Characterizing and understanding ensemble-based anomaly-detection

Gustavo de P. Avelar[1], Guilherme O. Campos[1], Wagner Meira Jr.[1]

Universidade Federal de Minas Gerais (UFMG), Brazil
{gustavopaula, gocampos, meira}@dcc.ufmg.br

Abstract.

Anomaly Detection (AD) has grown in importance in recent years, as a result of an increasing digitalization of services and data storage, and abnormal behavior detection has become a key task. However, discovering abnormal data that is mixed with the huge amount of data available is a daunting problem and the efficacy of the current methods depends on a wide range of assumptions. One effective strategy for detecting anomalies is to combine multiple models, which are called "ensembles", but the factors that determine their performance are often hard to determine, making their calibration and improvement a challenging task. In this paper we address these problems by employing a four-step method for the characterization and understanding of ensemble-based anomaly-detection task. We start by characterizing several datasets and analyzing the factors that make it hard to detect their anomalies. We then evaluate to what extent existing algorithms are able to detect anomalies in the same datasets. On the basis of both analyses, we propose a stacking-based ensemble that outperformed a state-of-the-art baseline, Isolation Forest. Finally, we examine the benefits and drawbacks of our proposal.

CCS Concepts: • **Computing methodologies** → **Anomaly detection**; **Machine learning algorithms**.

Keywords: anomaly detection, data mining, ensembles, machine learning, interpretability

1. INTRODUCTION

The Advances in information technology have led to the generation of huge amounts of data throughout the world. On top of this, large volumes of data are continuously being scraped by various smart devices and organizations. These data are being used to obtain an insights into current trends and social behaviors, among other applications.

Anomaly Detection (AD) has gained particular importance in such scenario [Hodge and Austin 2004]. A widely accepted definition[Hawkins 1980] is that "*An anomaly is an observation which deviates so much from the other observations as to arouse suspicion that it was generated by a different mechanism*". Discovering "rare" behavior or an anomalous event in a huge amount of data can lead to many insights, hidden information, potential trends and behavioral patterns within society.

As a result, the development of Anomaly Detection methods and techniques has recently received significant attention and several approaches were developed. Most of them are based on certain assumptions about what anomalous data are, and these assumptions also determine the effectiveness of the resulting method. As examples, We may mention methods that are *model-based* [Hawkins 1980], *distance-based* [Knorr and Ng 1998; Ramaswamy et al. 2000] and *density-based* [Breunig et al. 2000], and *clustering-Based* [He et al. 2003], among others.

One strategy to make anomaly detection more accurate and reliable is to combine several meth-

ods (called base detectors) in a single one. Such combination is called *ensemble* and it has been widely applied in the data classification and clustering scenarios with outstanding performance and results [Aggarwal 2013]. A well-known ensemble for anomaly detection is the Isolation Forest [Liu et al. 2008], which exploits sub-sampling of features to build an ensemble and will be used as a baseline in this work.

Ensemble approaches for anomaly detection vary according to the target scenario and the combination of *base-detectors* aims to maximize both accuracy and diversity [Dietterich 2000]. Furthermore, there is no clear procedure for combining several *base-detectors* and the difficulty of obtaining ground truth for some scenarios makes the decision-making process harder while combining several *base-detectors* [Aggarwal and Sathe 2015]. Another relevant issue is the understanding of the combination outcomes, which is necessary to improve both base detectors and the associated ensembles.

In this study, a new methodology based on the stacking ensemble technique is employed to combine the best of the most important anomaly detection algorithms. We take a step further, building on top of prior work [Campos et al. 2016]. Moreover, a section is devoted to discuss the interpretability and challenges of evaluating datasets incompatible with the standard definition of the Anomaly Detection (AD) problem. The experimental results obtained show that our proposed methodology is able to achieve a significant improvement in the effectiveness of the anomaly detection task.

We address these problems through a four-step method that enables the ensemble-based anomaly-detection task to be clearly characterized and fully understood. We start by characterizing several datasets and analyzing the difficulties of detecting their anomalies. We then evaluate how well-known algorithms can detect anomalies in the same datasets. On the basis of both these analyses, we propose a stacking-based ensemble that outperformed a state-of-the-art baseline, Isolation Forest. Finally, we took note of the benefits and drawbacks of the proposed ensemble.

