DATA-EFFICIENT AUTOMATIC MODEL SELECTION IN UNSUPERVISED ANOMALY DETECTION

Gautham Krishna Gudur, Raaghul R, Adithya K, Shrihari Vasudevan

Global AI Accelerator (GAIA), Ericsson R&D

ANOMALY DETECTION

- Process of identifying unexpected/unforeseen events in an application.
- Evolves over time – changes in data distributions, geographical constraints, business contexts.
- Types of anomalies – 1) Point, 2) Contextual, 3) Collective.

Unsupervised Anomaly Detection

- Most real-world AD applications are unsupervised, due to bottleneck of obtaining labelled data.
- Require domain-experts to identify anomalies.

MOTIVATION

- Approach recommendation for unsupervised AD is quite unexplored.
- Requires careful selection over all possible sets of relevant algorithms.
- Necessity to select top-k approaches which best fit the characteristics of data.
- Data-efficient methods necessary for querying least number of data points from domain expert.
- Prior knowledge in domain could accelerate model selection.

CONTRIBUTIONS

- A **generic model selection framework** to recommend top-k approaches using **Bayesian Inference** in unsupervised AD.
- Leverage multiple **existing/novel acquisition functions** to identify **most-informative** data points/subsets to be acquired from domain expert (oracle).
- Propose a **novel ranking criterion** for selecting the best acquisition functions.
- Benchmark on various standard datasets in unsupervised AD settings.

BAYESIAN INFERENCE

- Categorical likelihood distribution over all initial models.
- If prior distribution is Dirichlet, then corresponding posterior distribution is also Dirichlet.
- Dirichlet distribution is a conjugate prior for the Categorical/Multinomial distribution.

$$\alpha_i + c_i, i = 1 \ldots k,$$

$$p_i | alpha_i, c_i \sim Dir(k, \alpha_i + c_i), \ \forall\ i = 1 \ldots k$$

- Dirichlet Priors – domain knowledge/apriori beliefs over models – like taxonomy for the given data.

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

Bayesian Inference

Initialize Categorical distribution (likelihood) over M models

Initialize Dirichlet Priors $p_i \sim Dir(\alpha_i), i = 1, .., M$ with concentrations α_i

Obtain probabilities from all M unsupervised AD models with $\mathcal{D}_{train}\{x\}$

Acquisition Functions for subset selection

Subset selection on $\mathcal{D}_{train}\{x\}$ using AF to obtain $\mathcal{D}_{subset}\{x\}$

Present $\mathcal{D}_{subset}\{x\}$ to oracle to obtain labeled subset $\mathcal{D}_{subset}\{x, y\}$

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

Evaluation Metrics and Ranking Score (η)

end for

Return top-k models with best hyperparameters using Grid Search

BAYESIAN INFERENCE FOR MODEL SELECTION

- We propose using Exact Inference and Stochastic Variational Inference for modelling posterior probabilities.
- Our framework supports uniform/non-uniform Dirichlet priors over AD models.
- Default setting – uniform (when taxonomies are unavailable)
- The samples of the Dirichlet posterior would give us probabilities of the Categorical distribution over all AD models, from which top-k models are chosen.

ACQUISITION FUNCTIONS

- Boundary – Selects points that are closest to the boundary threshold for each model, and are 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 closest to the boundary threshold, and then selects the points with Max disagreement.

- Max Entropy – 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 – 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.

EXPERIMENTS

- Five different anomaly detection datasets from DAMI.

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

- Subset sizes – 5%, 10%, 20%, 30%, 40%
- Objective –
  - **Use Ranking Score (η)** to choose best-performing acquisition function

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

  - **Select the top-5 unsupervised anomaly detection models** using Bayesian Inference for EI & SVI.

