


# AutoML for Explainable Anomaly Detection (XAD)

Nikolaos Myrtakis  

Department of Computer Science, University of Crete, Heraklion, Greece  
ETIS Laboratory, CY Cergy Paris Université, ENSEA, France

Ioannis Tsamardinos  

Department of Computer Science, University of Crete, Heraklion, Greece

Vassilis Christophides  

ETIS Laboratory, CY Cergy Paris Université, ENSEA, France

---

## Abstract

Numerous algorithms have been proposed for detecting anomalies (outliers, novelties) in an unsupervised manner. Unfortunately, it is not trivial, in general, to understand why a given sample (record) is labelled as an anomaly and thus diagnose its root causes. We propose the following *reduced-dimensionality, surrogate model approach* to explain detector decisions: approximate the detection model with another one that employs only a small subset of features. Subsequently, samples can be visualized in this low-dimensionality space for human understanding. To this end, we develop **PROTEUS**, an AutoML pipeline to produce the surrogate model, specifically designed for feature selection on imbalanced datasets. The PROTEUS surrogate model can not only explain the training data, but also the out-of-sample (unseen) data. In other words, PROTEUS produces **predictive** explanations by approximating the decision surface of an unsupervised detector. PROTEUS is designed to return an accurate estimate of out-of-sample predictive performance to serve as a metric of the quality of the approximation. Computational experiments confirm the efficacy of PROTEUS to produce predictive explanations for different families of detectors and to reliably estimate their predictive performance in unseen data. Unlike several ad-hoc feature importance methods, PROTEUS is robust to high-dimensional data.

**2012 ACM Subject Classification** Computing methodologies → Anomaly detection

**Keywords and phrases** Anomaly Explanation, Predictive Explanation, Anomaly Interpretation, Explainable AI

**Digital Object Identifier** 10.4230/OASICS.Tannen.2024.8

**Funding** The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant” (Project Number: 1941); the ERC under the European Union’s Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 617393.

## 1 Introduction

Detection of “anomalous” samples (records, instances), called *anomaly detection*, is an important problem in machine learning. It is conceptually related to outlier and novelty detection in several application settings. The anomalous samples may indicate mislabelled data, catastrophic measurements or data entry errors, bugs in data wrangling and preprocessing software, or other interesting phenomena.

Numerous **unsupervised** algorithms (e.g., IF [32], LOF [7], LODA [44]) to detect anomalies (hereafter **detectors**) have been proposed. The most advanced ones detect anomalies in a multi-dimensional fashion, simultaneously considering all feature values. Unfortunately, detectors, in general, do not explain why a sample was considered as abnormal, leaving human analysts with no guidance about their root causes, insight to take corrective actions, or remedy their effect.



© Nikolaos Myrtakis, Ioannis Tsamardinos, and Vassilis Christophides;  
licensed under Creative Commons License CC-BY 4.0

The Provenance of Elegance in Computation – Essays Dedicated to Val Tannen.

Editors: Antoine Amarilli and Alin Deutsch; Article No. 8; pp. 8:1–8:23

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Several methods for **explaining anomalies** have been proposed, hereafter **explainers**. *The explanations often take the form of a subset of features called a **subspace** in the literature. The idea is that by examining only the explaining features suffices to determine whether the sample is an anomaly or not according to the detector.*

Existing methods can be categorized to those that provide **local explanations** (point-based) that pertain to a single sample, or **global explanations** (a.k.a. set-based) to simultaneously explain all training samples. The latter is important in order to reduce the burden of human analysts to inspect possibly different explanations for each anomaly. We should stress that global explanation is different from clustering as the former’s objective is to provide a subspace segregating the anomalous from normal samples. Explainers may be **specific** to a detection algorithm or **detector-agnostic**, hence applicable post-hoc to any detection algorithm. As reported by several independent experimental studies, e.g. [17], there is no detector outperforming all others on all possible datasets. Hence, researchers cannot just design a specific explainer for the optimal detector; it may thus be preferable to design optimal agnostic explainers. Explainers may also be categorized as **descriptive** in the sense that they explain the samples used to train the detector. Explainers that return explanations that generalize to unseen data are **predictive** ones. The importance of predictive explanations has been recognised in Explainable AI to avoid recomputing explanations on every new batch of data.

Figure 1 illustrates how predictive explanations can be used in data validation pipelines monitoring the data fed to downstream ML models. Given that in real application settings it is difficult or even impossible to label data as anomalous or normal [17], unsupervised detectors are initially used to spot anomalies. Then, a predictive anomaly explainer could be used by human analysts to reveal the root causes of the detected anomalies and decide subsequent corrective actions. It is essentially a surrogate model <sup>1</sup>, trained with a small subset of the original features that serve as explaining feature subspace. Depending on the quality of the approximation of the decision boundary of an unsupervised detector, the surrogate model can be also used to detect anomalies in fresh data, i.e., new batches of data, by completely bypassing the need to rerun the detector.

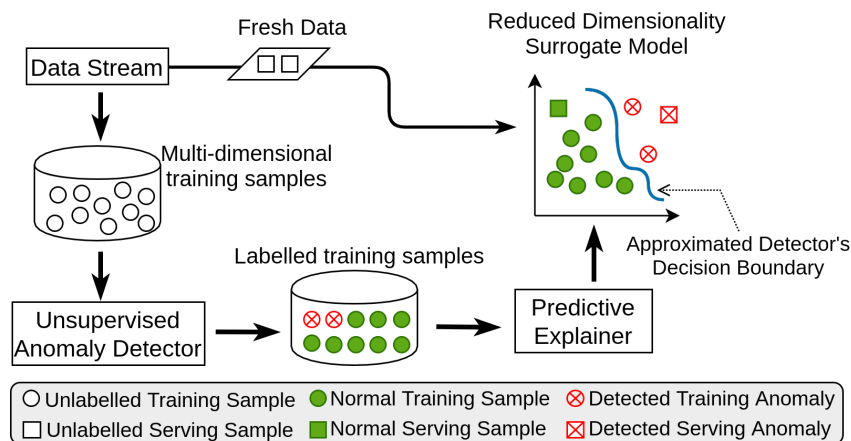
In this paper, we propose a **novel method to produce global, predictive explanations** called **PROTEUS**<sup>2</sup>. PROTEUS is *detector agnostic*, and can be used to approximate the decision boundary of any detector. We should stress that prior work on detector agnostic explainers like CA-Lasso [38] and SHAP [34] but also detector specific explainers like LODA [44] produce explanations that are only **local and descriptive**.

PROTEUS essentially constructs a *reduced-dimensionality, surrogate model* that approximates the behavior of a detector with fewer features. Since the detector is labelling the samples as anomalies or not, the problem of finding such a model reduces to a *supervised predictive modeling with feature selection* problem. In order for the surrogate model to also explain unseen samples, it has to approximate the detector’s decision boundary and not simply interpolate the anomalies (overfit) in the training data. To this respect, *the quality of approximation should be estimated using out-of-sample performance estimation protocols* like *K-fold cross validation (CV)*. To build the model, any combination of feature selection algorithm with a classifier could be employed. However, ideally one should optimize the combination of algorithms and their hyper-parameter values to achieve the best approximation with the samples at hand.

---

<sup>1</sup> A surrogate model is an interpretable model that is trained to approximate the predictions of a black box model [39]

<sup>2</sup> Proteus or *Πρωτεύς* in Greek, means “first” and is a minor sea God and son of Poseidon.



■ **Figure 1** Predictive Anomaly Explanation Pipeline.

The above requirements for tuning and estimating generalization performance of predictive models are nowadays addressed in the automated machine learning (**AutoML**) systems [23]. In this respect, **producing predictive anomaly explanations can be solved as an AutoML problem**. Unfortunately, the majority of existing tools such as auto-sklearn do not perform feature selection. In addition, they do not exploit the fact that the data can be augmented with new samples (pseudo-samples) that can be labelled by the detector, to improve performance. Finally, their performance estimates are often overestimated [51], particularly for imbalanced datasets. To address the above issues, PROTEUS makes the following contributions:

- (1) In Section 2, we introduce a novel AutoML engine specifically designed to support feature selection and classification on imbalanced datasets. Unlike existing explainers, PROTEUS outputs not only a *small-sized feature subset serving as explanation* but also a *surrogate model fitted on this subset* to explain unseen samples, as well as a *reliable out-of-sample (predictive) performance estimation*.
- (2) To produce such output, PROTEUS AutoML relies on *advanced design choices* described in Section 3, such as supervised oversampling, group-based stratification, and a special variant of Cross-Validation with Bootstrap Bias Correction (**BBC**) [53].
- (3) Thorough computational experiments presented in Section 4 we show the efficacy and robustness of PROTEUS in synthetic and real datasets of increasing dimensionality. Last but not least, our experiments show that PROTEUS approximates accurately the performance of a specific explainer (LODA) in a detector-agnostic fashion.

This is a substantial extension of our short paper published at ICDE 2021 [41]. This paper's additional contributions are four-fold:

- (4) We formally define in Section 2 descriptive and predictive explanations originally introduced in our work.
- (5) We assess in Section 5 the merit of the idea to use PROTEUS to correct the decisions of the unsupervised anomaly detectors. Specifically, we study the disagreements of classification to anomalies between the surrogate PROTEUS model and the detector. We show that PROTEUS can often correct the false positives of false negatives as identified by the detector.
- (6) We propose a new visualization method for presenting the global explanations found by PROTEUS as spider charts. The visualizations provide insight regarding the combination of feature values that lead to calling a sample as anomalous or not.