2. METHODOLOGY

The methodology employed in this work consists of four steps:

Dataset characterization — Several datasets are characterized, by assessing the inherent challenges they raise for the anomaly detection task through a distribution of similarity distances between inliers and outliers. In particular, we seek to determine whether or not the outliers are similar to inliers.

Base-detectors Analysis — The base detectors are analyzed from the standpoint of two dimensions. The first dimension concerns their combination, as either conjunctive or disjunctive detectors. The second dimension refers to the recall of anomalies in the TOP-Z% of the instances considered for ranking outliers, as discussed by [Jin et al. 2001].

Ensemble proposal — On the basis of the findings regarding datasets and base detectors, we implemented a stacking-based ensemble, which is a meta-classifier that uses the outcomes of the base detectors as inputs to a logistic-regression based meta-classifier. We decided to use logistic-regression because of its simplicity and popularity. We compared our results with those of the Isolation Forest ensemble [Liu et al. 2008].

Performance Understanding — The individual and combined performances of the base detectors were assessed through effect plots, which quantify the contribution of each base detector when identifying both the inliers and outliers.

3. DATASET DESCRIPTION

There are a plethora of datasets available in online repositories like UCI ([Dua and Graff 2017]), but only a few of them can be used for the purpose of anomaly detection. In this study, we examine datasets

from two groups, and take account of their features, as described by [Campos et al. 2016]: (i) Literature — datasets which are widely used by the community for the task of anomaly detection/outlier detection, and (ii) Semantically meaningful — datasets that are derived from classification datasets, where the category that occurs least frequently is regarded as the one that describes rare/abnormal behavior. For instance, in a medical dataset about hypothyroidism, the patient's condition is likely to be classified as normal, hyper-function, and subnormal function. In this case, a patient's condition that is other than normal is considered to be an anomaly.

Dataset \ Features	# instances	# outliers	% outliers	References	AUC(and)	AUC(or)	Max Distanc. CDFs	Ensembles results	
								Isolation Forest (IF)	Proposed Stacking △
Ionosphere	351	126	35.90	Literature	0.90	0.92	0.47	0.76 ± (0.11)	**0.99** ± (0.02)
Lymphography (1-of-n)[a]	148	6	4.05	Literature	0.99	0.99	0.66	0.78 ± (0.05)	**1.00** ± (0.00)
PenDigits	9,868	20	0.2	Literature	0.98	0.97	0.38	0.75 ± (0.05)	**0.98** ±(0.03)
WBC	454	10	2.2	Literature	0.99	0.94	0.80	0.94 ± (0.01)	**0.99** ± (0.01)
WDBC	367	10	2.72	Literature	0.93	0.93	0.60	**0.94 ± (0.02)**	**0.95** ± (0.05)
WPBC	198	47	23.74	Literature	0.60	0.63	0.00	**0.47 ± (0.06)**	**0.56** ± (0.11)
Arrhythmia	450	206	45.78	Semantic	0.70	0.69	0.25	0.56 ± (0.01)	**0.75** ± (0.04)
Cardiotocography	2,126	471	22.04	Semantic	0.63	0.64	0.33	0.56 ± (0.01)	**0.83** (± 0.01)
InternetAds	3,264	454	18.72	Semantic	0.73	0.75	0.23	0.50 ± (0.00)	**0.94** ± (0.01)
PageBlocks	5,473	560	9.46	Semantic	0.798	0.82	0.42	0.75 ± (0.04)	**0.90** ± (0.05)
Pima	768	268	34.90	Semantic	0.68	0.72	0.22	0.56 ± (0.02)	**0.78** ± (0.03)

[a] Categorical features were converted using 1-of-n normalization which presented the best results for this dataset.

Table I: Summary of dataset features, Maximum AUC for and/or operations, Maximum distance between CDFs, and AUC results of the proposed ensembles compared to our baseline Isolation Forest(IF). The best results per dataset are presented in bold.

Table I summarizes the main features of each dataset used in this work. We focused on normalized datasets without duplicates, as they had led to the best results in a previous study [Campos et al. 2016]. Feature engineering tasks were carried out such as (i) feature scaling, (ii) removal and transformation of categorical attributes, and (iii) removal of duplicate elements.

