Leave-One-Out-, Bootstrap- and Cross-Conformal Anomaly Detectors

1st Oliver Hennhöfer
Intelligent Systems Research Group
University of Applied Sciences Karlsruhe
Karlsruhe, Germany
oliver.hennhoefer@h-ka.de

2nd Christine Preisach
Intelligent Systems Research Group
University of Applied Sciences Karlsruhe
Karlsruhe, Germany
christine.preisach@h-ka.de

***Abstract*—The requirement of uncertainty quantification for anomaly detection systems has become increasingly important. In this context, effectively controlling Type I error rates (α) without compromising the statistical power ($1 - \beta$) of these systems can build trust and reduce costs related to false discoveries. The field of *conformal anomaly detection* emerges as a promising approach for providing respective statistical guarantees by model calibration. However, the dependency on calibration data poses practical limitations — especially within low-data regimes.**
In this work, we formally define and evaluate *leave-one-out-*, *bootstrap-*, and *cross*-conformal methods for anomaly detection, incrementing on methods from the field of conformal prediction. Looking beyond the classical *inductive* conformal anomaly detection, we demonstrate that derived methods for calculating *resampling*-conformal p-values strike a practical compromise between statistical efficiency (*full-conformal*) and computational efficiency (*split-conformal*) as they make more efficient use of available data. We validate derived methods and quantify their improvements for a range of one-class classifiers and datasets.

***Index Terms*—Conformal Inference, Anomaly Detection, Uncertainty Quantification, False Discovery Rate**

I. Introduction

The field of *anomaly detection* comprises methods for identifying observations that either deviate from the majority of observations or do otherwise not *conform* to an expected state of *normality*. Domains of application comprise *cyber security* [1], *fraud detection* [2], *predictive maintenance* [3], [4] and *healthcare* [5] — emphasizing the relevancy of anomaly detection systems in mission-critical industry applications.

This work focuses on the unsupervised approach of *one-class classification*. This approach is particularly suitable when a representative set of anomalous observations is unavailable, as expected in most anomaly detection settings.
A major limitation that one-class classifiers share is the lack of statistical guarantees regarding their estimates. Therefore, an estimator's uncertainty is by default unquantified, undermining its reliability and trustworthiness. Furthermore, the general lack of *non-parametric* models, often subject to *a priori* assumptions, and the abundance of *parameter-laden* algorithms — both prone to misspecification and overfitting [6] — often result in subpar anomaly estimates and thresholds.

Conformal anomaly detection (CAD) [7], [8] seeks to address these problems by leveraging the non-parametric and model-agnostic framework of *conformal prediction* [9]–[11] to provide a principled way to uncertainty quantification.
CAD computes valid *p*-values from arbitrary anomaly scores as obtained from any given one-class classifier. Respective *p*-values enable statistical hypothesis testing to determine whether an observation is an *inlier* [12] while controlling the (*batch-wise*) marginal False Discovery Rate (FDR).

Problem 1. *Let $\mathcal{D}$ be a set of observations (inliers) sampled from an arbitrary distribution P. Given a new batch of observations $\mathcal{B} = \{x_1, \ldots, x_n\}$, we aim to test the null hypothesis $\mathcal{H}_{0,i}$ for each $x_i \in \mathcal{B}$ as $\mathcal{H}_{0,i} : x_i$ is drawn from P (i.e., is an inlier). The objective is to determine which observations in $\mathcal{B}$ can be considered outliers while controlling the FDR for the batch at a specified nominal level α.*

The standard conformal procedure, splits available (*non-anomalous*) training data $\mathcal{D}$ into a *proper training set* $\mathcal{D}_{\text{train}}$ and a *calibration set* $\mathcal{D}_{\text{calib}}$. After fitting a scoring function $\hat{s}$ with an algorithm $\mathcal{A}$ on $\mathcal{D}_{\text{train}}$, respective anomaly scores (*conformity scores*) $\hat{s}(\mathcal{D}_{\text{calib}})$ are calculated. The *p*-values of unseen observations are computed as the relative rank of the obtained score among the scores as calculated for the *calibration set*, cf. [13]. Resulting statistical guarantees hold when *inliers* in training and test data are *exchangeable* — a term related to but *weaker* than the assumption of IID, as it only requires invariance to permutation without independence.

Contributions. Within the given context, the contributions of this work may be summarized as follows:

- We formally define *leave-one-out-*, *bootstrap-* and *cross*-conformal methods for anomaly detection. Respective (*resampling*-)conformal methods make more efficient use of available training data than the classical *inductive* (*split-conformal*) approach while yielding larger calibration sets — impacting the range of possible *p*-values to be obtained. We discuss respective theoretical foundations, guarantees, and implications.

This work was conducted as part of the research projects *Biflex Industrie* (grant number 01MV23020A) and *AutoDiagCM* (grant number 03EE2046B), funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK).

979-8-3315-0882-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICKG63256.2024.00022

- We empirically evaluate the marginal FDR and statistical power of derived methods to the *split-conformal* procedure for *Isolation Forest* [14], *Local Outlier Factor* [15] and *Principal Component Analysis* [16] on ten benchmark datasets (see [17]) after the adjustment of obtained p-values by the *Benjamini-Hochberg* procedure [18].

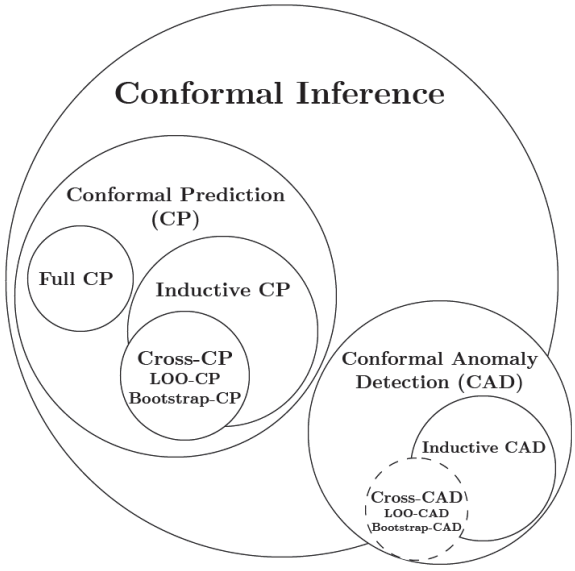

Fig. 1. Non-exhaustive taxonomy of the field of conformal inference with conformal prediction, conformal anomaly detection, and the derived family of *resampling*-conformal methods for anomaly detection.

