given data in an efficient manner is hard, particularly in an unsupervised setting*. Solution developers predominantly rely on experimentation and/or trial-and-error on the whole data to decide the best approaches for their use cases. The problem at hand is two-fold – first, *efficiently choosing the right subset of informative data points to be identified as anomalies*; second, *choosing algorithms that best fit this data*.

Preprint. Under review.

The main contributions of our paper are, **(1)** We propose a *model selection framework for unsupervised anomaly detection* using *Bayesian Inference*, and propose a novel *ranking criterion* for selecting the best models. **(2)** We address the labeled data scarcity problem in unsupervised anomaly detection via *subset selection* wherein, labels for a small fraction of most-informative data points are acquired from an oracle (human), by exploiting multiple *existing and novel acquisition functions*. **(3)** We benchmark our proposed model selection framework using various standard datasets to showcase its effectiveness in unsupervised anomaly detection settings with both uniform and non-uniform priors over models, which operate in synchronicity with human-augmented user feedback.

We demonstrate our framework as a viable automation approach to developing unsupervised AD solutions for real-world applications with minimal supervision. We discuss more related work in Appendix A.

2 Our Approach

The following are the steps in our proposed unsupervised anomaly detection model selection system.

Algorithm 1 Our Proposed Framework

Input: Train Dataset $\mathcal{D}_{train}$, Total unsupervised anomaly detection models M, Total Bayesian Inference iterations I, Acquisition Function AF, Subset Dataset $\mathcal{D}_{subset}$
Output: Top-k models chosen K, Ranking Score η

Initialize Categorical distribution (likelihood) over M models
Initialize Dirichlet Priors $p_i \sim Dir(\alpha_i), i = 1, .., M$ with uniform/non-uniform concentration α_i
Obtain probabilities from all M unsupervised models with $\mathcal{D}_{train}\{x\}$
Subset selection on $\mathcal{D}_{train}\{x\}$ using AF to obtain $\mathcal{D}_{subset}$
Present $\mathcal{D}_{subset}\{x\}$ to oracle to obtain $\mathcal{D}_{subset}\{x, y\}$
Choose corresponding best model for $\mathcal{D}_{subset}\{x, y\}$
for $i = 1$ **to** I **do**
  Update model posterior $p_i|\alpha_i, c_i \sim Dir(\alpha_i + c_i), i = 1, .., M$ based on best model for $\mathcal{D}^i_{subset}\{x, y\}$
end for
Select top-k models (K) based on the model posterior
for $k = 1$ **to** K **do**
  Calculate F1-score, Accuracy, Average Precision Score,
  AUC ROC, η of model *k*
end for
Return top-k models with best hyperparameters using
Grid Search

2.1 Bayesian Inference Framework for Model Selection

To perform model selection over unsupervised anomaly detection approaches, we propose a Bayesian inference framework using Exact Inference (EI), Stochastic Variational Inference (SVI), or Markov Chain Monte Carlo (MCMC), for modeling posterior probabilities [10]. We do not consider MCMC in this scenario since the time complexity is extremely high, in comparison to SVI and EI.

Given a categorical likelihood distribution over all AD models, if the prior distribution is Dirichlet, then the posterior distribution is also a Dirichlet distribution since the Dirichlet distribution is a conjugate prior for the Categorical/Multinomial distribution [7]. The samples of the Dirichlet posterior would give us probabilities of the Categorical distribution over all the AD models. Hence, it is straightforward to use Exact Inference. If alternative prior distributions are necessitated by the domain, EI may not be feasible and SVI and MCMC options could be explored.

2.1.1 Discussion on Dirichlet Priors

There are multiple ways to incorporate apriori beliefs in the form of priors for the data in consideration. We propose defining priors over anomaly detection approaches based on a taxonomy of, **(1)** Types of

anomalies like point, contextual and collective. **(2)** The given type/distribution of data (tree-based, density-based, etc.). **(3)** By adding domain knowledge on the approaches (along with other meta-data like priors over features), that the user believes will perform well on the given data.

In the scenario where priors in the form of taxonomies are unavailable, we initialize Dirichlet priors over the set of models in consideration to be *uniform*, which is typically the default setting.

The acquisition functions during subset selection (discussed in Section 2.2) – where the user provides feedback on anomalous behavior for a chosen subset of data, also enhances the priors over the anomaly detection approaches for the next iteration. Incorporating such priors improve convergence rate (faster convergence), thereby also potentially reducing the total number of iterations required to obtain efficient posteriors.