4. EXPERIMENTAL SETTINGS

In this section we describe in detail the various settings of the experiments that are outlined below.

Base detectors — We employed the same as *base-detectors* [Campos et al. 2016; Campos et al. 2018]. These include strategies that take account of global features like *K-Nearest Neighbors* (KNN), *K-Nearest Neighbors Weighted* (KNNW) and *Outlier Detection using Indegree Number* (ODIN). In addition, accurate and efficient local approaches such as *Local Outlier Factor* (LOF), *Connectivity-based Outlier Factor* (COF), *Influenced Outlierness* (INFLO), *Local Density Factor* (LDF), *Fast Angle-based Outlier Detection* (FastABOD), *Local Outlier Probabilities* (LoOP), *Local Distance-based Outlier Factor* (LDOF), *Simplified Local Outlier Factor* (SimplifiedLOF), and *Kernel Density Estimates Outlier Score* (KDEOS).

Input Parameters — Most of the techniques require as input the parameter k, which defines the size of the neighborhood when calculating the "outlierness" score. We selected the best parameter k for each dataset, according to the best AUC ROC (*area under the curve of the receiver operating characteristic*), as shown in [Campos et al. 2016].

Score Normalization — The determination of the *outlierness* score differs among strategies and we had to normalize and standardize the final scores of each algorithm, as recommended by [Kriegel et al. 2011]. These kinds of procedures also help achieve better results for ensembles during the combination phase.

Evaluation — We defined the AUC ROC (AUC) metric for evaluating the effectiveness of anomaly detection methods. This metric relies on the well-known Receiver Operation Characteristics (ROC), which is defined by the True Positive Rate (TPR) versus False Positive Rate (FPR). TPR is equal to $\frac{TP}{TP+FP}$ and FPR is given by $\frac{FP}{TN+FP}$. This metric also handles well imbalanced datasets, which is the case of the anomaly detection task. The values of AUC range from 0 and

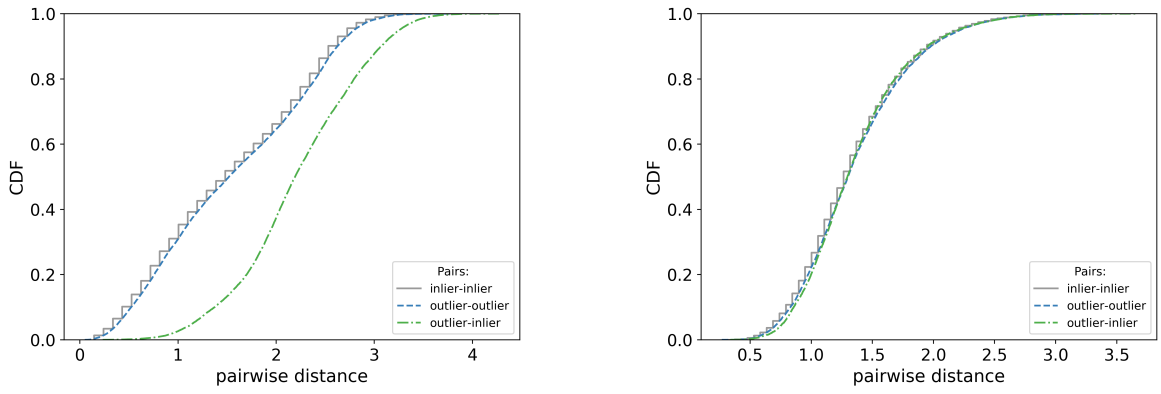
1. A perfect result gives an AUC with a value of 1, while, the worst result yields a value near 0. In this work, AUC also describes the probability of an outlier/abnormal object being recognized before than a normal object [Hanley and McNeil 1982].

5. EXPERIMENTAL RESULTS

In this section, we show how our methodology can be applied to the chosen datasets.