Initial AD algorithms used –

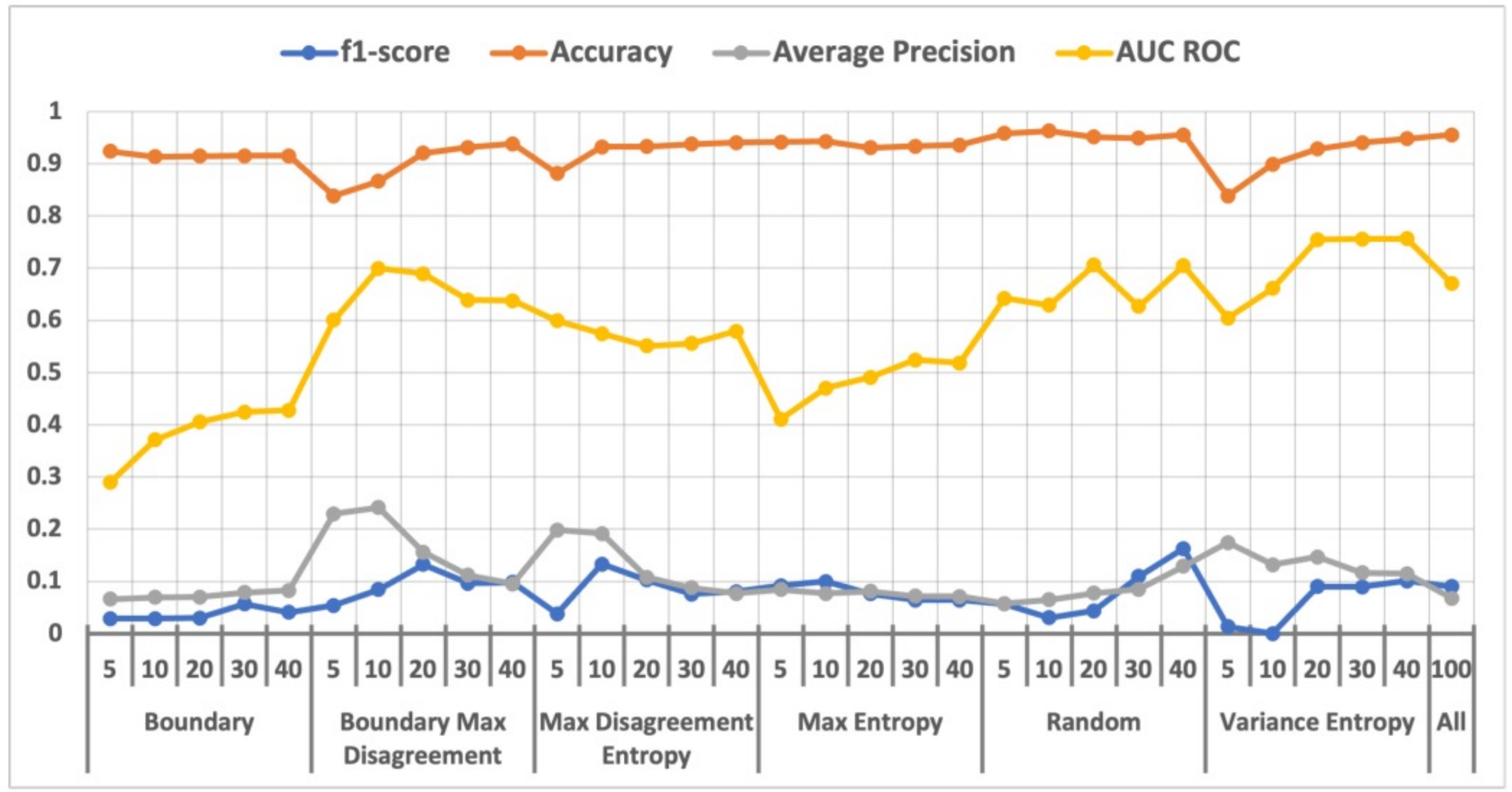
- COF
- Isolation Forest
- CBLOF
- LOF
- OCSVM
- KNN
- HBOS
- ABOD
- LODA