2.2 Acquisition Functions for Subset Selection

Acquisition functions are used in subset selection to choose the most informative set of data points to be queried from the overall data D_{train}. We examine and propose the following acquisition functions,

Boundary: This acquisition function selects points that are close to the boundary threshold for each model, and are considered the most uncertain.

$$abs(p_{ij} - threshold)$$

Max Disagreement: Selects data points wherein each model's disagreement against consensus probabilities (mean probabilities across models) is the largest for some learners.

Boundary Max Disagreement: Combines Boundary and Max Disagreement acquisition functions, wherein it first selects the data points that are closest to the boundary threshold, and then selects the points with Max disagreement.

Max Entropy: This acquisition function chooses data points that maximize the predictive entropy.

$$-\sum_{c} p(y = c|x, D_{train}) \log p(y = c|x, D_{train})$$

Variance Entropy: This acquisition function selects data points where the probability distribution across various models has the highest variance.

$$\sigma^2 = \frac{\sum_{j=1}^{M} (p_{ij} - \mu)^2}{M}$$

Random: This acquisition function chooses data points uniformly at random.

2.3 Ranking Score (η)

The ranking score, defined by η, is a position-weighted aggregated similarity metric between the top-k model recommendations of a given subset, and the top-k model recommendations of the entire 100% dataset, i.e., with and without subset selection. The score depends on the presence of a model in top-k, as well as the position of the rank.

$$\eta = 1/k * \sum_{r=1}^{k} r_{subset}/(r_{subset} + abs(r_{subset} - r_{full}))$$

where r_{subset} is the rank of a model for the given acquisition function, while r_{full} is the rank of a model with the entire dataset. η indicates the effectiveness of the top-k model recommendations for different acquisition functions with respect to the entire 100% data.

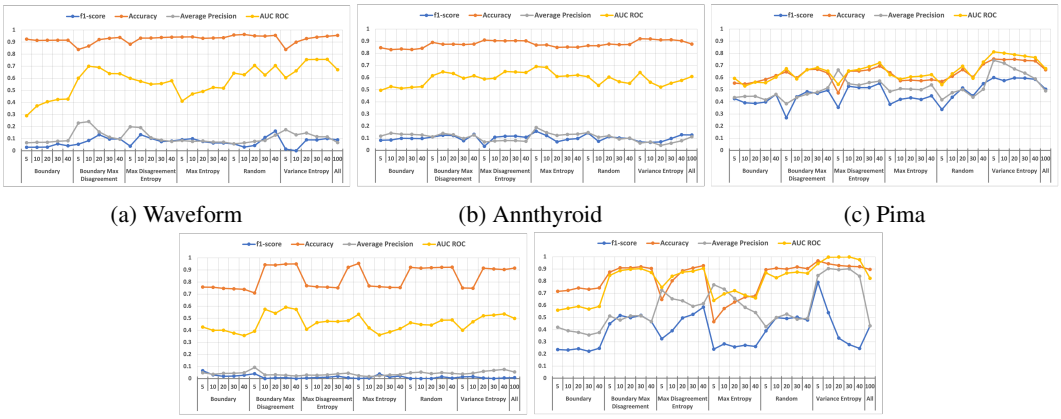
Table 1: Evaluation Metrics for two Bayesian Inference techniques (Exact Inference (EI) and Stochastic Variational Inference (SVI)) for all datasets, each averaged across all Acquisition Functions and subset sizes, along with corresponding times taken.

Dataset	Accuracy (%)		F1-score		Avg Precision		AUC ROC		Time (in sec)	
	EI	SVI	EI	SVI	EI	SVI	EI	SVI	EI	SVI
Waveform	92.395	92.365	0.072	0.062	0.111	0.106	0.577	0.564	0.322	1092.113
Annthyroid	87.621	87.526	0.1	0.102	0.108	0.103	0.591	0.59	0.294	995.956
Pima	63.513	62.762	0.467	0.462	0.52	0.506	0.652	0.661	0.181	614.822
Wilt	83.375	83.265	0.013	0.011	0.041	0.039	0.467	0.458	0.184	630.476
PageBlocks	81.864	81.766	0.394	0.388	0.591	0.588	0.805	0.798	0.217	746.831

3 Experiments and Results

We train and evaluate our experiments on five different anomaly detection datasets as observed in Appendix B Table 3. These datasets from DAMI[1] [4], are often used in the unsupervised anomaly detection benchmarking literature.