5.1 Dataset characterization

Our dataset characterization is based on an estimate of the distribution, for each dataset, of three sets of pairwise distances: (1) **Inlier-Inlier** ($P_{i,i}$) — Distance between points/objects with normal behaviour; (2) **Inlier-Outlier** ($P_{i,o}$) — Distance between one point with normal behaviour and one anomalous point; (3) **Outlier-Outlier** ($P_{o,o}$) — Distance between anomalous points. The rationale is that the larger the Inlier-Outlier distances compared with Inlier-Inlier distances, the more detectable the outliers are. The Outlier-Outlier distances indicate the extent to which the outliers are clustered. We then plot the Cumulative Distribution Function (CDF) for each set of pairwise distances and measure the largest frequency difference for a given distance. Figure 1 shows the CDFs for two datasets, Ionosphere and WPBC, where it is clear that the outliers of Ionosphere should be easier to detect than those from WPBC. In particular, Fig. 1a shows that 60 % of the Ionosphere's $P_{i,i}$ and $P_{o,o}$ pairs are shorter than the $P_{i,o}$ pairs, while the distributions for WPBC basically overlap. Table I shows the maximum distance between the CDFs for each dataset, which is an indication of their difficulty in terms of anomaly detection. In the next section, we evaluate the ability of the base detectors to detect outliers in the datasets chosen.



(a) Ionosphere (b) WPBC

Fig. 1: CDF for pairwise distances.

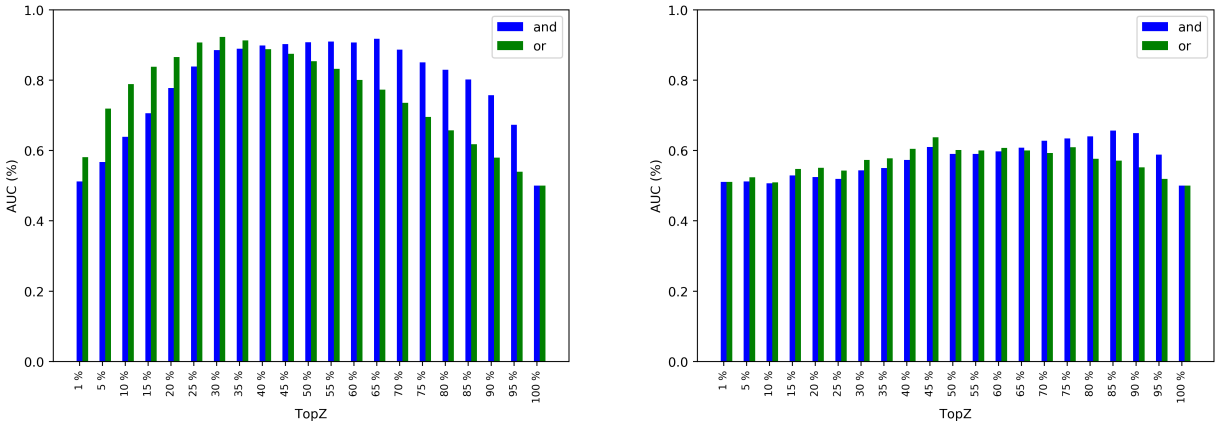
5.2 Base-detectors Analysis

Our base-detectors analysis examines the ranking of the TOP-Z% outliers for each base detector and each dataset, that is the first Z% of the points in terms of *outlierness*. We assess two combination strategies of base detectors: (i) *and* – an outlier is present in all rankings (conjunctive or intersection), and (ii) *or* – an outlier is present in at least one ranking (disjunctive or union).

Since there are 12 *base-detectors*, making a total of $\sum_{i=1}^{12} C_i^{12}$ possible combinations per Top-Z%, we take note of the best results for each algorithm (i.e., the k that provides the best AUC performance [Campos et al. 2016]).

The results for the Ionosphere and WPBC datasets can be seen in Figure 2, where there are tradeoffs between the two combination strategies, in particular for Ionosphere, where the disjunctive strategy is able to provide a better AUC for smaller values of Z; however, as Z increases, its AUC decreases, as a result of an increasing number of inliers in the rankings.