(7) We survey in Section 6 several categories of related work on explaining anomalies in unsupervised and supervised settings, positioning the predictive explanation of PROTEUS w.r.t. each category.

Finally, in Section 7 we conclude the paper and discuss directions of future research.

## 2 Problem Definition

In this section, we formalize the notion of descriptive explanations inspired by [19] and we introduce the novel concept of predictive explanations.

Let  $D = \{x_1, \dots, x_n\}$  be a dataset of  $n$  samples, where each sample  $x \in \mathbb{R}^d$ . An **Anomaly Detector**  $A$  is essentially a function that scores the “anomalousness” of samples in  $D$  according to an unsupervised **Anomaly Model**:  $\omega_A : \mathbb{R}^d \rightarrow \mathbb{R}$ . Continuous scores are then converted into dichotomous decisions using a threshold choice method [55]. Given a threshold  $T$  and sample  $x \in D$ , a **binary** Anomaly detector is a function  $\omega_A^l : \mathbb{R} \rightarrow \{0, 1\}$  defined as follows:  $\omega_A^l(x) = \mathbb{1}[\omega_A(x) > T]$ . The value  $\omega_A^l(x) = 1$ , semantically denotes the identification of an anomaly.

► **Definition 1.** *The descriptive explanation  $\mathcal{D}$  of a set of anomalies  $O = \{x \mid \omega_A^l(x) = 1, x \in D\}$ , is a subset of features  $S$ , where  $|S| = b \ll d$ , that maximizes the **cumulative score** for a set of anomalies:*

$$\begin{aligned} \mathcal{D} = \operatorname{argmax}_S \quad & \sum_{x \in O} \omega_A(x[S]) \\ \text{s.t.} \quad & |S| = b \end{aligned} \quad (1)$$

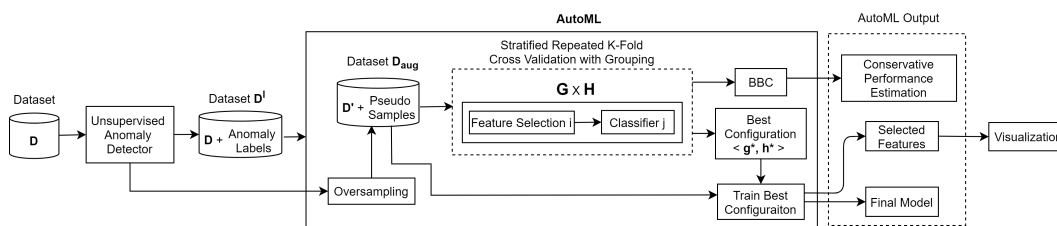
where  $[\cdot]$  denotes the projection of  $x$  over the features in  $S$  composing its explanation.

A global descriptive explanation algorithm strives to reveal the subspace that maximizes the cumulative anomaly score for a set of identified anomalies, given a specific anomalousness criterion such as distance, isolation etc. Such explanations are called *descriptive* as they are computed for every new batch of anomalous and normal samples. In order to make explanations also *discriminative* for unseen data, we need to consider *predictive* explanations i.e., a hyperplane of reduced dimensionality that separates the anomalies from the normal samples when training a classifier over the output of an unsupervised anomaly detector. To produce explaining hyperplanes we need to evaluate alternative surrogate models built using different classification algorithms  $h \in H$ , where  $h$  is fitted in a lower dimensional space, produced in turn by different feature selection algorithms  $g \in G$  that consider the labels returned by different anomaly detectors.

In a nutshell, *predictive* explanations are produced by solving an AutoML problem [23]. We denote the combination of an algorithm  $g$  and  $h$  with their respective hyper-parameter values  $a$  and  $b$  as a configuration  $\theta$ , which is a function  $f = h(g(\cdot, a), b)$ . The function  $f$  first applies the specified feature selection algorithm  $g$  with hyper-parameters  $a$  to some input data and the result is then used to train a classifier  $h$  with hyper-parameters  $b$ . Let  $D^l = \{(x_1, \hat{y}_1), \dots, (x_n, \hat{y}_n)\}$  be the augmented dataset  $D$  enriched with the anomaly labels as indicated by the detector model:  $\hat{y}_i = \omega_A^l(x_i)$ .

► **Definition 2.** *The predictive explanation  $\mathcal{P}$  is the hyperplane that comprises of a minimal subset of features  $S$  leading to an optimal surrogate model  $h$  w.r.t. a performance metric  $Q$ :*

$$\mathcal{P} = \operatorname{argmax}_S \max_h Q(h(D^l[S]))$$



■ **Figure 2** Proteus AutoML Pipeline for Anomaly Detection and Explanation.

Given the dataset  $D^l$ , the objective is to build a reduced-dimensionality surrogate model  $f$  trained with some data  $D_{train}^l$  to best approximate the detector's decision boundary. To assess the quality of the approximation,  $f$  has to generalize to unseen data  $D_{test}^l$  which were not used during the training of  $f$ . Therefore, the objective is to find the configuration  $\theta^*$  that contains the tuple  $\langle h^*, g^*, a^*, b^* \rangle$  maximizing a performance metric  $Q$ :

$$\theta^* = \underset{\theta}{\operatorname{argmax}} Q(f(D_{train}^l, \theta), D_{test}^l) \quad (2)$$

The last step is to train the best configuration using all available data, to produce the final surrogate model  $f(D^l, \theta^*)$  i.e., a model  $h^*(D^l[S], b^*)$  that is used to predict the “anomalousness” of unseen samples using only a subset of features  $S = g^*(D^l, a^*)$ .

As anomalies are rare, the quality of performance of a predictive explanation requires evaluation metrics that are insensitive to the class distribution. In this respect, PROTEUS relies on optimizing the area under the Receiver Operating Characteristic (ROC AUC) curve (hereafter **AUC**). Given a minimal subset of features and a classifier, *AUC equals the probability that the classifier scored higher an anomalous than a normal sample*. Discovering such minimal subset is a challenging task as the search space is exponential and features in the input dataset may be both *irrelevant* or *redundant* w.r.t. to the predictive outcome. PROTEUS relies on effective and efficient *feature selection* algorithms [31, 52, 50] to extract predictive explanations in a supervised setting.

### 3 Producing Global, Predictive Explanations with PROTEUS

Figure 2 illustrates the main steps of the pipelines automatically generated by PROTEUS. We proceed with explaining each step as well as the underlying design choices.

**Producing Predictive Explanations as a Supervised Task.** First, the anomaly detector runs in dataset  $D$  for producing the anomaly scores which are then transformed into binary labels (anomaly or not) in dataset  $D^l$ . Producing a surrogate model of lower dimensionality becomes a supervised, binary classification task with feature selection, where the outcome is the label of the unsupervised detector. We note that *data are standardized* for subsequent steps.

**Oversampling.**  $D^l$  is expected to be highly imbalanced (w.r.t. the outcome), as anomalies are rare. Imbalanced datasets are statistically challenging for any ML classifier. One technique to alleviate the problem is *oversampling* the minority class. We focus on *synthetic minority oversampling*, i.e., the samples are perturbed by adding noise to the values of the features, creating new samples called *pseudo-samples*. In common (unsupervised) oversampling methods, for small enough perturbations the pseudo-samples are *assumed* to remain in the

minority class. An assumption that strongly depends on the definition of what is considered “small-enough”. However, one can take advantage of the detector model produced in the first step is available to query regarding the label of a pseudo-sample. In other words, PROTEUS oversampling is *supervised* as in case of explanation methods for black-box predictive models [45]. Intuitively, oversampling probes the region around the anomalies and perturbs these samples to examine if they cross the detector’s decision boundary or not. It thus effectively increases the available sample size for the classification, potentially increasing the quality of the approximation with the surrogate model. For each anomalous sample  $a$  it produces  $ps$  pseudo-samples per anomaly by adding a perturbation vector  $p$  to  $a$ :  $a' \leftarrow a + p$ . Each  $p$  follows a multi-variate ( $d$ -dimensional) normal distribution with zero mean and an isotropic, diagonal, covariance matrix  $\sigma I$ ;  $\sigma$  is a hyper-parameter of the algorithm which we set to 0.1 for all the computational experiments. If  $a'$  is labelled as an anomaly it is appended to the oversampled dataset  $D_{aug}$ , otherwise, another pseudo-sample is produced.

**Hyper-Parameter Optimization Space.** To produce small-sized explanations, PROTEUS relies on feature selection algorithms, while to produce the surrogate model, a classifier is required. Most classification algorithms also accept a set of hyper-parameter values that also need to be tuned. We will call a combination of feature selection and classification algorithms and their hyper-parameters values as a *configuration*. *Each configuration is a pipeline that accepts a dataset and produces a classification model and corresponding selected features.* PROTEUS searches the configuration space for the one that leads to an optimal model by performing a simple grid search. This is, the search space of configurations is formed by the Cartesian product  $\mathcal{G} \times \mathcal{H}$  (see Figure 2) where  $\mathcal{G}$  ( $\mathcal{H}$ , respectively) is the set of all feature selection (classification) algorithms with bounded hyper-parameter values. As our choices for feature selection algorithms, we include the Statistical Equivalent Signatures (SES) [31], Forward-Backward with Early Dropping (FBED) [52], and Lasso. All of them guarantee to return the optimal feature subset (Markov Blanket in Bayesian Networks) under certain broad (but different for each algorithm) conditions, removing not only *irrelevant*, but also *redundant* features. In general, SES and FBED tend to return smaller feature subsets than Lasso, with a small drop in predictive performance [52].