II. Related Work

Beyond the seminal works regarding conformal inference [9]–[11], [19], the term *conformal anomaly detection* was first introduced in [7].

In [7], [8] CAD was initially applied for detecting anomalous trajectories in maritime surveillance applications. The works formalized and discussed the principles of conformal prediction applied to an anomaly detection task. As part of [8], *conformal anomaly detection*[1] and *inductive conformal anomaly detection* were defined.

The work of [12] further advanced the field of CAD by demonstrating that conformal p-values are *positive regression dependent on a subset* (PRDS) [21] and do not break FDR control via the *Benjamini-Hochberg* procedure. The respective work also proposed to explore potentially more powerful variations, beyond the *inductive* approach.

In the context of *conformal prediction*, the inductive approach was initially extended by *cross-conformal* methods as a "hybrid of the methods of inductive conformal prediction and cross-validation" [22], primarily to make more efficient use of available data, inducing a higher degree of stability into the calibration procedure. With that, this work mainly builds upon the general idea of "cross-conformal predictors" [22] and several extensions of the underlying concept — namely *Jackknife* [23], [24], *Jackknife+*, *CV*, *CV+* [22], [25] and *Jackknife+-after-Bootstrap* [26].

The concept behind *leave-one-out-* and *cross-conformal* methods for anomaly detection was first formally applied by [13] for the computation of *integrative p-values*. In this work, one-class classifiers were separately trained on both, available *inliers* and *outliers* to leverage information of both classes to integrate independently obtained p-values into a single scalar statistic. In this context, *transductive cross-validation+* (TVC+), based on CV+ [25], was proposed. Specifically, the fundamental *transductive* (*full-conformal*) approach, due to its theoretical advantages over e.g. the *inductive* or *cross-conformal* approach, was applied.

Other works relying on the application of conformal anomaly detection and related conformal concepts are [27]–[29] complemented by [30]–[32], dedicated to the online setting.

The works of [33]–[36] used (cross-)conformal *predictors* for anomaly detection using a forecasting approach.

None of these works explicitly and formally defined, referred to, or empirically evaluated *leave-one-out-*, *bootstrap-*, or *cross*-conformal anomaly detectors.

III. Background

Consider a set of data $\mathcal{D}$ comprising *n* observations $X_i \in \mathbb{R}^d$ in a *d*-dimensional feature space for $i \in [n] = \{1, 2, ...n\}$ that were sampled from an unknown continuous, discrete, or mixed distribution P_X. The goal is to answer the null hypothesis $\mathcal{H}_0$ of whether a new observation X_{n+1} was drawn from P_X under the assumption of *exchangability*, i.e. can be considered to be an *inlier*.

Definition III.1. A sequence $X_1, X_2, ...X_n$ is subject to *exchangability*, when for any finite permutation σ of the indices $1, 2, ..., n$ the joint probability distribution of a permuted sequence $X_{\sigma(1)}, X_{\sigma(2)}, ..., X_{\sigma(n)}$ is identical to the joint probability distribution of the original sequence.

We aim to compute *marginal* and *superuniform* (*conservative*) p-values $\hat{u}(X_{n+1})$ under $\mathcal{H}_0$ for all $\alpha \in (0, 1)$ with

$$\mathbb{P}_{\mathcal{H}_0}[\hat{u}(X_{n+1}) \leq \alpha] \leq \alpha. \tag{1}$$

Resulting p-values are considered to be *marginally valid* as they depend on a subset $\mathcal{D}_{\text{calib}} \subseteq \mathcal{D}$ for calibration and X_{n+1}, both considered to be random in 1. With that, *marginal* p-values are only valid *on average* yet allow for the reliable control of the marginal FDR [12].

Inductive Conformal Anomaly Detection. Given $\mathcal{D}$ containing only *inliers*, the *split-conformal* (also *inductive*) approach splits $\mathcal{D}$ into two disjoint subsets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{calib}}$.

[1] In [8] the term *conformal anomaly detection* refers to *transductive conformal prediction* [20], otherwise known as *full-conformal prediction* (compare Figure I). Full conformity is an important theoretical concept in conformal inference and the most statistically efficient approach. It does not require a dedicated calibration set but the fitting of models during inference, deeming it impractical for most real-world applications.

Data as part of the *proper training set* $\mathcal{D}_{\text{train}}$ is utilized to fit a one-class classifier to learn a function $\hat{s}(X)$ suitable to compute an *anomaly score* (or *conformity score*). In this work, small values of $\hat{s}(X_{n+1})$ are indicative of X_{n+1} being a potential *outlier*. However, given formulas may be adjusted to suffice the case of increasing scores.

Following the principles of *conformal inference* the score $\hat{s}(X_{n+1})$ of a new observation gets compared to the empirical distribution of $\hat{s}(X_i)$ as computed for the calibration data $\mathcal{D}_{\text{calib}}$, indexed by $i \in \mathcal{D}_{\text{calib}}$. With that, conformal *p*-values are computed as the *normalized rank* of $\hat{s}(X_{n+1})$ in $\hat{s}(X_i)$ [13]. For any given $X_{n+1} \in \mathbb{R}^b$ the corresponding marginal conformal *p*-value is defined as

$$\hat{u}(X_{n+1}) = \frac{|\{i \in \mathcal{D}_{\text{calib}} \ : \ \hat{s}(X_i) \leq \hat{s}(X_{n+1})\}|}{n} \tag{2}$$

Given the intention to control the FDR of obtained anomaly estimates, a conservative correction by the function

$$p(x) = \frac{nx + 1}{n + 1} \tag{3}$$

must be applied to ensure *super-uniformity*[2] of obtained *p*-values, although decreasing their statistical power (especially for smaller calibration sets [12], [13]).

Definition III.2 (**Super-Uniformity**)**.** A random value X within $[0, 1]$ is said to be *super-uniform* if [its *cumulative distribution function* (CDF) is given by] $\mathbb{P}_{\mathcal{H}_0}(X \leq t) \leq t$ for all $t \in [0, 1]$. This implies that X is super-uniformly distributed.

Under the assumption of *exchangability* the computed conformal *p*-value $\hat{u}(X_{n+1})$ is valid for testing given $\mathcal{H}_0$.