The conjuntive strategy is more effective for higher values of Z, since it is more selective. On the other hand, in the case of WPBC, it can be seen that the strategies operate in quite a similar way and their performance does not vary significantly as a function of Z; this means that it is much harder to detect outliers and there is not much of an opportunity for an ensemble to make improvements. It is also worth noting that the best value of Z varies for combination strategies and datasets, which makes it hard to define a general Z threshold that is effective. We also found that the best AUCs are found in combinations of about 3 base detectors, which is a relatively small fraction of the available algorithms. Moreover, the set of base detectors that achieves the best results also differs among the datasets, which again makes it hard to select the algorithms that can be combined in advance. It was also found that, in most cases, increasing the value of Z above 50% does not make a significant improvement in terms of AUC, since there begins to be a risk of misclassifying objects. The scenario is even worse for datasets that only contain a few outliers, and where larger Z values just treat inliers as outliers. Again, setting a threshold for all the algorithms does not seem to be doable. Table I shows the best values for conjunctive, column “AUC(and)", and disjunctive “AUC(or)" combinations for each dataset, which further demonstrates that some datasets are inherently harder to handle.



(a) Ionosphere

(b) WPBC

Fig. 2: Best combination for Union(or) and Intersection (and) operations given an specific Top-Z%.

5.3 Stacking Evaluation

After characterizing datasets and assessing the capabilities of base detectors, it can be concluded that combining the results, in terms of outlier rankings, is not an easy task, in particular with regard to base detectors, their parameters and combination settings. We have thus decided to use stacking, a meta-classifier, which will learn on the basis of the outcomes of the base detectors. In practice, we build a meta-classifier for each dataset, where its inputs are the “outlierness" scores of each base detector, and we can learn from this whether this combination of scores refers to an inlier or an outlier.

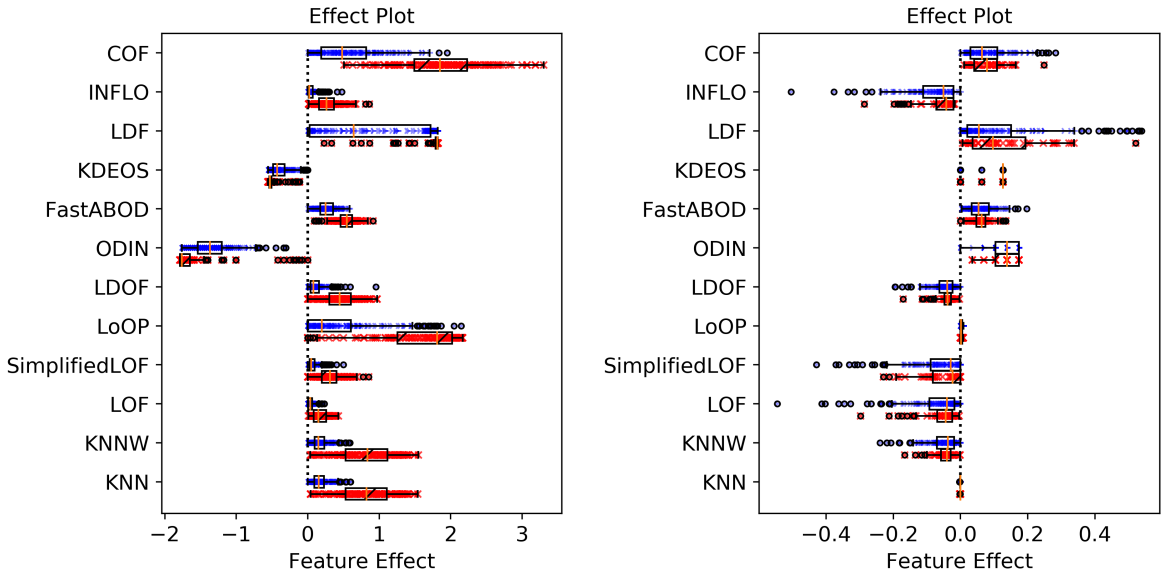
Logistic regression is used as a learning function and a stratified 5-fold cross validation procedure is employed since some datasets have a smaller number of outliers. Parameter tuning was calibrated within the training set, and the degree of accuracy was determined in the test stage. Hence, all results show the average and standard deviation among 5 executions. For purpose of comparison, we carried out the same experiments using the Isolation Forest algorithm [Liu et al. 2008] instead.

Table I shows the performance of Isolation forest compared with that of our approach considering the AUC metric. Our stacking method provided the best results in most datasets. Both techniques have a similar performance in the WPBC dataset because of the difficulties described earlier.