Moreover, as anomaly explanation targets human analysts, *we limit the number of features selected up to 10*, ranking them based on their score given by the corresponding algorithm (e.g. Lasso coefficients). We selected linear as well as non-linear classifiers considering two facts (a) the extensive experimental results of [12], (b) the fact that deep neural network architectures are almost certain to overfit in very low sample sizes, both in terms of total sample size and the size of the rare class. The present selection of classifiers comprises of: (i) Support Vector Machines, (ii) Random Forest and (iii) K-Nearest-Neighbors. Due to space constraints, we report the hyper-parameters in our GitHub repository<sup>3</sup>. Finally, the number of pseudo-samples to create per anomaly, called  $ps$  is also tuned as a hyper-parameter taking values in  $\{0, 3, 10\}$ . Of course, additional classifiers and feature selection algorithms can be easily integrated in PROTEUS. In total, PROTEUS tried 1800 configurations.

**Estimating Performance for Tuning.** What is considered as the optimal configuration, out of all tried, is *the one that leads to models with the highest expected out-of-sample (unseen samples) predictive performance*. It is important to estimate this quantity accurately, i.e., with small variance. *A smaller variance of estimation increases the probability that the*

<sup>3</sup> <https://github.com/myrtakis/PROTEUS>



*truly optimal configuration will be selected, and thus improves the quality of the final model.* Estimation is challenging when there are only few anomalies in the dataset. Indicatively, the synthetic dataset used in our experiments (see Section 4.1) contains 10 anomalies out of 867 samples.

To estimate the expected out-of-sample performance, PROTEUS employs a *Stratified, R-Repeated K-fold Cross Validation with Grouping* protocol. We now explain each part of the protocol. We assume that the reader is familiar with the Standard *K*-fold Cross Validation (**CV**, hereafter). The *Stratified CV* is a variant where the partitioning to folds is performed under the constraint that the distribution of the classes in each fold is approximately the same as the one in the full dataset [53]. Stratification reduces the variance of estimation for imbalanced data and classes with very few samples (*ibid*). To further reduce the variance of estimation we repeat the CV process multiple times *R* and take the average (*R-Repeated CV*). Multiple repeats reduce the variance component due to the stochasticity of the specific partitioning [53]. Finally, we come to *Grouping*. By CV with Grouping we indicate a variant of CV that handles grouped samples (a.k.a. as clustered samples in statistics, not to be confused with clustering of samples). These are samples that are not independently sampled and may be correlated given the data distribution. Such samples are repeated measurements on the same subject, as an example. In our context, *an anomaly and its pseudo-samples are grouped*: information from a pseudo-sample in the training set *leaks* to predicting the corresponding anomaly in the test fold. To avoid information leakage, CV with grouping partitions to folds with the constraint that all samples of a group remain in the same fold. In our experiments, we set the number of folds  $K = 10$  and the repeats  $R = 5$ . Hence, each application of the current version of PROTEUS trains  $(K \cdot R \cdot \# \text{ Configurations} + 1) \cdot ps = 90,003$  models.

**Producing the Final Surrogate Model and Feature Subset.** The final model is trained using all available samples (the full  $D_{aug}$ ) with the best configuration found, denoted with  $\langle F^*, C^* \rangle$  in Figure 2. This configuration also produces the final subset selection (anomaly explanation). The reasoning is that most algorithms (and hence, configurations) are expected to produce better quality models and improved feature selection with more available sample. The models trained during the CV are only employed for selecting the optimal configuration and providing estimates.

**Estimating the Out-of-Sample Performance.** We now consider how the performance estimate of the final model is produced. Let us assume that 1000 configurations are tried and the best found has a CV estimate of 0.90 AUC. Unfortunately, *the CV estimate of the best configuration is optimistic and should not be returned*, i.e., the actual AUC is expected to be lower. The reason is that our estimate is the best out of 1000 tries [52, 24]. The phenomenon is conceptually similar to the multiple hypothesis testing problem in statistics. In small sample sizes, the over-optimism is particularly striking. Recent work shows that most AutoML tools do not correct for this optimism [51]. In this respect, we apply the Bootstrap Bias Correction (**BBC**, hereafter) to our CV estimates [53] that corrects for this optimism. *This leads to returning conservative estimates of performance on average.*

## 4 Experimental Evaluation

PROTEUS was implemented in Python 3.6 and evaluated on several synthetic and real-world datasets described subsequently. The code and the datasets used in our experiments are available in our GitHub repository. All experiments were performed in a Linux Desktop computer with a 4-core Intel i5 processor and 32GB of memory.

■ **Table 1** Characteristics of datasets and AUC performance of detectors during training. We denote the parent synthetic dataset as P. Synthetic, the number of features and samples as #F and #S and the anomaly ratio as A.R.

Dat. Name	#F	#S	A.R.	IF	LOF	LODA
P. Synthetic	5	867	1%	0.96	1.0	0.92
W. Br. Cancer	30	377	5%	0.95	0.94	0.96
Ionosphere	33	358	36%	0.85	0.93	0.87
Arrhythmia	257	452	15%	0.80	0.74	0.75

#### 4.1 Synthetic and Real Datasets

We focus on datasets where the samples are *independent and identically distributed (i.i.d.)* and contain numerical features. We employ a *synthetic* dataset, where anomalies have been simulated so that a minimal, global, predictive explanation (feature subset) is both achievable and known. The presence of this gold-standard allows us to evaluate how well PROTEUS identifies it. Specifically, we selected randomly one of the 100-dimensional datasets introduced in [25]. Some anomalies have been generated in a way that makes them outliers according to a subset of 2 of these features, call it  $S_{2d}$ , and some according to a subset with 3 (other) features, call it  $S_{3d}$ . Thus, the subset of these 5 features  $S = S_{2d} \cup S_{3d}$  forms the *gold-standard of global explanation for all anomalies*. On this *parent* synthetic dataset, we added irrelevant features with randomly selected values following a normal distribution with zero mean and standard deviation of one. We ended up with 5 synthetic datasets having 20, 40, 60, 80 and 100 dimensions. All of them contain 867 samples with 10 anomalies i.e., the anomaly ratio is  $\approx 1\%$ . Such datasets have been frequently used in the literature of anomaly explanation [38, 10, 26, 43], because: (a) the features in an explaining subspace (e.g,  $S_{2d}$ ) are correlated so they cannot be selected independently; (b) anomalies are recognized as such either in  $S_{2d}$  or  $S_{3d}$ , but in no other strict subset. Thus, only multivariate detection algorithms and corresponding models will achieve high performance. Hence, PROTEUS must approximate a potentially more complex model.

We additionally consider *real-world datasets* that are widely-used in the evaluation of anomaly detectors. Specifically, we selected the Wisconsin-Breast Cancer, Ionosphere and Arrhythmia, originally from the UCI Machine Learning repository, as defined for anomaly detection purposes in Outlier Detection DataSets (ODDS) repository<sup>4</sup>. They were chosen to ensure that the detectors employed achieve reasonable performance, and thus explanation makes sense. The dataset characteristics and detector performances are shown in Table 1. Wisconsin-Breast Cancer and Ionosphere contain two classes. The minority classes in both datasets are considered as anomalies. For Arrhythmia, eight sub-classes were merged to form the anomaly class. Finally, we added irrelevant features following the procedure described in synthetic datasets constructing three additional datasets per real-world dataset with 30%, 60% and 90% irrelevant feature ratio.

#### 4.2 Experimental Setting

In our experiments, we selected three widely-used unsupervised anomaly detectors that employ different anomalousness criteria, namely Local Outlier Factor (LOF) [7] as a representative of *density-based*, Isolation Forest (IF) [32] as a representative of *isolation-based* and Lightweight On-line Detector of Anomalies (LODA) [44] as a representative of *projection-based* detectors.

<sup>4</sup> <http://odds.cs.stonybrook.edu/>



Regarding the hyper-parameters, for IF we used 100 trees and 256 sub-sample size, for LOF we used  $K = 15$  and for LODA we used 100 projection vectors as proposed by the respective authors. To assess the predictive power of a surrogate model produced by PROTEUS we stratified and splitted each dataset into 70% for training and 30% was held out for testing. In each dataset, the detectors run on training and test set before adding irrelevant features. The anomaly threshold  $T$  is set as the anomaly ratio for each dataset. The detectors performances are demonstrated in Table 1.

### 4.3 Feature Importance Alternatives

We compare the original PROTEUS system, employing feature selection methods (call it PROTEUS<sub>fs</sub>), with the PROTEUS pipeline instantiated only with feature importance methods from related explanation methods. We note that these alternatives have been developed to provide *descriptive* explanations; within the PROTEUS pipeline, they are coupled with a classification model, hyper-parameter values are optimized, and they are turned into predictive explanations.

