ADBench: Anomaly Detection Benchmark

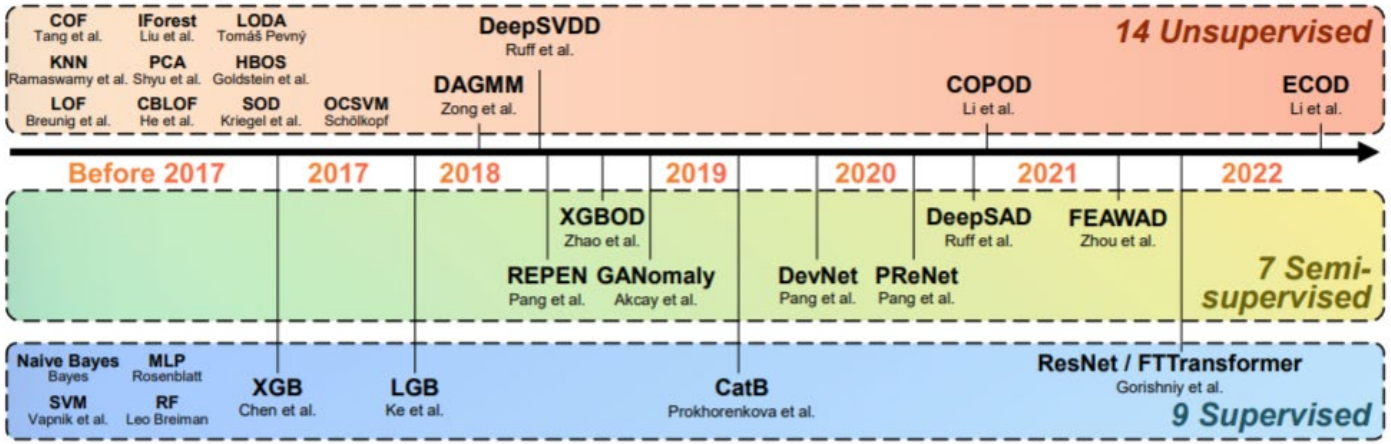
Songqiao Han[1], Xiyang Hu[2], Hailiang Huang[1], Minqi Jiang[1], Yue Zhao[2]
1 Shanghai University of Finance and Economics
2 Carnegie Mellon University

NEURAL INFORMATION PROCESSING SYSTEMS

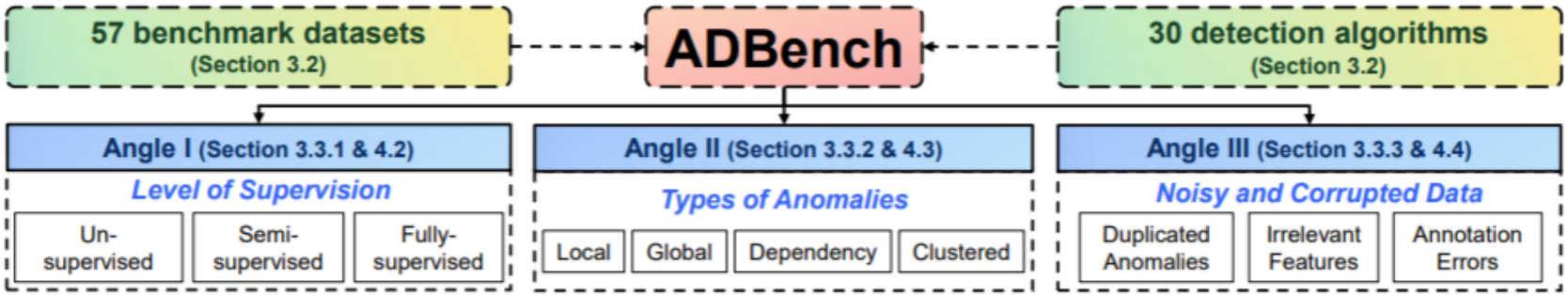
What did ADBench do?