The subset selection experiments are performed with combinations of 5 different incremental percentages – 5%, 10%, 20%, 30%, 40%, and 6 acquisition functions – Boundary, Max Disagreement Entropy Boundary, Max Disagreement, Max Entropy, Variance Entropy, Random, as discussed in Section 2.2. The experiments are executed for each combination of subset size (%) and acquisition function. In addition, we perform these experiments with full (100%) data for benchmarking.



(a) Waveform (b) Annthyroid (c) Pima

(d) Wilt (e) PageBlocks

Figure 1: Acquisition Functions (along with corresponding acquisition subset sizes (in %)) vs Evaluation Metrics for all datasets.

To evaluate our proposed model selection framework, we experiment with nine commonly used unsupervised anomaly detection algorithms in literature [9, 4, 20], starting with hyperparameters sampled at random. For consistency, we use the PyOD package [19] for implementing all models . The initial unsupervised AD algorithms used are, **(1)** COF **(2)** IsoForest **(3)** CBLOF **(4)** LOF **(5)** OCSVM **(6)** KNN **(7)** HBOS **(8)** ABOD **(9)** LODA. We choose the top-k models from this initial set of nine AD models.

The Bayesian Inference model selection framework, as discussed in Section 2.1, uses Exact Inference (EI) and Stochastic Variational Inference (SVI) to select the top-5 AD models across different subset sizes (%) and acquisition functions. Here, we perform our experiments with both uniform and non-uniform Dirichlet priors over all AD models. We use the Pyro package [2] for our Bayesian Inference experiments. Further, we perform hyperparameter tuning for the top-5 models selected with maximum Average Precision Score [4] as the criterion to choose the best hyperparameters. Our experiments are performed on an 8-Core Intel Core i9 @ 2.3 GHz, with 16 GB memory.

[1] https://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI

Table 2: Acquisition Functions vs top-5 recommended Anomaly Detection models along with their respective Ranking Scores for all datasets with their best corresponding subset sizes, set against 100% data with no acquisition criteria.

Dataset with Best Subset Size	Acquisition Criterion	Ranking (top-5) 1	2	3	4	5	Ranking Score (η)
Waveform (100% data)	No Acquisition	LOF	LODA	OCSVM	HBOS	ABOD	–
Waveform with 40% subset	Boundary	LOF	LODA	OCSVM	HBOS	ABOD	**1.0**
	Boundary Max Disagreement	LOF	LODA	OCSVM	HBOS	ABOD	**1.0**
	Max Disagreement Entropy	LOF	LODA	OCSVM	HBOS	ABOD	**1.0**
	Max Entropy	LOF	LODA	OCSVM	HBOS	ABOD	**1.0**
	Random	LOF	LODA	OCSVM	HBOS	ABOD	**1.0**
	Variance Entropy	LOF	LODA	OCSVM	HBOS	ABOD	**1.0**
Annthyroid (100% data)	No Acquisition	LOF	OCSVM	ABOD	KNN	HBOS	–
Annthyroid with 30% subset	Boundary	LOF	ABOD	OCSVM	KNN	HBOS	0.883
	Boundary Max Disagreement	LOF	OCSVM	ABOD	KNN	HBOS	**1.0**
	Max Disagreement Entropy	LOF	OCSVM	ABOD	HBOS	KNN	0.926
	Max Entropy	LOF	ABOD	KNN	OCSVM	HBOS	0.816
	Random	LOF	OCSVM	ABOD	KNN	HBOS	**1.0**
	Variance Entropy	LOF	OCSVM	HBOS	KNN	ABOD	0.863
Pima (100% data)	No Acquisition	ABOD	LOF	LODA	OCSVM	KNN	–
Pima with 40% subset	Boundary	ABOD	LOF	OCSVM	LODA	CBLOF	0.801
	Boundary Max Disagreement	ABOD	LOF	LODA	OCSVM	CBLOF	**0.89**
	Max Disagreement Entropy	LOF	ABOD	OCSVM	LODA	KNN	0.743
	Max Entropy	ABOD	LOF	LODA	OCSVM	KNN	**1.0**
	Random	LOF	ABOD	LODA	OCSVM	CBLOF	0.733
	Variance Entropy	LOF	ABOD	OCSVM	LODA	KNN	0.743
Wilt (100% data)	No Acquisition	KNN	OCSVM	HBOS	ABOD	LOF	–
Wilt with 20% subset	Boundary	OCSVM	KNN	LOF	ABOD	HBOS	0.696
	Boundary Max Disagreement	KNN	OCSVM	HBOS	ABOD	LOF	**1.0**
	Max Disagreement Entropy	OCSVM	KNN	HBOS	LOF	ABOD	0.76
	Max Entropy	OCSVM	KNN	LOF	HBOS	ABOD	0.678
	Random	KNN	OCSVM	HBOS	ABOD	LOF	**1.0**
	Variance Entropy	OCSVM	HBOS	KNN	LOF	ABOD	0.678
PageBlocks (100% data)	No Acquisition	OCSVM	LOF	HBOS	KNN	ABOD	–
PageBlocks with 10% subset	Boundary	LOF	ABOD	OCSVM	HBOS	CBLOF	0.551
	Boundary Max Disagreement	OCSVM	LOF	HBOS	KNN	ABOD	**1.0**
	Max Disagreement Entropy	LOF	ABOD	OCSVM	IsoForest	KNN	0.539
	Max Entropy	ABOD	OCSVM	LOF	KNN	HBOS	0.662
	Random	OCSVM	LOF	HBOS	KNN	ABOD	**1.0**
	Variance Entropy	LOF	KNN	OCSVM	IsoForest	HBOS	0.535