The research question to study is whether *methods specifically developed for explanations in the form of feature importance scores offer additional advantages over the feature selection methods*, everything else being equal (i.e., the rest of the PROTEUS pipeline). All alternative methods produce *local* explanations, i.e., for individual samples. Importance scores for a given feature are calculated for each sample (local scores). We compute the local scores only for the anomalous samples. To incorporate them into PROTEUS and select features for global explanations, the local scores are averaged out for each feature to produce a final feature importance score, as proposed in [33]. As a final feature selection, we select the top- $K$  features with the highest importance scores. In our experiments,  $K$  is set to 10, which is the maximum number of features allowed to be selected by PROTEUS<sub>fs</sub> and the feature importance methods. Regarding the hyper-parameters for the feature importance alternatives, we used the ones proposed by the respective authors. We evaluate the following alternatives:

- (1) Lightweight On-line Detector of Anomalies or **LODA**, hereafter, [44] is an anomaly detector that also returns local feature importance scores. LODA is included as it has shown an excellent trade-off between computational efficiency and anomaly detection performance as a detector [37]. As a feature importance method is selected as a **representative of a detector-specific explanation method**. As such, the results of its explanation method are shown only for the experiments where LODA is also used as the detector. We should stress that when comparing with LODA, the objective is to approximate its performance as the explanation is strongly coupled to the detection process. The resulting PROTEUS variant is called PROTEUS<sub>LODA</sub>.
- (2) Kernel **SHAP** (stands for SHapley Additive exPlanations) [34] is a model-agnostic method for local explanation of predictive models producing local feature scores. It is considered state-of-the-art, having outperformed LIME [45]. As Kernel SHAP does not produce a predictive model itself we consider it as a descriptive method. We use the proposed kernel as in the original publication of SHAP. Kernel SHAP is included as a representative of a **model-agnostic feature importance** method, leading to the variant PROTEUS<sub>SHAP</sub>.
- (3) **CA-Lasso** [38], is a representative of a **model-agnostic, local feature importance specifically pertaining to anomaly explanation**. It selects  $k$ -nearest neighbors per outlier  $a_i$  and  $k$  other random samples. To overcome the class imbalance, the authors oversample  $a_i$  adding pseudo-samples around it, labelling them as anomalies by

assumption, until the two classes are balanced. The explanation problem is then turned into binary classification per outlier solved with Lasso. The feature importance of each feature for  $a_i$  corresponds to the Lasso coefficients. Rather than learning the decision boundary of individual anomalies PROTEUS builds a binary classifier to explain all the anomalies spotted by an unsupervised detector. In that sense, feature selection in [38] generates local explanations per anomaly that do not generalize to unseen anomalies. Moreover, PROTEUS oversampling is supervised (by the detector) while numerous feature selection and classification algorithms along with the optimization of their hyper-parameter values. Finally, out-of-sample (predictive) performance is estimated by PROTEUS using AUC for subset selection instead of accuracy as originally proposed in [38]. The resulting PROTEUS variant is called PROTEUS<sub>CA-Lasso</sub>.

#### 4.4 PROTEUS Performance Estimation

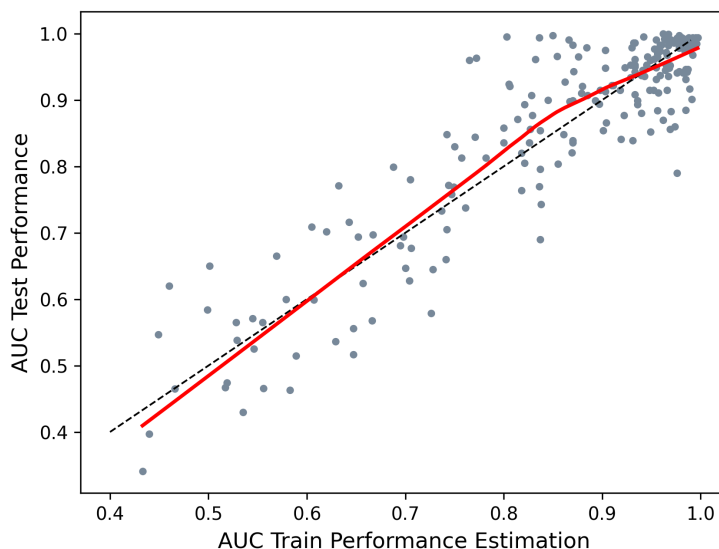
The objective of this experiment is to assess the effect of PROTEUS design choices, specifically the BBC and Grouping, to provide an accurate performance estimation. Figure 3 depicts the train estimates and test performance when PROTEUS is employed with the design choices described in Section 3, i.e., BBC and CV with Grouping. The dashed black diagonal line indicates the zero bias: points above the diagonal indicate underestimation (negative bias) and below overestimation (optimistic bias). To show the accuracy of the estimation of PROTEUS design choices, we fit a loess curve<sup>5</sup> on train and test performances for every combination (258 in total) of datasets (synthetic and real), detectors (IF, LOF and LODA) and feature selection methods (general purpose and feature importance methods). Ideally, we would want the loess curve to fit exactly the diagonal. Observe that with lower AUC performances PROTEUS tends to overestimate while with higher performances PROTEUS returns a more conservative estimation. In both cases, the points are close to the ideal diagonal line.

To further show the efficacy of the proposed design choices to provide an accurate performance estimation, in Figure 4 we compare the loess curves for train and test estimates for (i) BBC and Grouping (our design choices), (ii) no BBC (i.e., CV estimate) and Grouping (iii) BBC and no Grouping and (iv) no BBC and no Grouping. To quantify the bias for each of the four alternatives, we use the Residual Sum of Squares (RSS) to measure the discrepancy between the train and test performance. When PROTEUS is employed with BBC and Grouping (i), it gives the most accurate estimation of out-of-sample performance (with  $RSS_{(i)} = 0.05$ ) than when using any of the three alternative design choices (with  $RSS_{(ii)} = 0.88$ ,  $RSS_{(iii)} = 0.11$  and  $RSS_{(iv)} = 0.25$ ).

#### 4.5 Relevant Features Identification Accuracy

The goal of this experiment is to verify whether the features discovered during the training phase by PROTEUS<sub>fs</sub> and the feature importance alternatives are part of the gold-standard feature subset  $S$ . For this experiment we used the *synthetic* datasets. To assess the quality of the global explanation  $E$  in terms of features, we compute  $precision(S, E) = \frac{|S \cap E|}{|E|}$  and  $recall(S, E) = \frac{|S \cap E|}{|S|}$ . As we select the top-10 features to form the explanation and  $S$  contains 5 features, the *precision* for the feature importance alternative methods will be up to 0.5. The recall and precision curves are depicted in Figure 5. Feature selection

<sup>5</sup> [https://en.wikipedia.org/wiki/Local\\_regression](https://en.wikipedia.org/wiki/Local_regression)



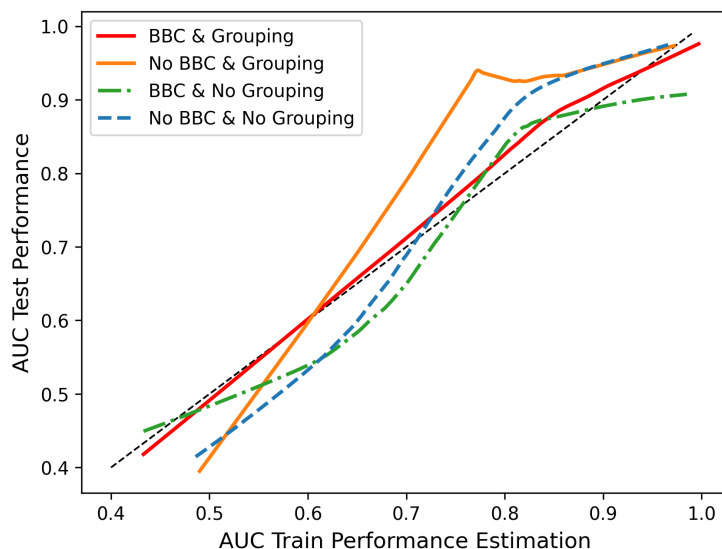
■ **Figure 3** Bias between train and test AUC performances of PROTEUS implemented with BBC and CV with grouping.

methods employed by  $\text{PROTEUS}_{fs}$  exhibit the highest precision never dropping below 0.5, independently of the employed detector or dataset dimensionality. We observed that precision is 0.5 when Lasso is selected and higher when FBED is selected. We should stress that SES was never selected by PROTEUS for the synthetic datasets. FBED removed most of the irrelevant features leading to a predictive model with less than 10 features to approximate the decision boundary of the corresponding detector.  $\text{PROTEUS}_{fs}$  achieves almost optimal recall regardless of the dimensionality and the employed detector. A slight drop in recall is observed when the precision higher than 0.8 (achieved only by FBED), while recall is optimal when Lasso is selected. Moreover,  $\text{PROTEUS}_{fs}$  feature selection methods are robust to increasing data dimensionality and irrelevant feature ratio where CA-Lasso and SHAP seem to be particularly sensitive.

#### 4.6 PROTEUS Generalization Performance

The objective of this experiment is to assess the generalization performance of PROTEUS without ( $\text{PROTEUS}_{full}$ ) and with feature selection ( $\text{PROTEUS}_{fs}$ ) as well as with the various feature importance alternatives, ( $\text{PROTEUS}_{CA-Lasso}$ ,  $\text{PROTEUS}_{SHAP}$ ,  $\text{PROTEUS}_{LODA}$ ). Figure 6 depicts the AUC performance for each method in test set. Regarding the synthetic datasets,  $\text{PROTEUS}_{fs}$  achieves very high AUC across the increasing data dimensionality with a minimum of 0.96. CA-Lasso and SHAP instead exhibit lower performances as they do not retrieve, as showed in the previous experiment, many of the relevant features. Observe that in the synthetic dataset  $\text{PROTEUS}_{fs}$  generalizes better than  $\text{PROTEUS}_{full}$ , i.e., when using all the available features.