Given **a long list of anomaly detection algorithms** developed in the last few decades, how do they perform with regard to:

(i) varying levels of supervision;
(ii) different types of anomalies;
(iii) noisy and corrupted data?



Based on the above questions, we design the proposed ADBench via **3 Angles**:



Contribution of ADBench

Data	# Samples	# Features	# Anomaly	% Anomaly	Category	Reference
ALOI	49534	27	1508	3.04	Image	[42]
annthyroid	7200	6	534	7.42	Healthcare	[141]
backdoor	95329	196	2329	2.44	Network	[119]
breastw	683	9	239	34.99	Healthcare	[173]
campaign	41188	62	4640	11.27	Finance	[131]
cardio	1831	21	176	9.61	Healthcare	[12]
Cardiotocography	2114	21	466	22.04	Healthcare	[12]
celeba	202599	39	4547	2.24	Image	[131]
census	299285	500	18568	6.20	Sociology	[131]
cover	286048	10	2747	0.96	Botany	[18]
donors	619326	10	36710	5.93	Sociology	[131]
fault	1941	27	673	34.67	Physical	[42]
fraud	284807	29	492	0.17	Finance	[131]
glass	214	7	9	4.21	Forensic	[43]
Hepatitis	80	19	13	16.25	Healthcare	[36]
http	567498	3	2211	0.39	Web	[145]
InternetAds	1966	1555	368	18.72	Image	[25]
Ionosphere	351	33	126	35.90	Oryctognosy	[163]
landsat	6435	36	1333	20.71	Astronautics	[42]
letter	1600	32	100	6.25	Image	[48]
Lymphography	148	18	6	4.05	Healthcare	[26]
magic.gamma	19020	10	6688	35.16	Physical	[42]
mammography	11183	6	260	2.32	Healthcare	[176]
mnist	7603	100	700	9.21	Image	[90]
musk	3062	166	97	3.17	Chemistry	[37]
optdigits	5216	64	150	2.88	Image	[10]
PageBlocks	5393	10	510	9.46	Document	[113]
pendigits	6870	16	156	2.27	Image	[9]
Pima	768	8	268	34.90	Healthcare	[145]
satellite	6435	36	2036	31.64	Astronautics	[145]
satimage-2	5803	36	71	1.22	Astronautics	[145]
shuttle	49097	9	3511	7.15	Astronautics	[145]
skin	245057	3	50859	20.75	Image	[42]
smtp	95156	3	30	0.03	Web	[145]
SpamBase	4207	57	1679	39.91	Document	[25]
speech	3686	400	61	1.65	Linguistics	[23]
Stamps	340	9	31	9.12	Document	[25]
thyroid	3772	6	93	2.47	Healthcare	[142]
vertebral	240	6	30	12.50	Biology	[17]
vowels	1456	12	50	3.43	Linguistics	[82]
Waveform	3443	21	100	2.90	Physics	[107]
WBC	223	9	10	4.48	Healthcare	[114]
WDBC	367	30	10	2.72	Healthcare	[114]
Wilt	4819	5	257	5.33	Botany	[25]
wine	129	13	10	7.75	Chemistry	[2]
WPBC	198	33	47	23.74	Healthcare	[114]
yeast	1484	8	507	34.16	Biology	[66]
CIFAR10	5263	512	263	5.00	Image	[81]
FashionMNIST	6315	512	315	5.00	Image	[178]
MNIST-C	10000	512	500	5.00	Image	[120]
MVTec-AD	See Table B2.				Image	[16]
SVHN	5208	512	260	5.00	Image	[121]
Agnews	10000	768	500	5.00	NLP	[192]
Amazon	10000	768	500	5.00	NLP	[63]
Imdb	10000	768	500	5.00	NLP	[111]
Yelp	10000	768	500	5.00	NLP	[192]
20newsgroups	See Table B3.				NLP	[86]