3.1 Discussion on Results

Table 1 presents the various evaluation criteria like F1-score, Accuracy, Average Precision Score, AUC ROC for two model selection techniques – Exact Inference (EI) and Stochastic Variational Inference (SVI), averaged across different acquisition functions, and subset sizes for all datasets. These metrics are widely used for evaluating supervised/unsupervised anomaly detection models in literature [4]. We can clearly observe that EI and SVI have comparable performance across different evaluation criteria, however, the time taken for SVI is exponentially higher (at least 3000x seconds higher for each dataset) than EI. Hence, we report only the EI results in the forthcoming experiments.

From Figure 1, we can observe the evaluation criteria across all acquisition functions and subset sizes for all datasets with EI, including 100% (All) data. The baseline metrics obtained in our experiments are similar for all datasets from [4]. We primarily focus on F1-score, Average Precision score, while accuracy is mostly not emphasized since AD datasets are highly imbalanced. An interesting observation from Figure 1 is that for an unsupervised setting, random acquisition can perform as well as other acquisition functions.

Figure 1 shows that the average precision of Variance Entropy decreases across subset sizes for datasets with higher anomalies (like Pima), and decreases for datasets with lower anomalies increases in general. We also observe that Boundary Max Disagreement effectively converges towards optimal F1-scores and average precision scores as the subset size increases across all datasets.

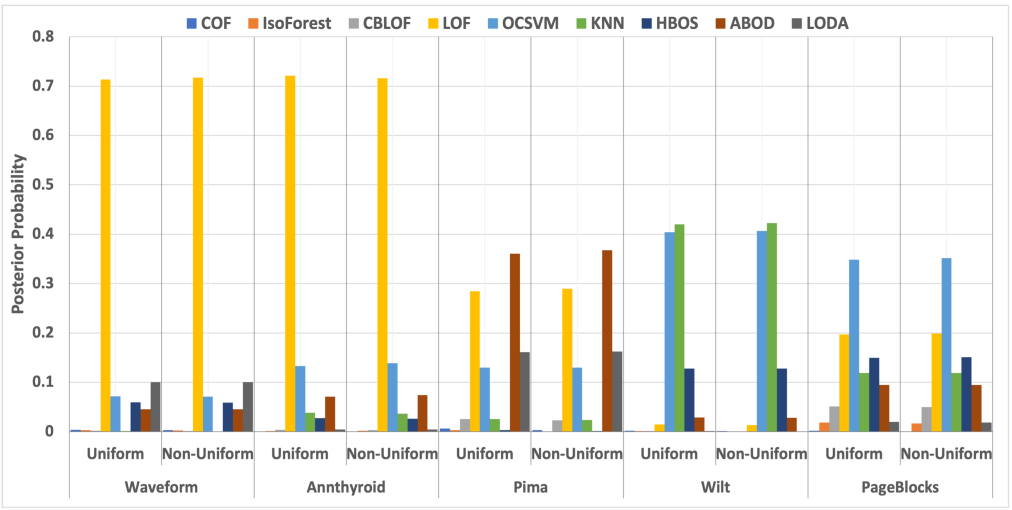


Figure 2: Posterior Probabilities obtained after Exact Inference initialized with Uniform and Non-Uniform Dirichlet Priors across all datasets with the best performing Boundary Max Disagreement Acquisition Function and best corresponding subset size for each dataset as reported in Table 2.