Proposition III.3 (e.g. from [12])**.** *If the inliers in* $\mathcal{D}_{calib}$ *are exchangable with themselves and with* X_{n+1}*, then* $\mathbb{P}_{\mathcal{H}_0}[\hat{u}(X_{n+1}) \leq \alpha] \leq \alpha$*. for all* $\alpha \in (0, 1)$*.*

Besides the *marginal validity* of computed *p*-values, also the *range* of possible *p*-values to be obtained depends on $\mathcal{D}_{\text{calib}}$ with the lower bound limited by $1/|\mathcal{D}_{\text{calib}}|+1$. This poses critical limitations (especially in *low-data regimes*) as *p*-values might not sufficiently reflect a model certainty, resulting in overly *conservative* yet occasionally *anti-conservative p*-values [12].

The stated limitations, motivate the formal definition and empirical evaluation of *leave-one-out-*, *bootstrap-*, and *cross-conformal* anomaly detectors that systematically yield more powerful anomaly detectors.

IV. Leave-one-out-, Bootstrap- and Cross-Conformal Anomaly Detection

Leave-one-out-, bootstrap- and cross-conformal anomaly detection extends the standard *split-conformal* approach by resampling schemes regarding model training and calibration. Without the need for a dedicated calibration set $\mathcal{D}_{\text{calib}}$, respective approaches make more efficient use of the available training data $\mathcal{D}$ that may be difficult or expensive to obtain in certain contexts. Resulting anomaly detectors are less prone to unstable estimates due to *unlucky* splits that may induce bias to the calibration procedure. With that, they mitigate implications posed by the dependence on a single subset $\mathcal{D}_{\text{cal}}$.

In principle, the *resampling-conformal* anomaly detectors divide $\mathcal{D}$ into k subsets $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_k$ (whether *disjoint* folds or *intersecting* bootstrap samples) to learn a scoring function s_k on $\mathcal{D} \setminus \mathcal{D}_k$ by any given algorithm $\mathcal{A}$ suitable for one-class classification.

Algorithm 1 formally defines the *generalized resampling-conformal anomaly detector* that can be parameterized to yield different variants of conformal anomaly detectors.

Algorithm 1 Resampling-Conformal Anomaly Detection

Input: Training data $\mathcal{D}$ (*inliers*), One-class algorithm $\mathcal{A}$, Test data X_{n+1}, Number of *disjoint* folds (or *intersecting* bootstrap samples) K, Method variant (basic/+)

1: Resample $\mathcal{D}$ into K sets, $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K$.
2: **for** $k \in 1, 2, ..., K$ **do**
3: Fit a scoring function s_k with $\mathcal{A}$ on $\mathcal{D} \setminus \mathcal{D}_k$.
4: Calculate and retain conformity scores $s_k(\mathcal{D}_k)$.
5: **if** Method+ **then**
6: Retain fitted s_k.
7: **end if**
8: **end for**
9: **if not** Method+ **then**
10: Fit a final scoring function $\hat{s}$ on all $\mathcal{D}$ for inference.
11: **end if**
12: **return** Compute the *p*-value as the normalized rank of $\hat{s}(X_{n+1})$, or an aggregation of $\{s_k(X_{n+1})\}_{k=1}^K$ (for Method+), among all $\{s_k(\mathcal{D}_k)\}_{k=1}^K$ as in 2.

Output: Conformal *p*-value $\hat{u}(X_{n+1})$

Given Proposition III.3 holds for the resampling procedure during every iteration, the theoretical guarantees of the *split-*conformal approach also hold for the *resampling-conformal* approaches, as defined in the following. Meanwhile, the *resampling-conformal* approaches yield larger sets of calibration scores and more powerful (lower) *p*-values.

Proposition IV.1 (cf. [13])**.** *If the inliers in any drawn subset* $\mathcal{D}_k \subseteq D$ *are exchangeable with themselves and with* X_{n+1}*, then* $\mathbb{P}_{\mathcal{H}_0}[\hat{u}(X_{n+1}) \leq \alpha] \leq \alpha$ *for all* $\alpha \in (0, 1)$*.*

In the following, (*i*) Jackknife$_{\text{AD}}$, and Jackknife+$_{\text{AD}}$, (*ii*) CV$_{\text{AD}}$, and CV+$_{\text{AD}}$, and (*iii*) Jackknife-after-Bootstrap$_{\text{AD}}$ and Jackknife+-after-Bootstrap$_{\text{AD}}$ are formally defined, based on their respective equivalents from the field of conformal prediction.

A. Jackknife$_{AD}$ and Jackknife+$_{AD}$

The term *Jackknife* [37]–[39] denotes a statistical procedure encompassing general resampling techniques for estimating

[2] *Super-uniformity* is required to validly apply *Benjamini-Hochberg Procedure* [18] for FDR-control, see Section V.

bias and *variance* of a (statistical) estimator [40]. In contrast, the well-known *leave-one-out validation* can be viewed as a specific implementation of *Jackknife* for model evaluation in machine learning.

Following the *Jackknife* procedure, we define n leave-one-out-sets X_{-i}. For the standard Jackknife$_{\text{AD}}$ (J$_{\text{AD}}$), described in [23]–[25] for *predictive* tasks, we fit a scoring function

$$\hat{s}_{-i} := \mathcal{A}(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n) \tag{4}$$

and compare each resulting calibration score $\hat{s}_{-i}(X_i)$ to the test score $\hat{s}(X_{n+1})$ as obtained by a function $\hat{s}$, subsequentially fitted on all available $X_i \in \mathcal{D}$ as

$$\hat{u}(X_{n+1}) = \frac{|\{i \in \mathcal{D} : \hat{s}_{-i}(X_i) \leq \hat{s}(X_{n+1})\}| + 1}{n+1}. \tag{5}$$

For the Jackknife+$_{\text{AD}}$ (J+$_{\text{AD}}$), we additionally *retain* all n fitted scoring functions $\hat{s}_{-i}$ for inference on future X_{n+1} and use an aggregation function $\varphi(\cdot)$ (e.g. MEDIAN$[\cdot]$) to form a scalar score with $\varphi(\{\hat{s}_{-1}(X_{n+1}), \hat{s}_{-2}(X_{n+1}), ..., \hat{s}_{-n}(X_{n+1})\})$ as

$$\hat{u}(X_{n+1}) = \frac{|\,\{i \in \mathcal{D} : \hat{s}_{-i}(X_i) \leq \varphi(\{\hat{s}_{-i}(X_{n+1})\})\,| + 1}{n+1}. \tag{6}$$

The J+$_{\text{AD}}$ is an extension to J$_{\text{AD}}$ primarily seeking to stabilize obtained anomaly estimates by centering them around the corresponding leave-one-out-estimates $\hat{s}_{-i}(X_{n+1})$ instead of a single $\hat{s}(X_{n+1})$. As long as the estimator is not highly sensitive to certain observations (or subsets) of $\mathcal{D}$ the results are similar [25]. Despite the small adjustment, J+$_{\text{AD}}$ possesses a stronger theoretical foundation due to the congruent model calibration and inference procedure.