5.4 Understanding the Performance

The last step in our methodology aims to understand the contribution made by each base detector, which is achieved through the use of interpretability assessment procedures. Here we use effect plots, which show the distribution of weighted scores to the meta-classifier outcomes [Molnar 2018]. The X-axis shows the feature that effects the box-plot in the score of outliers/inliers for a specific base method, while the Y-axis shows the base-detector name followed by the parameter k of the best result in the dataset. We then produce two "rows", for each base detector, one for inliers and one for outliers. Each row contains a boxplot of the weighted scores and also the actual points that generate the boxplot. Blue "$\succ$" are inliers, whereas red "$\times$" are outliers of the dataset. The input for the effect plot was normalized between 0-1 as part of the pre-processing stage, as well as for ease of comprehension.

Since the logistic regression is an additive interaction in the log space, we can visualize how each base detector contributes to the anomaly detection task. Figure 3 displays the outlier effect plots for Ionosphere (Fig 3a) and WPBC (Fig 3b). It should be noted that some base-detectors have a negative effect (that is, their box plot and points appear to the left of the no effect dotted line) and some have a positive one, which shows how they contribute to the outcome. A negative effect may indicate that the outcomes of the associated base detector are the inverse of what is desired, and they are used as "inhibitors". Note that that these inhibitor base detectors vary across datasets, which confirms that it is not possible to identify which detectors provide the best result a priori.



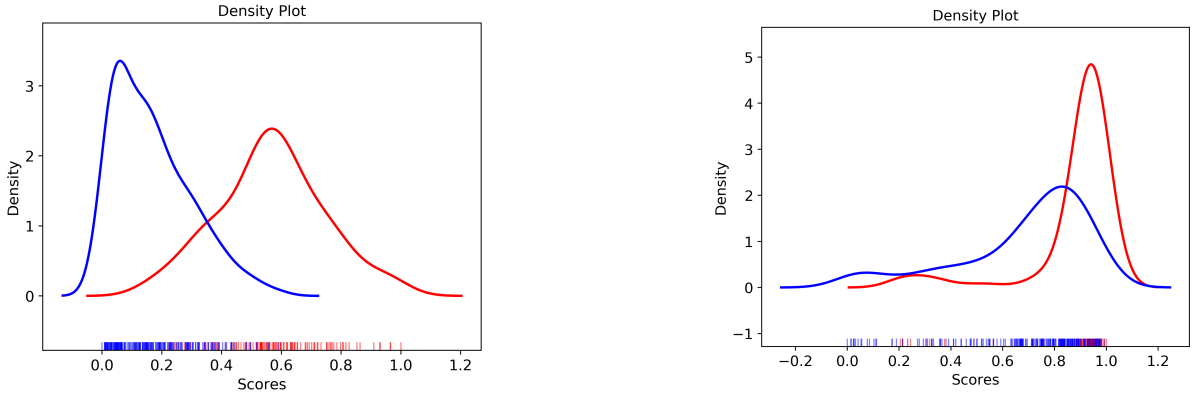
(a) (b)

Fig. 3: Effect plot by class for Ionosphere (a) and WPBC (b). Outliers are presented as red "$\times$" and inliers as blue "$\succ$". The feature effect plot shows the distribution of effects per base-detector.

By comparing the boxplots from the inliers with the outliers, we can have a clearer understanding of why our stacking ensemble performed better for Ionosphere. In this case, it can be seen that the boxplots for outliers and inliers do not overlap, which suggests that the base detectors are able to

distinguish between them. On the other hand, when 3b is checked, it is evident that the boxplots for all base detectors do overlap for WPBC, that is, none of our base-methods is able to distinguish properly between outliers and inliers.

The base detector performance can be more clearly understood by checking the score distributions, as depicted for two Ionosphere cases in Figure 4, where the "red" line represents the outliers and the "blue" lines the inliers. The leftmost graph, from COF, shows that it was able to separate inliers from outliers, where the rightmost graph, from KDEOS, shows that the score distributions overlap significantly, which explains its performance.