We then filter the best corresponding subset size and report the top-k recommended models across all acquisition functions, along with the entire data (100% with no acquisition) in Table 2. Here, we observe that the Ranking Score (η) (as observed in Section 2.3 is mostly high for Boundary Max Disagreement, which implies consistent performance in identifying top-k models along with random.

We also showcase the posterior probabilities obtained with EI when initialized with uniform and non-uniform Dirichlet priors in Figure 2, with the Boundary Max Disagreement criterion with the corresponding best subset sizes from Table 2. The non-uniform priors are sampled at random across models, simulating a scenario as discussed in Section 2.1.1. From the figure, we can infer that the posterior probabilities obtained when initialized with non-uniform priors perform almost similarly to posterior probabilities when initialized with uniform priors across all datasets, indicating the effectiveness of our framework with both uniform and non-uniform priors. The top-k corresponding recommended models with uniform and non-uniform priors are also shown in Appendix C Table 4, which show that even with custom non-uniform Dirichlet priors over models, our framework efficiently recommends the top-k models.

4 Conclusion

In this paper, we present three important contributions pertaining to unsupervised anomaly detection. First, we identify the most important subsets of data points by systematically analyzing various existing and novel acquisition functions. Second, we successfully showcase the effectiveness of our unified data-efficient Bayesian Inference model selection framework, demonstrated by the evaluation criteria. Finally, we also formulate a Ranking Score (η) to rank our top-k models selected using Bayesian inference which operates in settings with both uniform and non-uniform priors over models, thereby enabling end-users to easily use the selected/recommended models.

References

[1] Charu C Aggarwal. An introduction to outlier analysis. In *Outlier analysis*, pages 1–34. Springer, 2017.

[2] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20:973–978, 2019.

[3] Evgeny Burnaev, Pavel Erofeev, and Dmitry Smolyakov. Model selection for anomaly detection. In *Eighth International Conference on Machine Vision (ICMV)*, volume 9875, pages 445–450. SPIE, 2015.

[4] Guilherme O Campos, Arthur Zimek, Jörg Sander, Ricardo JGB Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E Houle. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30(4):891–927, 2016.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):1–58, 2009.

[6] Hongmei Deng and Roger Xu. Model selection for anomaly detection in wireless ad hoc networks. In *2007 IEEE Symposium on Computational Intelligence and Data Mining*, pages 540–546. IEEE, 2007.

[7] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

[8] Gautham Krishna Gudur, Prahalathan Sundaramoorthy, and Venkatesh Umaashankar. Active-harnet: Towards on-device deep bayesian active learning for human activity recognition. In *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications*, pages 7–12, 2019.

[9] Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[10] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

[11] Dan Pelleg and Andrew Moore. Active learning for anomaly and rare-category detection. *Advances in Neural Information Processing Systems*, 17, 2004.

[12] Tiago Pimentel, Marianne Monteiro, Adriano Veloso, and Nivio Ziviani. Deep active learning for anomaly detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[13] Abhijith Ragav and Gautham Krishna Gudur. Bayesian active learning for wearable stress and affect detection. *arXiv preprint arXiv:2012.02702*, 2020.

[14] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[15] Stefania Russo, Moritz Lürig, Wenjin Hao, Blake Matthews, and Kris Villez. Active learning for anomaly detection in environmental data. *Environmental Modelling & Software*, 134, 2020.

[16] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2012.

[17] Jonas Soenen, Elia Van Wolputte, Lorenzo Perini, Vincent Vercruyssen, Wannes Meert, Jesse Davis, and Hendrik Blockeel. The effect of hyperparameter tuning on the comparative evaluation of unsupervised anomaly detection methods. In *Proceedings of the KDD'21 Workshop on Outlier Detection and Description*, pages 1–9, 2021.

[18] Yuanxiang Ying, Juanyong Duan, Chunlei Wang, Yujing Wang, Congrui Huang, and Bixiong Xu. Automated model selection for time-series anomaly detection. *arXiv preprint arXiv:2009.04395*, 2020.

[19] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019.

[20] Yue Zhao, Ryan Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

A Related Work

Conventionally, most anomaly detection/outlier detection works have rigorously focused on improving existing anomaly detection algorithms or proposing new interesting ones [1]. Such works deal with multiple types of anomalies, which can broadly be classified as point, contextual and collective anomalies [5].