Both methods become computationally prohibitive when working with larger datasets, with J+$_{\text{AD}}$ also being computationally prohibitive during inference.

B. CV$_{AD}$ and CV+$_{AD}$

CV$_{\text{AD}}$ and CV+$_{\text{AD}}$ can be seen as a generalizations of J$_{\text{AD}}$ and J+$_{\text{AD}}$ by creating K disjoint folds $S_1, S_2, ...S_K$, each of size $m = n/K$, with $\bigcup_{i=1}^{K} S_i = \mathcal{D}$ fitting K scoring functions

$$\hat{s}_{-S_K} := \hat{s}(X_i : i \in \{1, 2, ..., n\} \setminus S_K) \tag{7}$$

to calculate CV$_{\text{AD}}$ in analogy to J$_{\text{AD}}$ as

$$p(X_{n+1}) = \frac{|\{S_k \subset \mathcal{D} : \hat{s}_{-S_k}(S_k) \leq \hat{s}(X_{n+1})\}| + 1}{n+1} \tag{8}$$

and CV+$_{\text{AD}}$, respectively, in analogy to J$_{\text{AD}}$ as

$$p(X_{n+1}) = \frac{|\{S_k \subset \mathcal{D} : \hat{s}_{-S_k}(S_k) \leq \varphi(\hat{s}_{-i}(X_{n+1})\}| + 1}{n+1}. \tag{9}$$

The advantage of CV$_{\text{AD}}$ and CV+$_{\text{AD}}$ is naturally the lower computational costs, depending on the parameterization of K. Theoretically, this comes at the cost of decreased statistical power of resulting anomaly detectors, although the theoretical foundation is practically equivalent to J$_{\text{AD}}$ and J+$_{\text{AD}}$, cf. [25].

C. Jackknife- and Jackknife+-after-Bootstrap$_{AD}$

The originally conceived Jackknife+-after-Bootstrap (J+aB$_{\text{AD}}$) [26] is based on the idea of *Jackknife-after-bootstrap* [41] and may analogously be extended by the respective *non-retaining* variant. With that, JaB$_{\text{AD}}$ iteratively samples k overlapping bootstrap samples $B_1, B_2, ..., B_K$ and the complementing set of *out-of-bag* observations $-B_1, -B_2, ..., -B_K$ to fit K scoring functions

$$\hat{s}_{B_K} := \hat{s}(X_i : i \in \{1, 2, ...n)\} \setminus -B_K) \tag{10}$$

to calculate JaB$_{\text{AD}}$ as

$$\hat{u}(X_{n+1}) = \frac{|\{B_k \subset \mathcal{D} : \hat{s}_{B_k}(-B_k) \leq \hat{s}(X_{n+1})\}| + 1}{n+1} \tag{11}$$

and the function-retaining variant J+aB$_{\text{AD}}$ as

$$\hat{u}(X_{n+1}) = \frac{|\{B_k \subset \mathcal{D} : \hat{s}_{B_k}(-B_k) \leq \varphi\left(\{\hat{s}_{B_k}(X_{n+1})\}\right)\}| + 1}{k \times |-B_k| + 1}. \tag{12}$$

Both, JaB$_{\text{AD}}$ and J+aB$_{\text{AD}}$ have the advantage of theoretically yielding arbitrarily large calibration sets, depending on their parameterization. This may be a critical property for work in *low-data regimes* as the calibration set size for the other methods is limited to the size of the training data. Furthermore, when the batch size of new data during inference is large, measures for FDR control might get too conservative, limiting the statistical power of conducted tests (see Section V).

V. MULTIPLE TESTING, PRDS, AND THE BENJAMINI-HOCHBERG PROCEDURE

In order to control the FDR, the marginally valid *p*-values, as *simultaneously* obtained for any CAD method during *batch-wise* inference, need to be corrected for *multiple testing* [42]. In this work, we focus on the popular *Benjamini-Hochberg* (BH) procedure [18] that allows FDR control for a set of *super-uniformly distributed* and *independent p*-values (or test statistics) at a given nominal level.

The FDR is defined as the expected proportion of *false discoveries* (Q) as defined by the proportion of *total discoveries* (R) to *erroneous discoveries* (V), given $R > 0$ (else $\mathbb{E}(Q) = 0$) as

$$\text{FDR} = \mathbb{E}(Q) = \mathbb{E}\left[\left(\frac{\text{V}}{\text{R}}\right) \mid R > 0\right], \tag{13}$$

or put practically, in context of discussed *anomaly detection systems*, as

$$\text{FDR} = \mathbb{E}\left(\frac{\text{efforts wasted on false alarms}}{\text{total efforts}}\right), \tag{14}$$

both following the definitions as provided in [43].

The *Benjamini-Hochberg* procedure computes adjusted p-values $p^{\text{BH}}_{(i)}$ for m tested hypotheses $\{\mathcal{H}_{01}, \mathcal{H}_{02}, ..., \mathcal{H}_{0i}\} = \{\mathcal{H}_{0i}\}_{i=1}^{m}$ and corresponding p-values $\{p_1, p_2, ..., p_i\} = \{p_i\}_{i=1}^{m}$ by sorting respective p-values as $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$. Now let p_k be the largest value, so that $p_k \leq {k\alpha}/{m}$. In case no respective k exists, no actual discovery (*outlier*) is among the obtained results. Otherwise, in case a respective k exists, reject the k hypotheses $\mathcal{H}_{0i}$ that belong to $p_{(1)}, ..., p_{(k)}$, defined as

$$p^{\text{BH}}_{(i)} = \text{Minimum}\left(\left(\underset{j \geq i}{\text{Minimum}}\ {mp_{(j)}}/{j}\right), 1\right). \tag{15}$$

With that, $p^{\text{BH}}_{(i)} \leq \alpha$ only if respective $\mathcal{H}_i$ was among the discoveries prior to the adjustment.