Regarding the real datasets, similar trends are observed with  $\text{PROTEUS}_{fs}$  achieving consistently a very high generalization performance with a minimum of 0.8 in Arrhythmia in the presence of 2,570 dimensions and 90% irrelevant feature ratio.  $\text{PROTEUS}_{fs}$  seems to approximate in a detector-agnostic manner, the optimal performance of LODA's feature importance method when LODA is used as the detection algorithm. This is due to the

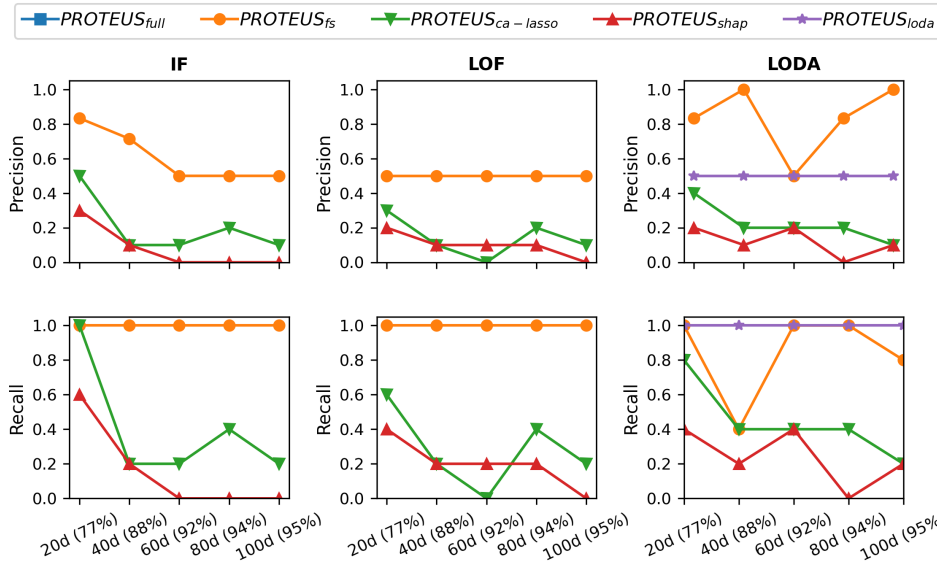


■ **Figure 4** Bias between train and test AUC performance of PROTEUS implemented with 4 alternatives.

fact that LODA’s explanations are tailored to its detection algorithm; however, if LODA’s detection performance was poor in a dataset, the provided explanation would be of less value for the analysts. The feature selection methods employed by  $\text{PROTEUS}_{fs}$ , are able to discover the relevant features leading to predictive models with very high performance regardless of the data dimensionality (and the increasing relevant feature ratio) and capture accurately the decision boundary of every employed unsupervised detector.

## 4.7 PROTEUS RunTime Performance

In the subsequent experiment, we demonstrate the execution time of the feature selection methods employed by PROTEUS. Figure 7 depicts the runtime comparison between the ad-hoc feature importance methods and the feature selection algorithms. The employed methods are specifically designed to search efficiently the exponential search space and thus require less than two seconds on average in 100-dimensions to select features, exhibiting a steady execution time. In contrast, SHAP is the most expensive method; as we had to explain the outcome of any employed detector, we used Kernel SHAP which is model-agnostic. Given the fact that SHAP is optimized only for particular families of algorithms, e.g. tree-based, its execution time is particularly sensitive to data dimensionality because the Shapley values must be calculated for all the input features. Recall that in PROTEUS we tried three classifiers resulting 30 combinations according to their hyper-parameters and three feature selection algorithms resulting 20 combinations according to their hyper-parameters including the full selector, i.e., when the full feature space is considered. Thus, the total number of configurations tried in PROTEUS is 600. Each configuration requires 2 seconds on average to complete regardless of the dataset dimensionality.



■ **Figure 5** Precision and Recall performance of discovered features when explaining IF, LOF and LODA on synthetic datasets w.r.t. increasing data dimensionality (% irrelevant feature ratio).

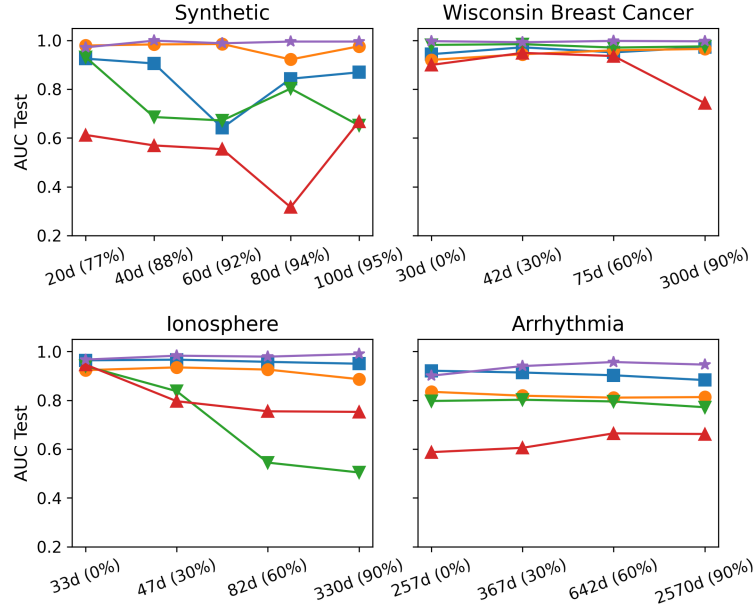
## 5 Contrasting PROTEUS Surrogate Models with Unsupervised Anomaly Detectors

In this section, we are investigating possible disagreements between PROTEUS’s surrogate models and unsupervised Anomaly Detectors, namely detected anomalies explained as normal points or vice-versa.

To assist human analysts in spotting such suspicious samples, we introduce an original visualization method based on *spider charts*. The proposed *Spider Anomaly Explanation (SAE) charts* is essentially a 2D visualisation of multivariate data projected over the explaining subspaces returned by PROTEUS. The chart has a “web-like” form with concentric circles and several spokes where each one corresponds to a specific feature. Extreme values of the features are depicted near the center or near the outermost circle. Then, a multi-dimensional sample is represented as an irregular polygon intersecting every spoke according to the quantile its feature values falls in.

In this work, we propose a variation of spider charts tailored to anomaly explanation. First, instead of plain feature values, we consider each concentric circle in the chart to represent one of the four quantiles where the center corresponds to the 0 quantile and the outermost circle corresponds to the 1 quantile. Then, every feature value is translated to a quantile ranging from 0 to 1. Hence, the normal region in the chart is the interquartile range (IQR) containing 50% of the values. Finally, we reverse the samples with extreme low values belonging to quantiles 0 - 0.25 to the 0.75 - 1 quantiles so that both low and high extremes can be identified near the outermost circle, far from the normal region. When a sample’s value intersect with a spoke in the quantiles 0.75 - 1, it means that at least 75% of values for the particular feature fall below the sample’s value.

In Figure 8 we demonstrate two SAE charts when explaining the LOF detector in the Ionosphere dataset. The explanation produced by PROTEUS comprises of 9 features. In Figure 8a both PROTEUS’s surrogate model and LOF agreed on the labels of these two samples. We can observe that the normal sample falls entirely into the normal (green) region



■ **Figure 6** AUC test performance averaged over the three detectors on synthetic and real datasets w.r.t. increasing dimensionality (% irrelevant feature ratio).

while the anomalous sample deviates significantly in every feature. Figure 8b illustrates two samples where PROTEUS’s surrogate model disagrees with LOF on their labels. LOF identified the blue sample as anomaly while PROTEUS identified it as normal. We can clearly observe that this sample was erroneously detected by LOF as it falls entirely into the normal region. For the other conflict (the red sample) it is not as obvious as in the former case because it deviates w.r.t. a subset of the features of the explanation. This sample is an anomaly according to the gold standard that was not detected by LOF. However, PROTEUS considered this sample an anomaly, extracting three features (Radar 10, 9 and 25) where it takes extreme values. We should finally stress that since PROTEUS strives to explain all the anomalies simultaneously, it is likely that an anomalous sample deviates w.r.t. a subset of the explaining subspace.

To quantify the utility of a PROTEUS explanation to reveal errors made by an unsupervised detector we introduce two metrics that rely on the gold standard available for each dataset. We consider as *conflicts* the suspicious samples for which the PROTEUS’s surrogate model predicts a different label than the detector. Subsequently, we define two sets of conflicts following the notation of Section 2 where  $\omega_A^l$  is the detector model,  $f(\cdot, \theta^*)$  is the PROTEUS’s surrogate model equipped with the best found hyper-parameters, “1” denotes an anomaly and “0” denotes a normal sample.