BAYESIAN INFERENCE –
Exact Inference vs Stochastic Variational Inference

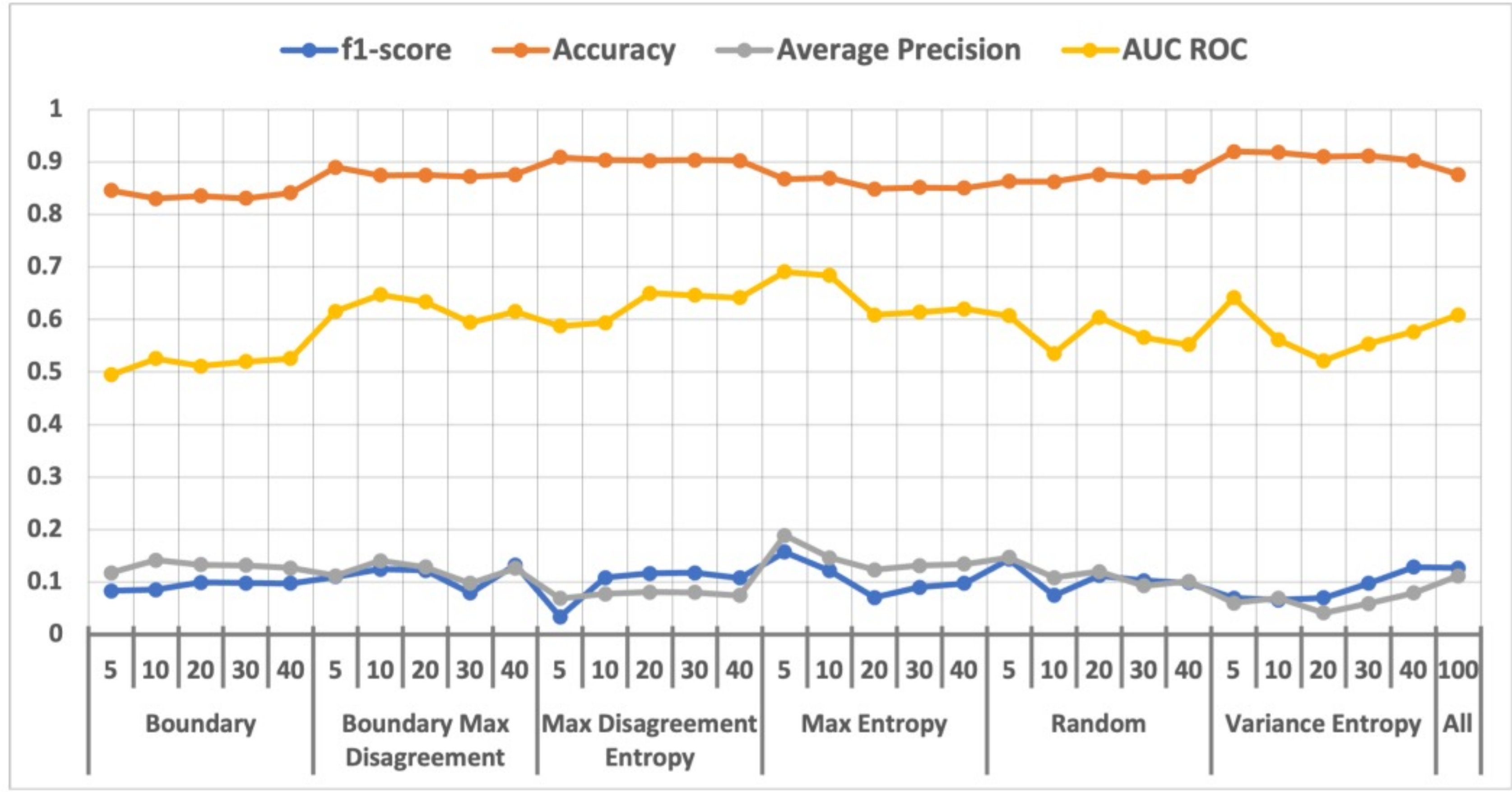
	Accuracy (%)		f1-score		Avg Precision		AUC ROC		Time (in sec)	
Dataset	**EI**	**SVI**	**EI**	**SVI**	**EI**	**SVI**	**EI**	**SVI**	**EI**	**SVI**
Waveform	92.395	92.365	0.072	0.062	0.111	0.106	0.577	0.564	0.322	1092.113
Annthyrroid	87.621	87.526	0.1	0.102	0.108	0.103	0.591	0.59	0.294	995.956
Pima	63.513	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

- Comparable performance in evaluation metrics across all datasets.
- Time taken for SVI is exponentially higher (at least 3000x).