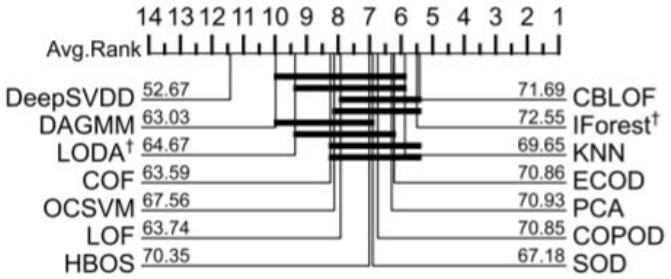
Benchmark	Coverage (§3.2)		Data Source		Algorithm Type		Comparison Angle (§3.3)		
	# datasets	# algo.	Real-world	Synthetic	Shallow	DL	Supervision	Types	Robustness
Ruff et al. [150]	3	9	✓	✓	✓	✓	✗	✓	✗
Goldstein et al. [53]	10	19	✓	✗	✓	✗	✗	✓	✗
Domingues et al. [38]	15	14	✓	✗	✓	✗	✗	✗	✓
Soenen et al. [164]	16	6	✓	✗	✓	✗	✗	✗	✗
Steinbuss et al. [166]	19	4	✗	✓	✓	✗	✗	✓	✗
Emmott et al. [42]	19	8	✓	✓	✓	✗	✗	✓	✓
Campos et al. [25]	23	12	✓	✗	✓	✗	✗	✗	✗
ADBench (ours)	57	30	✓	✓	✓	✓	✓	✓	✓

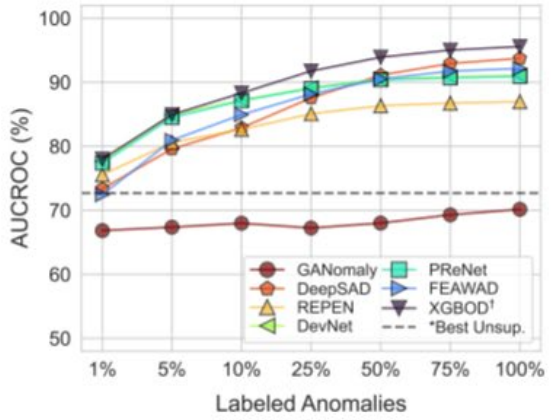
Compared to the existing AD benchmark, ADBench includes:

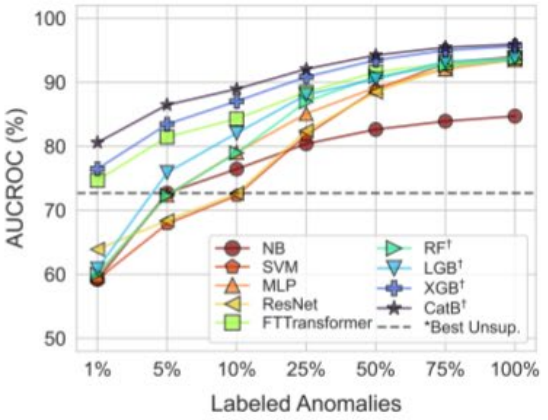
- Most datasets (57), including 10 **NLP** and **CV** datasets transformed by the pretrained model
- Most algorithms (30), including both **shallow** and **deep** learning algorithms
- Both **real-world** and **synthetic** datasets
- Multiple comparison angles
- Evaluation with both ML metrics and **statistical tests**
- Extensive experiments (**98, 436**)

Minqi Jiang (jiangmq95@163.com)

Angle I: Availability of Ground Truth Labels (Supervision)