In [12], *inductive* conformal p-values were shown to be both, *super-uniformly* distributed (see (III)) and *positive regression dependent* [*on each one from*] *a subset* (PRDS).

Definition V.1 (**PRDS**, e.g., from [21])**.** A vector of p-values $(p_1, p_2, ..., p_n)$ is said to exhibit *positive regression dependence on [each one from] a subset* $I_0 \subset \{1, 2, ..., n\}$ if, for any $i \in I_0$ and any non-decreasing set $\mathcal{D} \subset [0, 1]^n$, the conditional probability $\mathbb{P}_{\mathcal{H}_0}(p \in \mathcal{D} \,|\, p_i = p)$ is non-decreasing in p.

Although initially developed under the assumption of *independence* among obtained p-values, BH was proven to be also robust against this form of dependence among p-values.
Any set of p-values, as obtained for *resampling*-conformal methods, can likewise be considered to be PRDS as the same principle applies: " [...] larger scores in the calibration set make the p-values for all test points simultaneously smaller, and vice-versa" [12].

Theorem V.2 (cf. [12])**.** *Assume* $\hat{s}(X_i)$, $\hat{s}_{-i}(X_i)$ *for all* $X_i \in \mathcal{D}$, $\hat{s}_{-S_k}(S_k)$ *with* $\bigcup_{i=1}^{K} S_i = \mathcal{D}$ *or* $\hat{s}_{B_k}(-B_k)$ *with* $B_k \subseteq \mathcal{D}$ *to be continuously distributed. Consider* m *test points* $X_{n+1}, X_{n+2}, ...X_{n+m}$ *such that the inliers are jointly independent of each other and the data* $\mathcal{D}$*. Then the marginal (resampling-)conformal* p*-values* $(\hat{u}(X_{n+1}), \hat{u}(X_{n+2}), ..., \hat{u}(X_{n+m}))$ *are PRDS on the set of inliers.*

This proves that the FDR control for *resampling*-conformal p-values also offers *marginal guarantees* resulting in *marginal* FDR control.

Theorem V.3 (e.g., from [21])**.** *If the joint distribution of the* p*-values (or test statistics) is PRDS on the subset of* p*-values (or test statistics) corresponding to true* $\mathcal{H}_0$*, the Benjamini Hochberg procedure controls the FDR at levels* $\leq \frac{|\mathcal{H}_0|}{|\mathcal{H}|}\alpha$.

As it was shown that *resampling*-conformal p-values come with the same statistical properties as *inductive*-conformal p-values, the adjustment procedures as described in [12] — to obtain stronger (than average) guarantees in the form of *calibration-conditional* conformal p-values — can equally be applied.

VI. Evaluation

We assess derived leave-one-out, bootstrap- and cross-conformal methods in two separate experiments to assess their effectiveness in providing reliable uncertainty quantification in anomaly detection:

- **Experiment I**: Comparison between split-, leave-one-out and cross-conformal ($K = 10$) methods on ten benchmark datasets.
- **Experiment II**: Comparison between different calibration set sizes obtained for both bootstrap-conformal methods on ten benchmark datasets.

TABLE I
Key figures of the evaluation datasets as part of ADBench [17].

	Name	n	n_{feature}	n_{outlier}/n
Low Data	**Wine**	129	13	.078
	WBC	223	9	.350
	Ionosphere	351	32	.359
	Breast	683	9	.350
	Pima	768	8	.350
High Data	**Musk**	3 062	166	.032
	Annthyroid	7 200	6	.074
	Mammography	11 183	6	.023
	Shuttle	49 097	9	.072
	Fraud	284 807	29	.002

The conformal methods are applied to *Isolation Forest*, [14] *Local Outlier Factor* (LOF) [15] and *Principal Component Analysis* (PCA) [16]. The algorithms are used with their default parameters[3] as implemented by `PyOD` [44].

Ten benchmark datasets, as found in the benchmark collection *ADBench* [17], are used for evaluation. The datasets are selected to encompass diverse datasets in terms of (*i*) size, (*ii*) dimensionality, and (*iii*) class imbalance (see Table I).
Due to their computational costs, the leave-one-out-conformal

[3]For the Principal Component Analysis *n_components* was set to 3. For all algorithms *contamination* was set to the smallest possible *float* value due to their application for *one-class classification*.

TABLE II

PERFORMANCE OF SPLIT- AND RESAMPLING-CONFORMAL ANOMALY DETECTION METHODS USING ISOLATION FOREST. THE EVALUATION IS BASED ON THE (MARGINAL) FALSE DISCOVERY RATE ($\alpha = 0.2$) AND THE (MEAN) STATISTICAL POWER ($\bar{x}$), EXTENDING BEYOND THE MARGINAL CASE AT THE 90TH QUANTILE ($P90$) OF THE EMPIRICAL DISTRIBUTION AND THE RESPECTIVE STANDARD DEVIATION (σ).