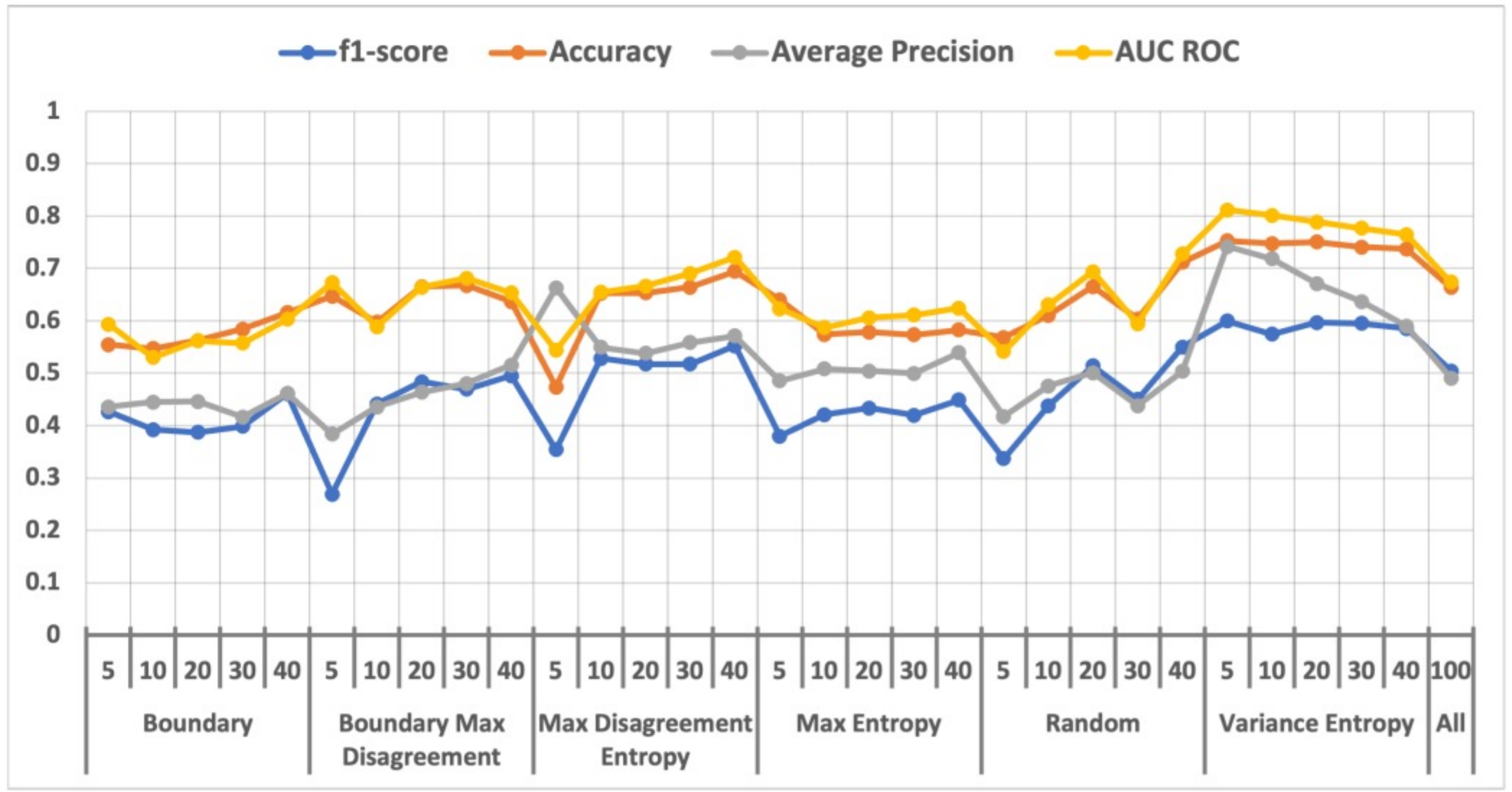
ACQUISITION FUNCTIONS – RESULTS



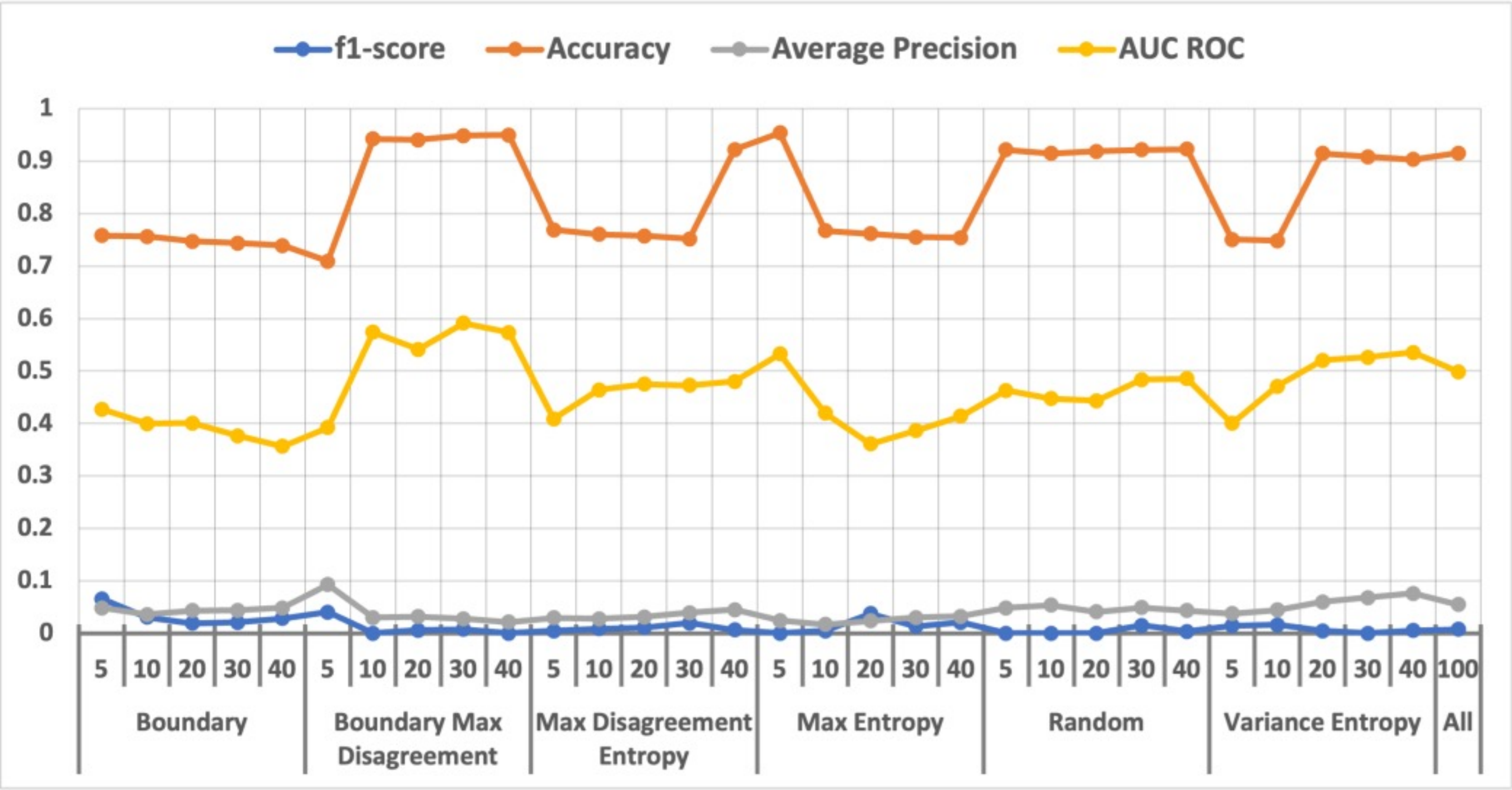
(a) Waveform



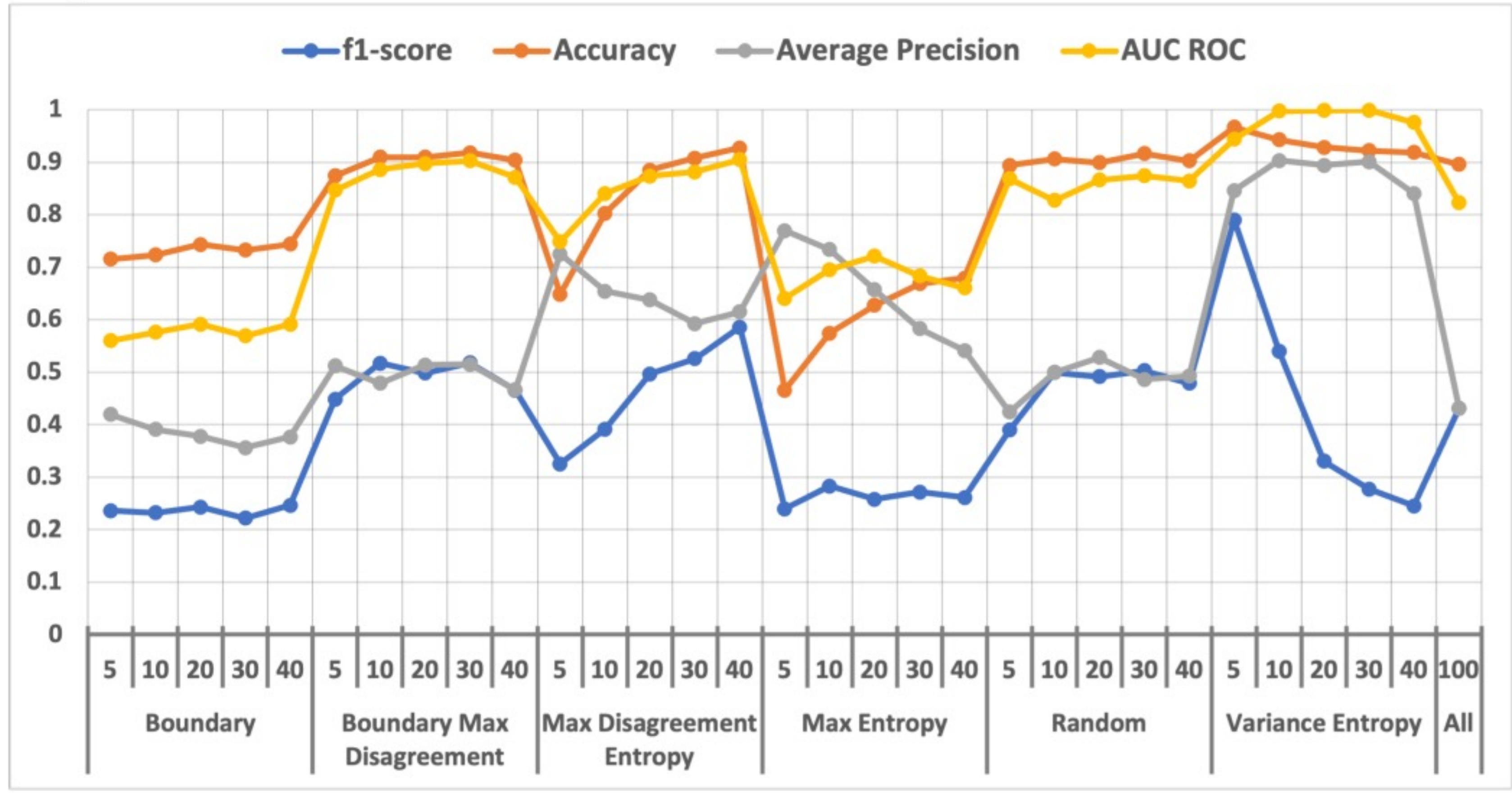
(b) Annthyroid



(c) Pima



(d) Wilt



(e) PageBlocks

ACQUISITION FUNCTIONS – RESULTS

Variance Entropy

- **Average Precision decreases across subset sizes** for datasets with high % of anomalies (like Pima).
- **Average Precision increases or is at least consistent across subset sizes** for datasets with low % of anomalies.

Boundary Max Disagreement

- converges towards optimal Average Precision, f1-score.
- comparable efficiencies to all (100%) data.

F1-score and Average Precision – primary metrics used (accuracy not emphasized since AD datasets are highly imbalanced).