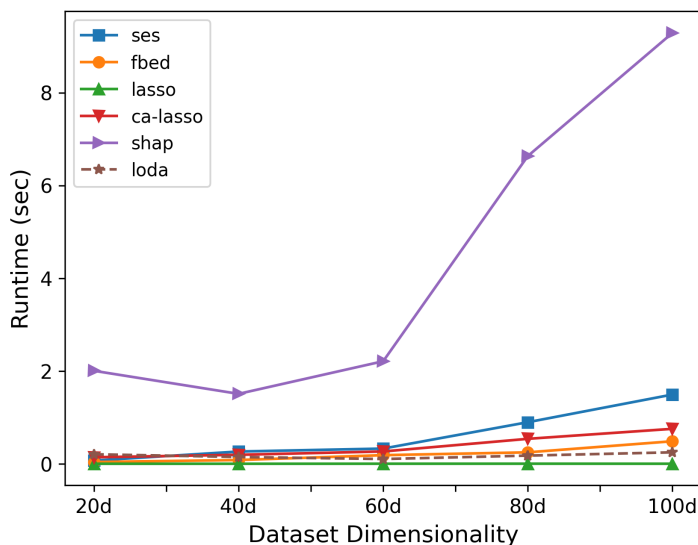
► **Definition 3.** Anomaly Normal Conflicts (ANC): *Each sample that the detector model labels as anomaly while PROTEUS’s surrogate model labels as normal.*

$$ANC = \{s \mid \omega_A^l(s) = 1 \wedge f(s, \theta^*) = 0\},$$

► **Definition 4.** Normal Anomaly Conflicts (NAC): *Each sample that the detector model labels as normal while PROTEUS’s surrogate model labels as anomaly.*

$$NAC = \{s \mid \omega_A^l(s) = 0 \wedge f(s, \theta^*) = 1\}$$





■ **Figure 7** Average runtime of feature selection/importance methods on synthetic datasets of increasing dimensionality.

Based on the two previous sets we define two metrics to quantify the utility of a PROTEUS explanation.

► **Definition 5.** True Normal Discovery (TND): *The ratio of conflicted samples that PROTEUS's surrogate model labelled correctly as normals according to the True Normals in the gold standard.*

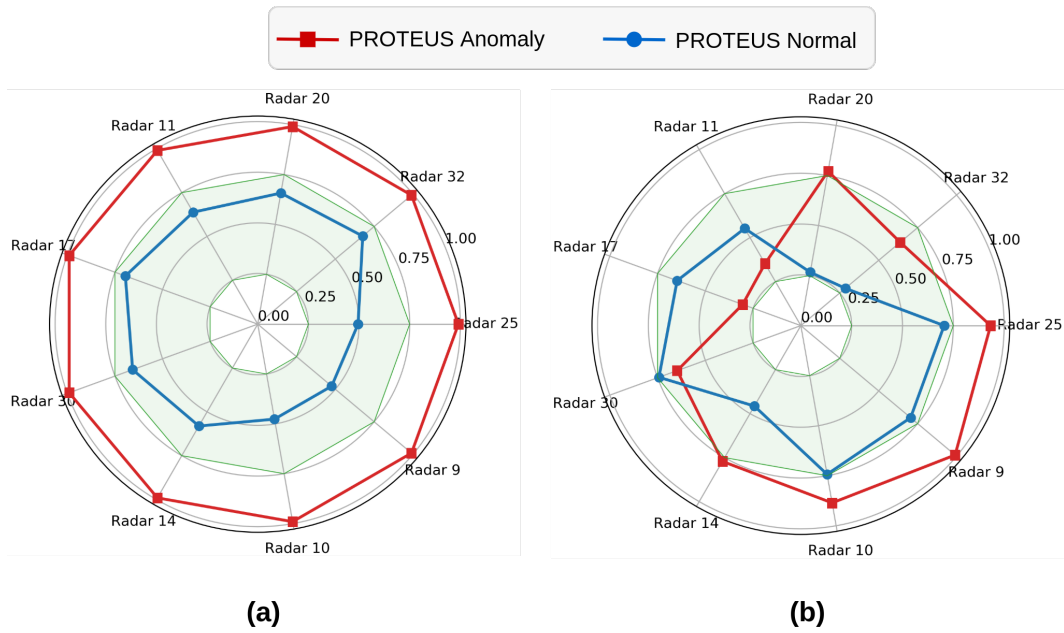
$$TND = \frac{|ANC \cap \text{True Normals}|}{|ANC|}$$

► **Definition 6.** True Anomaly Discovery (TAD): *The ratio of conflicted samples that PROTEUS's surrogate model labelled correctly as anomalies according to the True Anomalies in the gold standard.*

$$TAD = \frac{|NAC \cap \text{True Anomalies}|}{|NAC|}$$

When there are no conflicts, i.e., PROTEUS approximates perfectly the detector's decision boundary ( $AUC = 1$  on test set), the two metrics are not defined ( $ANC$  and  $NAC$  are empty). In case of conflicts,  $TND$  and  $TAD$  range between 0 and 1. Values close to 1 indicate that PROTEUS' surrogate model disagrees with the detector model and it labels suspicious samples correctly w.r.t. the gold standard. In contrast, values close to 0 indicate that PROTEUS disagrees incorrectly with the detector. Clearly, the number of conflicts is higher when PROTEUS exhibits low AUC performance.

Figure 9a contrasts the AUC of PROTEUS against the AUC of the three detectors used to analyze each real dataset. The former is computed on the test (holdout) set using the labels produced by a detector and serves as the approximation quality of its decision boundary. The latter is computed on the train set using the labels of the gold standard and reveals the effectiveness of a detector to identify anomalies in a dataset. We can easily observe that *the quality of the approximation of a detector's decision surface by PROTEUS decreases as*

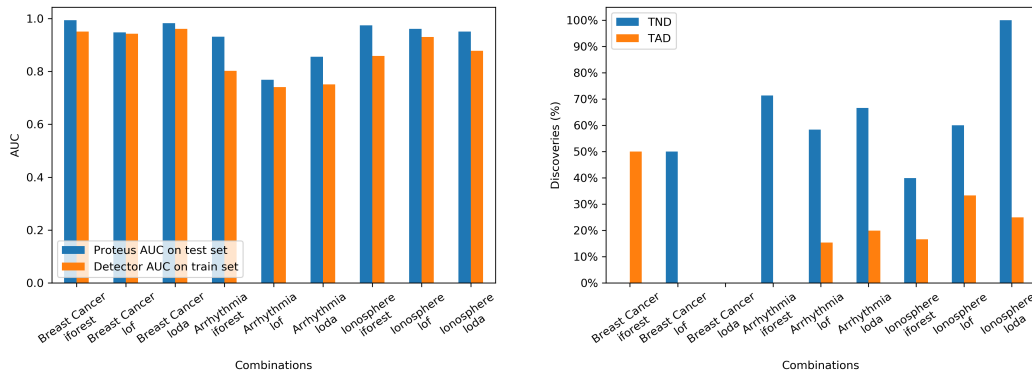


■ **Figure 8** Spider Anomaly Explanation Charts when explaining LOF in Ionosphere using PROTEUS.

*the detector’s effectiveness decreases.* For instance, in Arrhythmia we observe the lowest AUCs for the three detectors and also for the surrogate models of PROTEUS. This trend can be attributed to the fact that some misdetected samples are very hard to classify correctly without a very complex boundary. However, if the surrogate model needs to learn a more complex decision surface to segregate the misdetected samples from their neighbors, it makes the surrogate model prone to overfitting and thus reduces its generalization performance. Overfitting is avoided thanks to the CV protocol; PROTEUS will strive to select models that generalize well in unseen data optimizing the out-of-sample AUC performance.

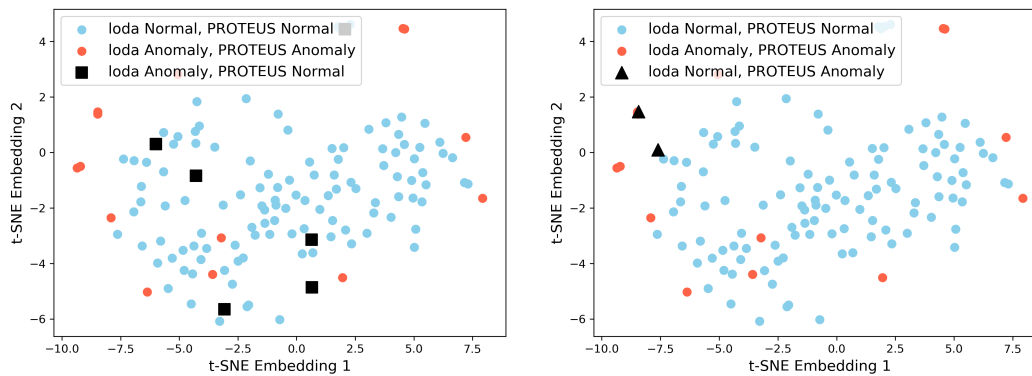
Figure 9b sheds some light on the percentage of conflicting samples between PROTEUS and unsupervised detectors per dataset. PROTEUS reveals more True Normals with an average TND  $\sim 50\%$  than True Anomalies with an average TAD  $\sim 18\%$ . In other words, *PROTEUS seems to be more effective in discovering false alarms.* To justify this claim we consider a 2D reduced visualization using t-SNE [35] of the Arrhythmia dataset projected over 10-dimensional PROTEUS explanation for LODA. Figures 10b and 10a depict the agreements between PROTEUS and LODA as circles and their disagreements as rectangles for *ANC* and triangles for *NAC*. Figure 10b illustrates the *ANC* samples contributing to the identification of *True Normals*. As expected, these samples are located within dense regions, surrounded by normal samples, requiring more complex boundaries to separate. In contrast, Figure 10a illustrates the *NAC* samples contributing to the identification of *True Anomalies*. These samples lie on sparse areas where less complex boundaries can be built to separate them. This is because less complex boundaries enable better generalizing models and thus PROTEUS can classify misdetected normals in sparse areas, not yielding many *True Anomalies*.