	Isolation Forest – False Discovery Rate														
	Split-Conformal$_{AD}$			10-CV$_{AD}$			10-CV+$_{AD}$			Jackknife$_{AD}$			Jackknife+$_{AD}$		
	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ
Wine	**.032**	.107	.088	**.063**	.218	.133	**.037**	.149	.091	**.047**	.136	.114	**.040**	.141	.081
WBC	**.119**	.355	.153	**.135**	.329	.128	**.138**	.335	.129	**.141**	.342	.132	**.146**	.348	.133
Ionosphere	**.044**	.138	.107	**.065**	.207	.117	**.062**	.204	.098	**.073**	.233	.111	**.064**	.193	.090
Breast	**.178**	.328	.112	**.180**	.304	.096	**.102**	.228	.096	**.184**	.302	.090	**.184**	.304	.092
Pima	**.058**	.172	.146	**.059**	.215	.110	**.047**	.112	.106	**.078**	.281	.124	**.046**	.155	.093
Musk	**.102**	.228	.096	**.099**	.324	.141	**.008**	.011	.025	—	—	—	—	—	—
Annthyroid	**.161**	.262	.091	**.159**	.250	.081	**.161**	.232	.060	—	—	—	—	—	—
Mammography	**.164**	.261	.077	**.153**	.296	.099	**.126**	.208	.053	—	—	—	—	—	—
Shuttle	**.175**	.209	.025	**.182**	.251	.059	**.182**	.251	.059	—	—	—	—	—	—
Fraud	**.178**	.218	.030	**.178**	.221	.039	**.173**	.182	.007	—	—	—	—	—	—
	Isolation Forest – Statistical Power														
	Split-Conformal$_{AD}$			10-CV$_{AD}$			10-CV+$_{AD}$			Jackknife$_{AD}$			Jackknife+$_{AD}$		
	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ
Wine	**.007**	.025	.019	**.018**	.071	.041	**.012**	.039	.034	**.011**	.035	.029	**.010**	.035	.025
WBC	**.174**	.476	.186	**.645**	.901	.239	**.616**	.877	.235	**.586**	.882	.256	**.614**	.887	.239
Ionosphere	**.027**	.112	.050	**.111**	.265	.100	**.110**	.244	.096	**.127**	.289	.108	**.119**	.232	.095
Breast	**.700**	.850	.165	**.797**	.909	.085	**.798**	.903	.082	**.800**	.899	.080	**.802**	.913	.083
Pima	**.022**	.072	.046	**.031**	.102	.042	**.026**	.068	.038	**.040**	.116	.052	**.025**	.064	.029
Musk	**.406**	.841	.370	**.253**	.828	.343	**.062**	.099	.196	—	—	—	—	—	—
Annthyroid	**.570**	.829	.280	**.584**	.819	.270	**.625**	.799	.183	—	—	—	—	—	—
Mammography	**.732**	.850	.127	**.787**	.912	.110	**.842**	.891	.047	—	—	—	—	—	—
Shuttle	**.825**	.856	.025	**.818**	.892	.059	**.835**	.853	.013	—	—	—	—	—	—
Fraud	**.821**	.859	.030	**.822**	.870	.039	**.827**	.834	.007	—	—	—	—	—	—

methods are only evaluated on datasets with less than 1000 observations (designated as *low-data regime*).

A. Setup

Following the evaluation setup as described in [12], we randomly create J distinct datasets $\mathcal{D}_1, ..., \mathcal{D}_j$ with $j \in J$, comprising only *normal* observations. Each dataset $\mathcal{D}_j$ represents an independent data set used for training and calibration. Each $\mathcal{D}_j$ comes with L test sets $\mathcal{D}^{\text{test}}_{j,1}, ..., \mathcal{D}^{\text{test}}_{j,l}$ with $l \in L$. While $\mathcal{D}_j$ is fixed regarding its L test sets, they are again drawn randomly and are not strictly disjointed.

For the evaluation, we are interested in the FDR conditional on $\mathcal{D}_j$ defined as the expectation value

$$\text{cFDR}(\mathcal{D}_j) := \mathbb{E}\left[\text{FDP}\left(\mathcal{D}^{\text{test}}; \mathcal{D}_j\right) \mid \mathcal{D}_j\right], \tag{16}$$

with $\text{FDP}\left(\mathcal{D}^{\text{test}}; \mathcal{D}_j\right)$ as the *False Discovery Proportion* (FDP) of inliers in the test set that was incorrectly reported as outliers.

The results for any given $j \in J$ will be evaluated by the FDR

$$\widehat{\text{cFDR}(\mathcal{D}_j)} := \frac{1}{L}\sum_{l=1}^{L} \text{FDP}\left(\mathcal{D}^{\text{test}}_{j,l}; \mathcal{D}_j\right) \tag{17}$$

and the statistical power

$$\widehat{\text{cPower}(\mathcal{D}_j)} := \frac{1}{L}\sum_{l=1}^{L} \text{Power}\left(\mathcal{D}^{\text{test}}_{j,l}; \mathcal{D}_j\right) \tag{18}$$

where $\text{Power}(\mathcal{D}^{\text{test}}_{j,l}; \mathcal{D}_j)$ is defined as the proportion of total outliers in $\mathcal{D}^{\text{test}}_{j,l}$ correctly identified as outliers.

Our experiments demonstrate that the *marginal* FDR as $\widehat{\text{mFDR}(\mathcal{D}_j)} = \frac{1}{J}\sum_{j=1}^{J} \widehat{\text{cFDR}(\mathcal{D}_j)}$ is controlled.

Respective evaluation metrics cover instances either incorrectly labeled as anomalies (a *false alarm*) or anomalies that go unrecognized (a *missed discovery*).

B. Implementation Details

For training and calibration during *Experiment I*, $n_{\text{Inlier}}/2$ observations were used with n_{train} and $n_{\text{cal}} = \min\{2000, {}^{n_{\text{train}}}/_{2}\}$. For training and calibration $J = 100$ subsets were drawn, each with $L = 100$ corresponding non-disjoint test sets $\mathcal{D}^{\text{test}}_{j,l}$ of size $n_{\text{test}} = \min\{1000, {}^{n_{\text{train}}}/_{3}\}$ were sampled, each with 90% inliers and 10% outliers. The FDR was controlled at the nominal level $\alpha = 0.2$ by the Benjamini-Hochberg procedure.

For *Experiment II*, JaB$_{AD}$ and J+aB$_{AD}$ were trained with a (re-)sampling ratio of 0.95 and increasing subsampling iterations to obtain different calibration set sizes $\{100, 200, \ldots, 1000\}$. The experiment protocol of *Experiment I* was followed for every set size.

TABLE III

PERFORMANCE OF SPLIT- AND RESAMPLING-CONFORMAL ANOMALY DETECTION METHODS USING LOCAL OUTLIER FACTOR. THE EVALUATION IS BASED ON THE (MARGINAL) FALSE DISCOVERY RATE ($\alpha = 0.2$) AND THE (MEAN) STATISTICAL POWER ($\bar{x}$), EXTENDING BEYOND THE MARGINAL CASE AT THE 90TH QUANTILE ($P90$) OF THE EMPIRICAL DISTRIBUTION AND THE RESPECTIVE STANDARD DEVIATION (σ).