(a) COF (Best $k = 100$)

(b) KDEOS (Best $k = 71$)

Fig. 4: Score density distribution by class for COF and KDEOS, respectively. The "red" line represents the outliers and "blue" lines inliers.

Finally, it should be noted that these tradeoffs are summarized in Table II where the medians for the weighted scores are shown for inliers and outliers, as well as for each dataset and base detector. The last row shows the number of base detectors that were able to differentiate between inliers and outliers for each dataset (these cases are highlighted in bold).

6. CONCLUSIONS AND FUTURE WORK

We set out a basic methodology to characterize and understand outlier detection, and this involved designing and evaluating a stacking ensemble. Our approach incorporates the best learning functions from multiple models as a means of improving its predictive capability in a wide range scenarios. Additionally, we evaluated the trade-offs in the parameter and threshold definitions and examined the datasets in datasets in order to understand the challenges they raise. Such evaluation may subsidize improvements on our stacking strategy, which we foresee as acting as a significant future work direction. Input files from previous resources and references are openly accessible through the repository which is available online at `https://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/`.

REFERENCES

Aggarwal, C. C. Outlier ensembles: Position paper. *SIGKDD Explor. Newsl.* 14 (2): 49–58, Apr., 2013.

Aggarwal, C. C. and Sathe, S. Theoretical foundations and algorithms for outlier ensembles. *SIGKDD Explor. Newsl.* 17 (1): 24–47, Sept., 2015.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. SIGMOD '00. Association for Computing Machinery, New York, NY, USA, pp. 93–104, 2000.

Base Detector	Class (I=Inlier, O = Outlier)	Datasets Effect plot median										
		Ionosphere	Lymphography (1-of-n)	PenDigits	WBC	WDBC	WPBC	Arrhythmia	Cardiotocography	InternetAds	PageBlocks	Pima
KNN	I	**0.1528**	**0.2869**	**0.2473**	**0.0559**	**0.0492**	-0.0005	0.1270	0.9894	0.6038	**0.1709**	0.2161
	O	**0.8234**	**0.5999**	**0.9615**	**1.0567**	**0.1797**	-0.0006	0.2337	1.3405	0.8277	**0.7722**	0.3883
KNNW	I	**0.1492**	**0.3422**	**0.2002**	**0.0299**	**0.0429**	-0.0372	0.1149	-0.0846	0.6410	**0.0728**	0.1519
	O	**0.8353**	**0.6557**	**0.7981**	**0.7710**	**0.1529**	-0.0402	0.2051	-0.1085	0.9368	**0.2718**	0.2733
LOF	I	**0.0342**	**0.1241**	**0.0044**	**0.0297**	**0.0519**	-0.0404	0.0242	-0.0026	0.7924	-0.0056	0.0165
	O	**0.1487**	**0.7161**	**0.0269**	**0.9985**	**0.2803**	-0.0445	0.0816	-0.0048	1.3657	-0.0218	0.0297
SimplifiedLOF	I	**0.0508**	**0.2794**	**0.0042**	**0.0876**	**0.0894**	-0.0288	0.0649	-0.1107	0.7245	0.0124	-0.1055
	O	**0.3093**	**0.7441**	**0.0143**	**0.8598**	**0.3125**	-0.0237	0.1110	-0.1492	1.0759	0.0290	-0.1448
LoOP	I	**0.1955**	**0.1575**	**0.5226**	**0.0456**	**0.3093**	0.0019	0.0377	-0.1665	-0.6335	0.1097	-0.1427
	O	**1.8109**	**0.7811**	**2.1764**	**0.4669**	**1.1428**	0.0016	0.1084	-0.3295	-0.9565	0.8688	-0.3000
LDOF	I	**0.0775**	**0.3195**	**-0.1341**	**0.0546**	**0.0991**	-0.0393	0.0305	-0.1509	-0.2788	**0.0723**	-0.2538
	O	**0.4469**	**0.7700**	**-0.2807**	**0.3049**	**0.3388**	-0.0387	0.0556	-0.1918	-0.4068	**0.2601**	-0.3137
ODIN	I	**-1.3692**	**0.3190**	**0.5798**	**0.1141**	**0.2988**	0.1387	0.0839	1.1738	-2.5461	-0.0260	0.4117
	O	**-1.7604**	**0.6479**	**0.8308**	**0.2184**	**0.4004**	0.1387	0.1093	1.3026	-2.5511	-0.0366	0.4729
FastABOD	I	**0.2492**	**0.2527**	**0.3271**	**0.1583**	**0.1186**	0.0536	0.0171	-1.5043	1.0542	-1.4073	1.2069
	O	**0.5461**	**0.3248**	**0.5860**	**0.4782**	**0.3809**	0.0630	0.0303	-1.5396	1.3390	-1.6357	1.7317
KDEOS	I	**-0.4266**	**0.2520**	**0.2719**	0.2494	**0.3004**	0.1266	0.6605	0.3482	0.0111	0.1991	0.2237
	O	**-0.5271**	**0.4128**	**0.3387**	0.2824	**0.3542**	0.1266	0.7074	0.3724	0.0118	0.2500	0.2237
LDF	I	**0.6443**	**0.1770**	**0.1350**	**0.0183**	**0.2273**	0.0544	0.0385	0.0652	-0.8990	0.0541	0.0152
	O	**1.8160**	**0.8448**	**1.2992**	**0.8444**	**0.9664**	0.0964	0.3141	0.1666	-1.6837	0.2601	0.0291
INFLO	I	**0.0241**	**0.1515**	**-0.0119**	**0.0216**	**0.0433**	-0.0504	0.0253	-0.0359	0.3026	-0.0391	-0.0340
	O	**0.2629**	**0.7726**	**-0.0530**	**0.8597**	**0.2400**	-0.0425	0.0636	-0.0647	0.4975	-0.0959	-0.0890
COF	I	**0.4810**	**0.4589**	**0.2447**	**0.2729**	**-0.0098**	0.0642	0.1155	0.5443	0.0373	-0.0386	0.3893
	O	**1.8523**	**0.7273**	**0.6673**	**0.7684**	**-0.0251**	0.0791	0.2012	0.6022	0.0463	-0.0520	0.7281
Total of Techniques where boxplot do not overlap		12	12	12	11	12	0	0	0	0	3	0