To conclude, PROTEUS constructs a reduced-dimensionality surrogate model that not only generalizes well to unseen samples but also provides valuable insights for identifying False Negatives and False Positives of unsupervised anomaly detectors.



(a) AUC of anomaly detectors followed by their approximation quality (AUC) from PROTEUS for real datasets. (b) Fraction of samples identified as TND and TAD according to PROTEUS explanations for different combinations of detectors and datasets.

Figure 9



(a) (b)

Figure 10 A 2-D reduced t-SNE visualisation of Arrhythmia according to PROTEUS 10-D explanation for LODA.

## 6 Related Work

In this section, we survey various categories of related work on explaining anomalies in unsupervised and supervised settings, partially inspired by [36]. We should stress that explanations of anomalies in temporal data is beyond the scope of this work [16, 5].

### 6.1 Explainable Anomaly Detectors

As unsupervised detectors assess the abnormality of multidimensional data on various feature subspaces, they can also report the subspaces that contributed the most to the anomaly score of a particular sample. A first example of such explainable anomaly detectors is LODA [44] which scores samples based on the average log density over an ensemble of one-dimensional histogram density estimators. Given that each histogram (with sparse projections) scores a randomly generated subspace, LODA explanations are essentially a list of features ranked according to their contribution to the anomalousness score of a sample.

LODI [11] and LOGP [10] seek an optimal subspace in which an anomaly is maximally separated from its neighbors. Both works exploit a dimensionality reduction technique to measure the anomalousness of a sample in a low-dimensional subspace capable of preserving the locality around its neighbors while at the same time maximizing its distance from this neighborhood. Then, the explanation of a sample is the top-k features with the largest absolute coefficient from the eigenvector with the largest eigenvalue.

In [47], an interactive explanation method is proposed that can be used for any density-based anomaly detector. [29] introduced a method to detect anomalies in axis-parallel subspaces, called SOD, that computes the anomaly score of a in a hyperplane w.r.t. to nearest neighbors in the full space. SOD hyperplanes that contribute most in the anomaly scores serve as explanations. CMI [6] and HiCS [25] rely on statistical methods to select subspaces of high-dimensional datasets, where anomalies exhibit a high deviation from normal samples. Both consider highly contrasting subspaces as explanations of all possible anomalies in a dataset.

*The previous works explain anomalies as a byproduct of an unsupervised detection method. Given that independent experimental evaluations showed that no detector outperforms all others for all possible datasets [17, 8, 13, 18], in our work we focus on learning the decision boundary of any unsupervised anomaly detector that could be used for a particular dataset. In contrast to the descriptive explanations provided by the aforementioned works, PROTEUS targets predictive explanations that could be successfully used to detect and explain anomalies also in unseen data.*

## 6.2 Post-hoc Anomaly Explainers

The primary focus of these methods is to specify a subset of features such that a sample may obtain a high anomaly score when projected onto these subspaces. Some authors have referred to this explanation task as “outlying aspects mining” [14, 43].

We first consider works providing local explanations. The seminal work [27] first introduced the problem of explaining individual outliers with “Intensional knowledge” under the form of minimal feature subspaces in which they show the greatest deviation from inliers. To find optimal subspaces, [30] formulates a constraint programming problem that maximizes differences between neighborhood densities of known outliers and inliers. [26] employs a search strategy aiming to find a subspace which maximizes differences in anomaly score distributions of all samples across subspaces while [38] measures the separability per anomaly using classification accuracy, and then apply Lasso to produce a local explaining subspace. OAMiner [14] finds the most outlying subspace where a sample is ranked highest in terms of a probability density measure and OARank [43] ranks features based on their potential contribution toward the anomalousness of a sample. *Rather than learning the decision boundary of individual anomalies, PROTEUS builds a classifier to explain simultaneously all the anomalies spotted by an unsupervised detector. Moreover, PROTEUS’s oversampling is supervised, optimizing the hyper-parameters of various feature selection and classification algorithms.*

Extending earlier work [2] on explaining individual anomalies, [3, 1] focus on explaining groups of anomalies for categorical data using contextual rule based explanations. Authors search for <context, feature> pairs, where the (single) feature can differentiate as many outliers as possible from inliers while sharing the same context. The anomalousness score of a sample in a subspace is calculated based on the frequency of the value that the outlier takes in the subspace. It tries to find subspaces E and S such that the outlier is frequent in one and much less frequent than expected in the other. To avoid searching exhaustively

all such rules, the method takes two parameters, and, to constrain the frequencies of the given sample in subspaces E and S, respectively. Similarly, [56] describes anomalies grouped in time. They construct explanatory Conjunctive Normal Form rules using features with low segmentation entropy, which quantifies how intermixed normal and anomalous samples are. They heuristically discard highly correlated features from the rules to get minimal explanations. The aforementioned works assume that anomalies are scattered and strive to explain them individually rather than to summarize the explanation of a collection of anomalies.

The following works perform explanation summarization aiming to explain a set of anomalies collectively rather than individually. LookOut [19] exploits a submodular optimization function to ensure concise summarization. xPACS [36] groups anomalies by generating sequential feature-based explanations providing a ranked list of feature-value pairs that are incrementally revealed until the human expert reaches a satisfactory level of confidence. *In contrast to the interactive explanations provided by xPACS, PROTEUS provides a global feature subspace that could potentially explain even unseen anomalies.*

### 6.3 Explaining Black-box Predictors

Several methods have been recently proposed to explain why a supervised model predicted a particular label for a particular sample [15, 28, 40, 45]. LIME [45] constructs a linear interpretable model that is locally faithful to the predictor. In this respect, it draws uniformly at random (where the number of such draws is also uniformly sampled) pseudo-samples per every sample to be explained. Note that LIME let the black-box classifier label the generated pseudo-samples. To the best of our knowledge, LIME has not been successfully used for imbalanced neighborhoods [54]. Other works [15, 28] explain the model by perturbing the features to quantify their influence on predictions. However, these works do not aim to explain multiple examples collectively, as the global explanation problem studied in our work.

Other works aim to produce explanations in the form of feature relevance scores, which indicate the relative importance of each feature to the classification decision. Such scores have been computed by comparing the difference between a classifier's prediction score and the score when a feature is assumed to be unobserved [46], or by considering the local gradient of the classifier's prediction score with respect to the features for a particular example [4]. [48, 49] considered how to score features in a way that takes into account the joint influence of feature subsets on the classification score, which usually requires approximations due to the exponential number of such subsets.

The aforementioned works require as input a supervised model rather than an unsupervised anomaly detector. However, in real application settings it is difficult or even impossible to label data as anomalous or normal examples [17]. Moreover, *PROTEUS provides global explanations returned by standard feature selection algorithms after learning the decision boundary of the unsupervised detector.*

### 6.4 Evaluation of Explainers

Existing approaches for evaluating explanation methods in both supervised and unsupervised settings are typically quite limited in their scope. Often evaluations are limited to visualizations or illustrations of several example explanations [4, 10] or to test whether a computed explanation collectively conforms to some known concept in the dataset [4], often for synthetically generated data. [47] proposes a larger scale quantitative evaluation methodology for anomaly explanations regarding sequential feature explanation methods. *Compared to this study, in our work we assess the predictive performance of a classifier given an explanation along with the correctness of the learned features of the explanation.*

## 6.5 Imbalanced Learning

One of the main challenges in supervised anomaly detection, is class imbalance: anomalies are largely underrepresented compared to normal examples. In the following we position PROTEUS w.r.t. the main imbalanced learning methods [22]. The imbalanced learning problem is concerned with the performance of learning algorithms in the presence of under-represented data and severe class distribution skews. We follow the same categorization of imbalanced learning methods as in [22].

*Random oversampling* augments the original dataset by replicating examples from the minority class, while *random undersampling* removes a random set of majority class examples. PROTEUS pipelines do not perform random under/over-sampling. The synthetic minority oversampling technique (SMOTE) [9] generates new minority class examples from the line segments that join the  $k$  minority-class nearest neighbors. Our pipeline generates synthetic examples close to the original minority examples by adding gaussian noise. SVM SMOTE [42] is a SMOTE variant that generates the synthetic examples concentrated in the most critical area, i.e., the boundary discovered by fitting an SVM classifier. Borderline-SMOTE [20] seeks to oversample the minority class instances in the borderline areas, by defining a set of “Danger” examples. Adaptive Synthetic Sampling (ADASYN) [21] algorithm uses a density distribution as a criterion to automatically decide the number of synthetic examples that need to be generated for each minority example. *In comparison to the aforementioned works, PROTEUS performs a supervised synthetic minority oversampling ensuring that new samples are anomalies according to the decision boundary of an unsupervised detector that is currently explained. In addition, we proposed a method to avoid information leakage in the CV protocol when synthetic oversampling is applied.*

## 7 Conclusion and Future Work

We propose the first methodology for producing predictive, global anomaly explanations in a detector-agnostic fashion. In particular, we show how with adequate design choices regarding rare class oversampling and unbiased performance estimation of ML pipelines, generating predictive, global anomaly explanations boils down to an AutoML problem. As derived from our experiments, PROTEUS is not only able to discover explaining subspaces of features relevant to anomalies, but it can also construct predictive models that approximate effectively and robustly the decision boundary of popular unsupervised detectors (e.g., IF, LOF, LODA). As future work, it would be interesting to approximate the decision boundary of a detector directly from the provided anomaly scores rather than converting them to binary labels. Hence, one could transform the explanation problem into regression with feature selection.