MODEL SELECTION – RESULTS

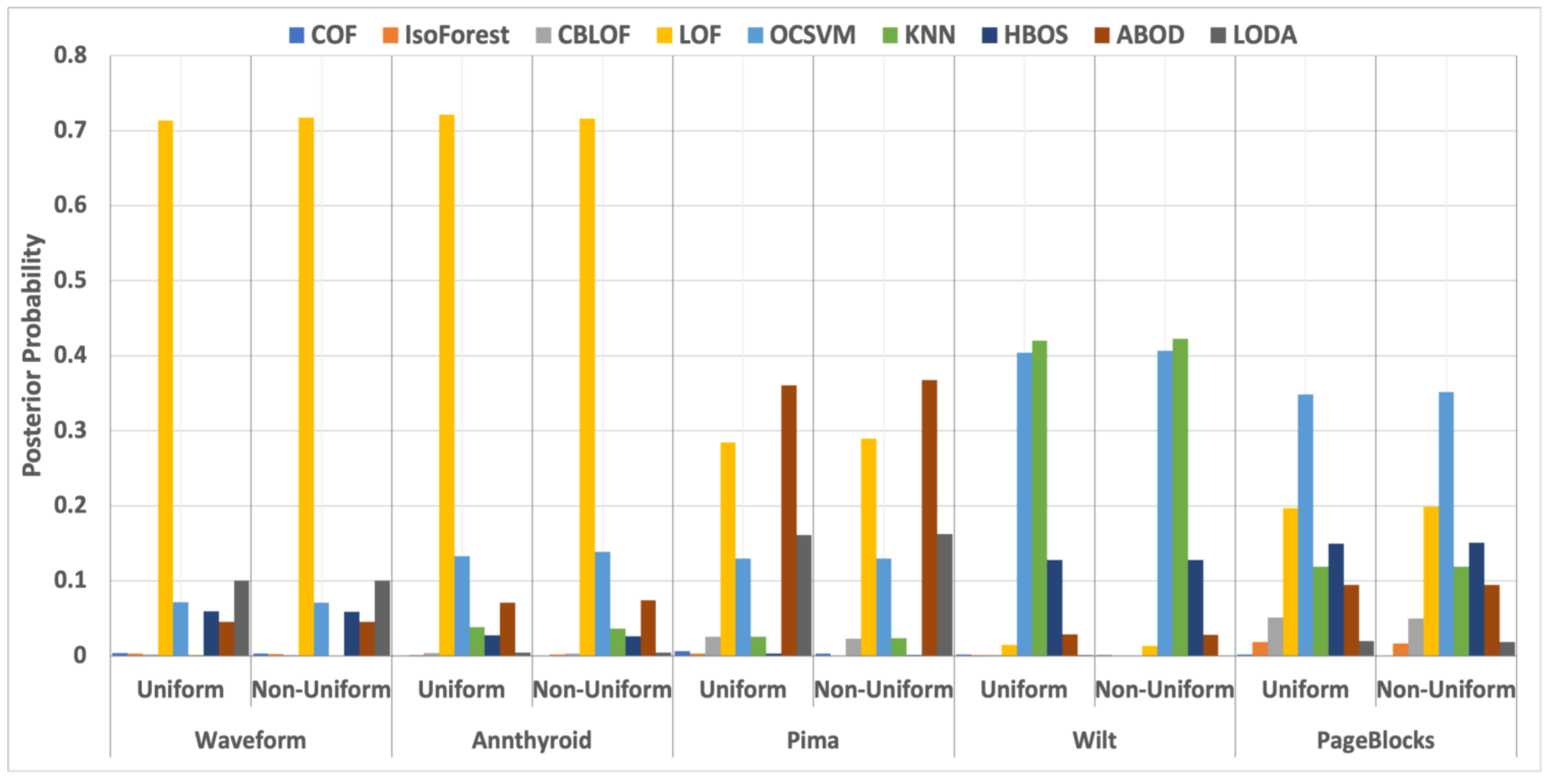
Dataset with Best Subset Size	Acquisition Criterion	Ranking (top-5)					Ranking Score (η)
		1	2	3	4	5	
Waveform (100% data)	No acquisition	LOF	LODA	OCSVM	HBOS	ABOD	–
Waveform with 40% subset	Boundary	LOF	LODA	OCSVM	HBOS	ABOD	**1.0**
	Boundary Max Disagreement	LOF	LODA	OCSVM	HBOS	ABOD	**1.0**
	Max Disagreement 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

MODEL SELECTION – RESULTS

- **Posterior Probabilities of non-uniform priors** can adapt as well as **posterior probabilities of non-uniform priors** with respect to the chosen subset data.

- Top-5 recommended AD algorithms are also efficiently recommended.

- Top-k AD models are efficiently chosen with any chosen Dirichlet Priors coupled with acquisition functions for subset selection.



Posterior Probabilities initialized with uniform/non-uniform Dirichlet Priors with Boundary Max Disagreement

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

Top-5 recommended algorithms with uniform/non-uniform Dirichlet Priors with Boundary Max Disagreement

THANK YOU!

Contact

gautham.krishna.gudur@ericsson.com

Gautham Krishna Gudur

Global AI Accelerator (GAIA),
Ericsson R&D