Table II: Effect plot median for all datasets. The values in bold means that effect plot for inlier and outlier differ. It indicates the detectors capability to distinguish both classes.

Campos, G., Zimek, A., and Meira Jr., W. An unsupervised boosting strategy for outlier detection ensembles. In *Advances in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science, vol. 10937. Springer, Germany, pp. 564–576, 2018. Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD ; Conference date: 03-06-2018 Through 06-06-2018.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., Assent, I., and Houle, M. E. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery* 30 (4): 891–927, Jul, 2016.

Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–15, 2000.

Dua, D. and Graff, C. UCI machine learning repository, 2017.

Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143 (1): 29–36, 1982.

Hawkins, D. M. *Identification of outliers*. Vol. 11. Chapman and Hall London ; New York, London [u.a.], 1980.

He, Z., Xu, X., and Deng, S. Discovering cluster-based local outliers. *Pattern Recognition Letters* 24 (9): 1641 – 1650, 2003.

Hodge, V. and Austin, J. A survey of outlier detection methodologies. *Artif. Intell. Rev.* 22 (2): 85–126, Oct., 2004.

Jin, W., Tung, A. K. H., and Han, J. Mining top-n local outliers in large databases. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '01. Association for Computing Machinery, New York, NY, USA, pp. 293–298, 2001.

Knorr, E. M. and Ng, R. T. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24rd International Conference on Very Large Data Bases*. VLDB '98. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 392–403, 1998.

Kriegel, H.-P., Kroger, P., Schubert, E., and Zimek, A. Interpreting and unifying outlier scores. In *Proceedings of the 2011 SIAM International Conference on Data Mining*. SIAM, SIAM / Omnipress, Mesa, Arizona, USA, pp. 13–24, 2011.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*. ICDM '08. IEEE Computer Society, USA, pp. 413–422, 2008.

Molnar, C. A guide for making black box models explainable. *URL: https://christophm. github. io/interpretable-ml-book* 1 (1): 1–303, 2018.

Ramaswamy, S., Rastogi, R., and Shim, K. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. SIGMOD '00. Association for Computing Machinery, New York, NY, USA, pp. 427–438, 2000.