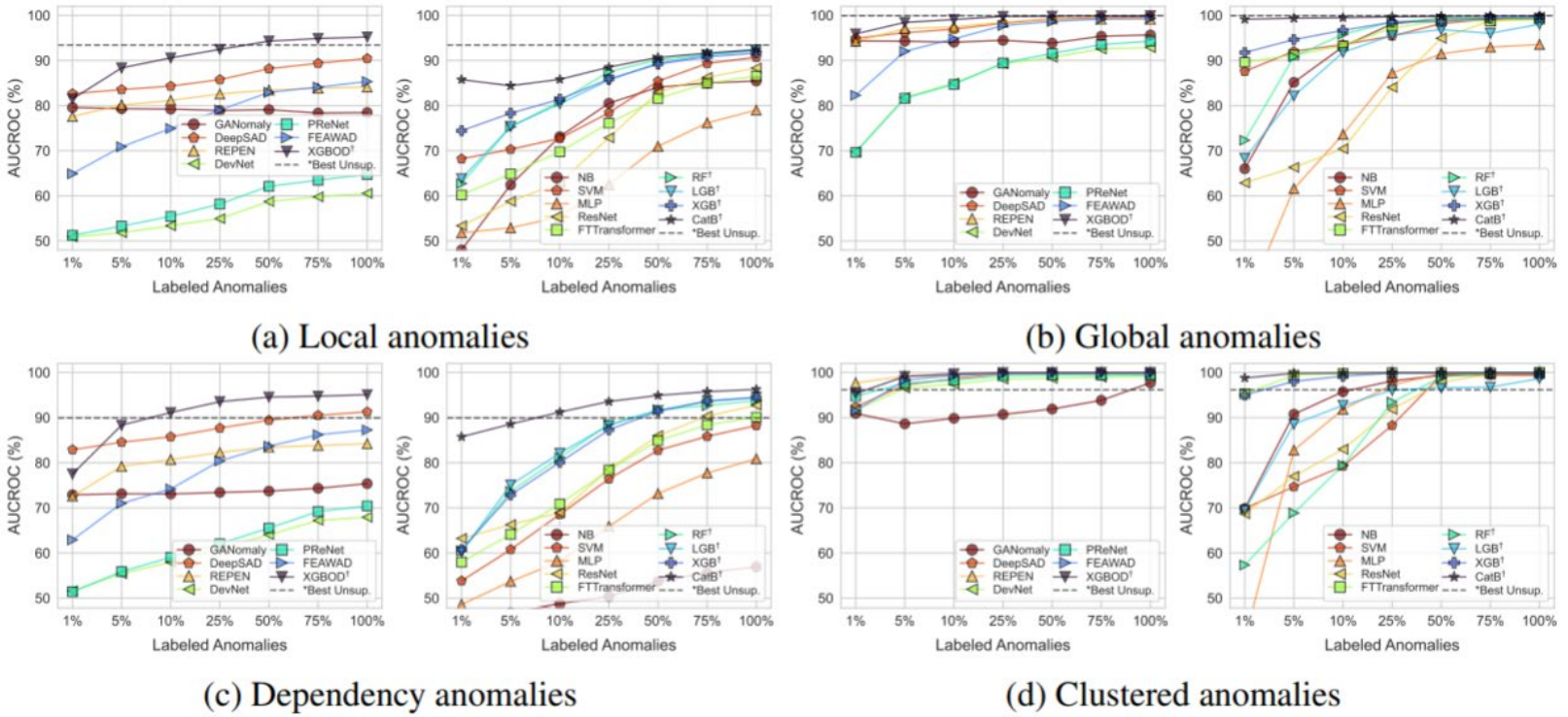




Surprisingly **none** of the benchmarked unsupervised algorithms is statistically better than others, emphasizing the importance of **algorithm selection**;

With merely 1% labeled anomalies, most semi-supervised methods can **outperform** the best unsupervised method, justifying the **importance of supervision**.