---

### References

- 1 F. Angiulli, Fabio Fassetto, L. Palopoli, and G. Manco. Outlying property detection with numerical attributes. *Data Mining and Knowledge Discovery*, 31:134–163, 2016.
- 2 Fabrizio Angiulli, Fabio Fassetto, and Luigi Palopoli. Detecting outlying properties of exceptional objects. *ACM Trans. Database Syst.*, 34(1):7:1–7:62, 2009.
- 3 Fabrizio Angiulli, Fabio Fassetto, and Luigi Palopoli. Discovering characterizations of the behavior of anomalous subpopulations. *IEEE Trans. Knowl. Data Eng.*, 25(6):1280–1292, 2013.



- 4 David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, 2010.
- 5 Aline Bessa, Juliana Freire, Tamraparni Dasu, and Divesh Srivastava. Effective discovery of meaningful outlier relationships. *ACM/IMS Trans. Data Sci.*, 2020.
- 6 Klemens Böhm, Fabian Keller, Emmanuel Müller, Hoang Vu Nguyen, and Jilles Vreeken. CMI: an information-theoretic contrast measure for enhancing subspace cluster and outlier detection. In *Proceedings of the 13th International Conference on Data Mining*, pages 198–206, 2013.
- 7 Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jorg Sander. Lof: identifying density-based local outliers. In *SIGMOD '00*, 2000.
- 8 Guilherme O. Campos, Arthur Zimek, Jörg Sander, Ricardo J. Campello, Barbora Micenkova, Erich Schubert, Ira Assent, and Michael E. Houle. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Min. Knowl. Discov.*, 30:891–927, 2016.
- 9 Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.
- 10 Xuan-Hong Dang, Ira Assent, Raymond T. Ng, Arthur Zimek, and Erich Schubert. Discriminative features for identifying and interpreting outliers. In *ICDE*, pages 88–99, 2014.
- 11 Xuan-Hong Dang, Barbora Micenkova, Ira Assent, and Raymond T. Ng. Local outlier detection with interpretation. In *ECML PKDD*, pages 304–320, 2013.
- 12 Manuel Fernández Delgado, Eva Cernadas, Senén Barro, and Dinani Gomes Amorim. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, 15(1):3133–3181, 2014. doi:10.5555/2627435.2697065.
- 13 Remi Domingues, Maurizio Filippone, Pietro Michiardi, and Jihane Zouaoui. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 74:406–421, 2018.
- 14 Lei Duan, Guanting Tang, Jian Pei, James Bailey, Akiko Campbell, and Changjie Tang. Mining outlying aspects on numeric data. *Data Mining and Knowledge Discovery*, 29:1116–1151, 2014.
- 15 Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision, ICCV*, pages 3449–3457, 2017.
- 16 Ioana Giurgiu and Anika Schumann. Additive explanations for anomalies detected from multivariate temporal data. In *CIKM*, pages 2245–2248, 2019.
- 17 Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS One*, 2016.
- 18 Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. Statistical analysis of nearest neighbor methods for anomaly detection. In *NeurIPS*, pages 10921–10931, 2019.
- 19 Nikhil Gupta, Dhivya Eswaran, Neil Shah, Leman Akoglu, and Christos Faloutsos. Beyond outlier detection: Lookout for pictorial explanation. In *ECML/PKDD*, 2018.
- 20 Hui Han, Wenyuan Wang, and Binghuan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *ICIC*, 2005.
- 21 Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks*, pages 1322–1328, 2008.
- 22 Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *TKDE*, 21:1263–1284, 2009.
- 23 Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors. *Automated Machine Learning: Methods, Systems, Challenges*. Springer, 2018.
- 24 David D. Jensen and Paul R. Cohen. Multiple comparisons in induction algorithms. *do. Learn.*, 38(3):309–338, 2000.
- 25 Fabian Keller, Emmanuel Müller, and Klemens Böhm. Hics: High contrast subspaces for density-based outlier ranking. In *ICDE*, pages 1037–1048, 2012.

- 26 Fabian Keller, Emmanuel Müller, Andreas Wixler, and Klemens Böhm. Flexible and adaptive subspace search for outlier analysis. In *CIKM*, pages 1381–1390, 2013.
- 27 Edwin M. Knorr and Raymond T. Ng. Finding intensional knowledge of distance-based outliers. In *VLDB*, pages 211–222, 1999.
- 28 Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, volume 70, pages 1885–1894, 2017.
- 29 Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Outlier detection in axis-parallel subspaces of high dimensional data. In *PAKDD*, volume 5476, pages 831–838, 2009.
- 30 Chia-Tung Kuo and Ian Davidson. A framework for outlier description using constraint programming. In *AAAI*, pages 1237–1243, 2016.
- 31 Vincenzo Lagani, Giorgos Athineou, Alessio Farcomeni, Michail Tsagris, Ioannis Tsamardinos, et al. Feature selection with the r package mxm: Discovering statistically equivalent feature subsets. *Journal of Statistical Software*, 2017.
- 32 Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *ICDM*, 2008. doi: 10.1109/ICDM.2008.17.
- 33 Scott M. Lundberg, Gabriel G. Erion, Hugh Chen, Alex J. DeGrave, Jordan M Prutkin, Bala G. Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2:56–67, 2020.
- 34 Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4765–4774, 2017. URL: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.
- 35 L. V. D. Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- 36 Meghanath Macha and Leman Akoglu. Explaining anomalies in groups with characterizing subspace rules. *Data Min. Knowl. Discov.*, 32(5):1444–1480, 2018.
- 37 Emaad A. Manzoor, Hemank Lamba, and Leman Akoglu. xstream: Outlier detection in feature-evolving data streams. In Yike Guo and Faisal Farooq, editors, *KDD*, pages 1963–1972, 2018. doi:10.1145/3219819.3220107.
- 38 Barbora Micenková, Raymond T. Ng, Xuan-Hong Dang, and Ira Assent. Explaining outliers by subspace separability. In *ICDM*, pages 518–527, 2013.
- 39 Christoph Molnar. *Interpretable Machine Learning*. independently published, 2019. URL: <https://christophm.github.io/interpretable-ml-book/>.
- 40 Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.*, 73:1–15, 2018.
- 41 Nikolaos Myrtakis, Ioannis Tsamardinos, and Vassilis Christophides. Proteus: Predictive explanation of anomalies. *ICDE*, 2021.
- 42 Hien M. Nguyen, Eric W. Cooper, and Katsuari Kamei. Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradigms*, 3:4–21, 2011.
- 43 Xuan Vinh Nguyen, Jeffrey Chan, James Bailey, Christopher Leckie, Kotagiri Ramamohanarao, and Jian Pei. Scalable outlying-inlying aspects discovery via feature ranking. In *PAKDD*, pages 422–434, 2015.
- 44 Tomás Pevný. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102:275–304, 2015.
- 45 Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *KDD*, 2016.
- 46 Marko Robnik-Sikonja and Igor Kononenko. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20:589–600, 2008.
- 47 Md Amran Siddiqui, Alan Fern, Thomas G. Dietterich, and Weng-Keen Wong. Sequential feature explanations for anomaly detection. *ACM Trans. Knowl. Discov. Data*, 13(1):1:1–1:22, 2019.

- 48 Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.*, 11:1–18, 2010.
- 49 Erik Strumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41(3):647–665, 2014.
- 50 Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 267–288, 1996.
- 51 Anh Truong, Austin Walters, Jeremy Goodsitt, Keegan E. Hines, C. Bayan Bruss, and Reza Farivar. Towards automated machine learning: Evaluation and comparison of automl approaches and tools. In *31st IEEE International Conference on Tools with Artificial Intelligence*, pages 1471–1479, 2019.
- 52 Ioannis Tsamardinos, Giorgos Borboudakis, Pavlos Katsogridakis, Polyvios Pratikakis, and Vassilis Christophides. A greedy feature selection algorithm for big data of high dimensionality. *Mach. Learn.*, 108(2):149–202, 2019.
- 53 Ioannis Tsamardinos, Elissavet Greasidou, and Giorgos Borboudakis. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach. Learn.*, 107(12):1895–1922, 2018. doi:10.1007/S10994-018-5714-4.
- 54 Adam White and Artur S. d’Avila Garcez. Measurable counterfactual local explanations for any classifier. In *European Conference on Artificial Intelligence, ECAI*, volume 325, pages 2529–2535, 2020. doi:10.3233/FAIA200387.
- 55 Jiawei Yang, Susanto Rahardja, and Pasi Fränti. Outlier detection: how to threshold outlier scores? In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing, AIIPCC 2019*, pages 37:1–37:6, 2019. doi:10.1145/3371425.3371427.
- 56 Haopeng Zhang, Yanlei Diao, and Alexandra Meliou. Exstream: Explaining anomalies in event stream monitoring. In *EDBT*, pages 156–167, 2017.