	Local Outlier Factor – False Discovery Rate														
	Split-Conformal$_{\text{AD}}$			10-CV$_{\text{AD}}$			10-CV+$_{\text{AD}}$			Jackknife$_{\text{AD}}$			Jackknife+$_{\text{AD}}$		
	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ
Wine	**.044**	.144	.123	**.165**	.444	.177	**.150**	.415	.170	**.153**	.415	.164	**.153**	.415	.164
WBC	**.130**	.316	.126	**.126**	.260	.117	**.129**	.263	.117	**.139**	.324	.122	**.138**	.324	.121
Ionosphere	**.077**	.238	.127	**.152**	.350	.144	**.122**	.310	.137	**.136**	.328	.140	**.136**	.328	.140
Breast	**.159**	.345	.127	**.092**	.276	.108	**.100**	.268	.108	**.101**	.268	.110	**.102**	.262	.110
Pima	**.032**	.060	.104	**.061**	.205	.138	**.065**	.204	.143	**.067**	.205	.146	**.067**	.205	.146
Musk	**.176**	.235	.046	**.175**	.222	.039	**.157**	.206	.038	—	—	—	—	—	—
Annthyroid	**.168**	.260	.082	**.161**	.240	.064	**.162**	.242	.065	—	—	—	—	—	—
Mammography	**.142**	.256	.088	**.160**	.230	.074	**.165**	.234	.075	—	—	—	—	—	—
Shuttle	**.181**	.210	.023	**.187**	.206	.013	**.178**	.196	.012	—	—	—	—	—	—
Fraud*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	Local Outlier Factor – Statistical Power														
	Split-Conformal$_{\text{AD}}$			10-CV$_{\text{AD}}$			10-CV+$_{\text{AD}}$			Jackknife$_{\text{AD}}$			Jackknife+$_{\text{AD}}$		
	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ
Wine	**.012**	.045	.033	**.113**	.263	.113	**.107**	.264	.113	**.113**	.263	.113	**.113**	.263	.113
WBC	**.222**	.492	.187	**.554**	.829	.231	**.628**	.877	.204	**.540**	.813	.231	**.541**	.821	.235
Ionosphere	**.075**	.210	.050	**.426**	.589	.119	**.398**	.535	.114	**.419**	.589	.122	**.419**	.589	.122
Breast	**.519**	.790	.265	**.168**	.553	.216	**.211**	.614	.235	**.197**	.600	.227	**.198**	.600	.227
Pima	**.010**	.030	.024	**.016**	.046	.034	**.018**	.050	.036	**.018**	.057	.036	**.018**	.057	.036
Musk	**.824**	.882	.046	**.825**	.880	.039	**.843**	.889	.038	—	—	—	—	—	—
Annthyroid	**.732**	.819	.082	**.807**	.865	.048	**.807**	.867	.048	—	—	—	—	—	—
Mammography	**.476**	.772	.287	**.509**	.770	.247	**.514**	.772	.245	—	—	—	—	—	—
Shuttle	**.819**	.846	.023	**.813**	.829	.013	**.822**	.837	.012	—	—	—	—	—	—
Fraud*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

**Fraud* was excluded due to the dataset's dimensionality (see Table I) and the $O(N^2d)$ time complexity of *Local Outlier Factor*.

VII. RESULTS AND DISCUSSION

The results in Table II, Table III and Table IV extend the findings in [12], as also resampling-conformal p-values allow for reliable control of the FDR for batch-wise inference in a marginal sense ($\bar{x} \leq 0.2$). The FDR for the split-conformal method tends to be overall lower but has to be contextualized by the observed statistical power as lower FDR and higher statistical power typically represent a trade-off.

With regard to statistical power, the resampling-conformal methods outperform the split-conformal approach with few exceptions. This can mainly be attributed to the larger calibration sets that the resampling-based methods yield, allowing for lower p-values that maintain significance after the multiple testing correction. As the size of the calibration set has a decreasing marginal benefit, the advantage of resampling-conformal methods becomes smaller in high-data regimes (for calibration sets ≥ 2000). The converging performance is also linked to the decreasing training set size for CV[+]$_{\text{AD}}$ with $K = 10$, as each ensemble model is trained on less data than the split-conformal model that caps its calibration set to 2000 instances for large datasets (see Subsection VI-B).

Looking at the bigger picture, several factors may be decisive for the performance of the conformal methods. First, the usefulness of obtained *non-conformity scores* (and resulting p-values) is primarily determined by the learned scoring function [45]. An unsuited algorithm $\mathcal{A}$ learning a deficient scoring function $\hat{s}$ will fail to produce powerful p-values. In this context, the ensemble approach of resampling-conformal methods might be counterproductive, as they potentially amplify the systematic bias of models that underfit a given dataset (i.e. dilute the little learning done). Second, the adherence or violation of the exchangeability assumption (for out-of-sample training data *and* for test data) may have an impact. For (perfectly) exchangeable data, Jackknife[+]$_{\text{AD}}$ may perform better as each ensemble model trains on more data, resulting in lower bias and higher variance. For more heterogeneous datasets, CV[+]$_{\text{AD}}$ may be more robust towards local patterns — although this would need to be confirmed empirically with respective experiment setups. Lastly, the impact of retaining trained classifiers (for the variants Jackknife+$_{\text{AD}}$ and CV+$_{\text{AD}}$) seems to be inconsistent throughout the results, although the differences for CV[+]$_{\text{AD}}$ are larger. The performance differences may be linked to the previously mentioned arguments regarding exchangeability, although the discussion is nuanced and would need further evaluation.

The key findings of Experiment I may be summarized as:

- Resampling-CAD offers reliable marginal FDR control and yields more powerful anomaly classifiers, especially within low-data regimes.
- The performance of CAD greatly depends on the effectiveness of the learned scoring function to separate between inliers and outliers to produce powerful p-values.

TABLE IV

PERFORMANCE OF SPLIT- AND RESAMPLING-CONFORMAL ANOMALY DETECTION METHODS USING PRINCIPAL COMPONENT ANALYSIS. THE EVALUATION IS BASED ON THE (MARGINAL) FALSE DISCOVERY RATE ($\alpha = 0.2$) AND THE (MEAN) STATISTICAL POWER ($\bar{x}$), EXTENDING BEYOND THE MARGINAL CASE AT THE 90TH QUANTILE ($P90$) OF THE EMPIRICAL DISTRIBUTION AND THE RESPECTIVE STANDARD DEVIATION (σ).