Angle II: Types of Anomalies



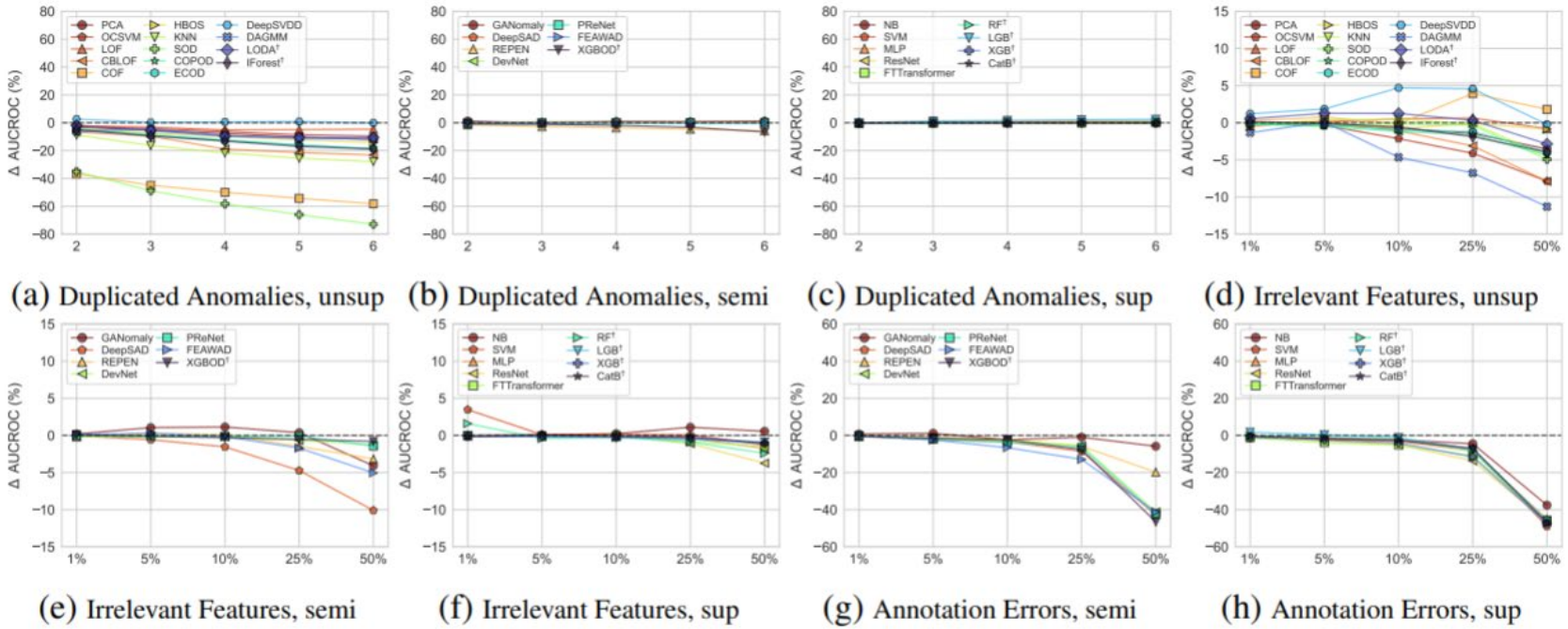
(a) Local anomalies (b) Global anomalies

(c) Dependency anomalies (d) Clustered anomalies

We observe that best unsupervised methods for specific types of anomalies are **even better than** semi- and fully-supervised methods;

That is to say, the **prior knowledge of data type** could be more valuable than that of **labeled anomalies**, revealing the necessity of understanding data characteristics.

Angle III: Model Robustness with Noisy and Corrupted Data



(a) Duplicated Anomalies, unsup (b) Duplicated Anomalies, semi (c) Duplicated Anomalies, sup (d) Irrelevant Features, unsup

(e) Irrelevant Features, semi (f) Irrelevant Features, sup (g) Annotation Errors, semi (h) Annotation Errors, sup

Semi-supervised methods show potential in achieving robustness in noisy and corrupted data, possibly due to their efficiency in using labels and feature selection.

Future Direction

Based on the experimental results and analysis, we give the several possible future direction for AD community:

- Unsupervised Algorithm Evaluation, Selection, and Design;
- Semi-supervised Learning;
- Leveraging Anomaly Types as Valuable Prior Knowledge;
- Noise-resilient AD Algorithms.

Thank You!

Paper

Github

Contact

Minqi Jiang (jiangmq95@163.com)

Yue Zhao (zhaoy@cmu.edu)