Point anomalies: Data instances are considered anomalous with respect to the rest of the data.

Contextual anomalies: Data instances are considered anomalous in a specific context, like a month/day of week/location.

Collective anomalies: Collection of continuous data instances is considered anomalous with respect to the entire dataset.

This necessitates an anomaly detection system with appropriate approaches to cater to all such anomaly types. Moreover, most anomaly detection systems in real-time are inherently unlabeled, thereby making them unsupervised in nature. These challenges make identifying all such anomalies a tedious process in addition to supervised settings, and require data labeling from experts well-versed in their respective domains.

However, algorithms selection and/or recommendation in an unsupervised AD setting has been relatively unexplored. Model selection in AD systems conventionally requires careful and rigorous selection over all possible sets of relevant algorithms. Moreover, finding a common metric for evaluation and effective comparison of these algorithms is hard, due to their unsupervised nature and requires the context of the problem. Hence, a robust framework that can be incorporated into the existing anomaly detection systems (consisting of diverse AD algorithms) becomes necessary to select/recommend the best algorithms for the given data.

We note some existing work on model selection for one-class models [3, 6], however they are limited only to a specific type of model class. METAOD [20] proposes an effective way to select the best approaches in unsupervised anomaly detection, however, it relies on training an offline meta-learner on various meta-train datasets, with hand-picked meta-features, which are computationally expensive to create. Similarly, [18] proposes using a pre-trained model selector and pre-trained parameter estimator for unsupervised anomaly detection which is again cumbersome to learn.

The conventional ways of model selection for any machine learning task, in our case, unsupervised anomaly detection, involve selecting the best hyperparameters from an exhaustive initial range of multiple hyperparameter values using Grid Search or similar mechanisms, with hold-out validation set [14, 17]. However, such mechanisms are again search algorithms, which end up taking massive amounts of time, and require a thorough knowledge of the hyperparameters to choose from.

Conventional existing works on active learning typically involve machine learning classification algorithms, and a few interesting applications [16, 13, 8]. There are also multiple active learning/subset selection works that leverage the use of an oracle (user feedback), which are particularly useful in unsupervised anomaly detection settings, where there are predominantly no labels in real-time applications. This helps alleviate massive labeling efforts for domain experts. Interesting works on finding useful anomalies using active learning, and treating them as a rare category [11], active learning for AD on environmental data [15] are noted. Deep active learning for unsupervised AD is explored in [12], however, none of the above is in the context of efficient model selection.

Hence, a unified framework for human-augmented (data-efficient) model selection in any given anomaly detection system, particularly unsupervised AD, makes it more convenient for the end-user to identify the appropriate best-fit algorithms for the problem at hand.

B Datasets

We report the characteristics of the unsupervised anomaly detection datasets used in the paper in Table 3.

Table 3: Characteristics of the Datasets

Dataset	Instances	Attributes	Outliers (%)
Waveform	3443	21	2.9
Annthyroid	7129	21	7.49
Pima	768	8	34.9
Wilt	4819	5	5.33
PageBlocks	5393	10	9.46

C Dirichlet Priors

Table 4: Top-5 AD algorithms from non-uniform Dirichlet Priors and their corresponding top-5 recommended algorithms from Posterior along with their respective Ranking Scores for all datasets with Boundary Max Disagreement Acquisition Function and best corresponding subset size.

Dataset with Best Subset Size	Dirichlet Prior	Ranking (top-5) 1	2	3	4	5
Waveform with 40% subset	**Uniform**	LOF	LODA	OCSVM	HBOS	ABOD
	Non-Uniform	LOF	LODA	OCSVM	HBOS	ABOD
Annthyroid with 30% subset	**Uniform**	LOF	OCSVM	ABOD	KNN	HBOS
	Non-Uniform	LOF	OCSVM	ABOD	KNN	HBOS
Pima with 40% subset	**Uniform**	ABOD	LOF	LODA	OCSVM	KNN
	Non-Uniform	ABOD	LOF	LODA	OCSVM	KNN
Wilt with 20% subset	**Uniform**	KNN	OCSVM	HBOS	ABOD	LOF
	Non-Uniform	KNN	OCSVM	HBOS	ABOD	LOF
PageBlocks with 10% subset	**Uniform**	OCSVM	LOF	HBOS	KNN	ABOD
	Non-Uniform	OCSVM	LOF	HBOS	KNN	ABOD