	Principal Component Analysis – False Discovery Rate														
	Split-Conformal$_{AD}$			10-CV$_{AD}$			10-CV+$_{AD}$			Jackknife$_{AD}$			Jackknife+$_{AD}$		
	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ
Wine	**.035**	.085	.103	**.048**	.147	.110	**.042**	.114	.098	**.045**	.111	.110	**.044**	.111	.108
WBC	**.101**	.267	.120	**.107**	.276	.115	**.095**	.270	.112	**.111**	.284	.119	**.109**	.282	.117
Ionosphere	**.069**	.264	.124	**.096**	.282	.115	**.096**	.258	.116	**.097**	.291	.121	**.087**	.288	.119
Breast	**.170**	.337	.123	**.148**	.316	.111	**.139**	.316	.114	**.150**	.316	.111	**.149**	.316	.111
Pima	**.043**	.148	.103	**.059**	.184	.125	**.055**	.159	.113	**.060**	.237	.120	**.058**	.201	.117
Musk	**.182**	.249	.050	**.181**	.234	.044	**.176**	.229	.044	—	—	—	—	—	—
Annthyroid	**.171**	.268	.071	**.165**	.236	.055	**.162**	.234	.056	—	—	—	—	—	—
Mammography	**.166**	.245	.066	**.174**	.231	.042	**.172**	.231	.042	—	—	—	—	—	—
Shuttle	**.174**	.207	.026	**.178**	.194	.015	**.172**	.192	.015	—	—	—	—	—	—
Fraud	**.180**	.220	.028	**.181**	.189	.006	**.180**	.188	.006	—	—	—	—	—	—
	Principal Component Analysis – Statistical Power														
	Split-Conformal$_{AD}$			10-CV$_{AD}$			10-CV+$_{AD}$			Jackknife$_{AD}$			Jackknife+$_{AD}$		
	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ	$\bar{x}$	P_{90}	σ
Wine	**.009**	.022	.024	**.018**	.053	.038	**.017**	.051	.036	**.017**	.052	.035	**.017**	.02	.035
WBC	**.160**	.422	.168	**.377**	.735	.277	**.358**	.725	.279	**.381**	.730	.256	**.377**	.727	.254
Ionosphere	**.068**	.229	.094	**.285**	.403	.080	**.282**	.406	.082	**.293**	.411	.074	**.286**	.396	.074
Breast	**.602**	.811	.248	**.747**	.898	.134	**.742**	.902	.139	**.781**	.896	.087	**.780**	.905	.087
Pima	**.012**	.043	.027	**.027**	.075	.042	**.025**	.072	.040	**.028**	.098	.045	**.027**	.087	.045
Musk	**.818**	.879	.050	**.819**	.880	.044	**.824**	.881	.044	—	—	—	—	—	—
Annthyroid	**.807**	.876	.056	**.825**	.879	.049	**.825**	.880	.047	—	—	—	—	—	—
Mammography	**.686**	.829	.195	**.810**	.850	.033	**.812**	.852	.033	—	—	—	—	—	—
Shuttle	**.826**	.855	.026	**.822**	.842	.015	**.828**	.848	.015	—	—	—	—	—	—
Fraud	**.820**	.855	.028	**.819**	.827	.006	**.820**	.826	.006	—	—	—	—	—	—

- Calibration set sizes have a decreasing marginal benefit, as the advantages of resampling- over split-conformal methods become smaller within *high-data regimes*.
- Impacts of *partial exchangeability* (only within certain subsets) and its implications for different method parameterizations are nuanced and need further investigation.

Experiment II (see Table V) tried to quantify the influence of the calibration set size as obtained by JaB$_{AD}$ and J+aB$_{AD}$, applied with *Principal Component Analysis* (PCA).
Overall the results confirm the general findings of Experiment I while yielding a noteworthy exception for the *Pima* dataset. As PCA clearly shows poor performance on the data, the marginal FDR is only maintained until a calibration set size of $n_{\text{calib}} = 500$. From this, we can conclude that the *powerless* p-values, produced by the subpar PCA model, get so low that they remain significant after the multiple testing correction — inflating the FDR beyond the marginal case. This demonstrates the *adverse effects* of scoring functions that fail to effectively separate between inliers and outliers, resulting in p-values that do not accurately reflect an instance's degree of outlyingness.

VIII. CONCLUSION

Resampling-conformal methods represent a natural and effective addition to conformal anomaly detection. Derived methods are particularly helpful for anomaly detection tasks in low-data regimes that require uncertainty quantification. The resampling-based methods outperform the split-conformal approach, although the impact of their parameterization and the characteristics of the underlying data (*partial exchangability*) is nuanced and suggests new directions for future research.

By framing (*batch-wise*) inference for anomaly detection as a multiple testing problem, the *marginal* FDR of resampling-conformal anomaly detectors can reliably be controlled by the Benjamini-Hochberg procedure, while typically exhibiting higher statistical power and estimator stability. Due to the inherent *model agnosticism* of conformal methods, they may easily be integrated into existing anomaly detection systems, offering high practicability. Constraints are mainly the increased need for computational capacities, at least at the model training stage, and the *exchangability* assumption. Furthermore, since FDR control is inherently tied to a multiple-testing perspective, the general approach does not directly apply to an online anomaly detection setting.
Beyond that, conformal anomaly detection methods integrate elegantly with anomaly detection algorithms like *Isolation Forest* that require a *threshold value* for detection.

Overall, the results provided in this work confirm the effectiveness of the overarching principles of conformal inference in a wider range of applications, beyond conformal prediction. In summary, the presented work formally defined the field of *leave-one-out*-, *bootstrap*- and *cross*-conformal (as *resampling*-conformal) anomaly detection in analogy to the existing field of *conformal prediction* and respective methods.

TABLE V

THE (MARGINAL) FALSE DISCOVERY RATE ($\alpha = 0.2$) AND THE (MEAN) STATISTICAL POWER ($\bar{x}$) FOR JACKKNIFE[+]-AFTER-BOOTSTRAP WITH DIFFERENT CALIBRATION SET SIZES (n_{CALIB}) USING PCA (RESAMPLING RATIO = 0.95). EACH RUN FOLLOWS THE PROTOCOL OF EXPERIMENT I.

		(Marginal) False Discovery Rate									
		$n_{\text{calib}} = 100$	200	300	400	500	600	700	800	900	1,000
Wine	JaB	.185	.179	.173	.179	.179	.183	.185	.187	.186	.184
Wine	J+aB	.180	.177	.168	.172	.170	.173	.173	.178	.176	.175
WBC	JaB	.145	.161	.162	.167	.170	.172	.178	.176	.177	.181
WBC	J+aB	.139	.152	.153	.157	.161	.165	.170	.162	.168	.171
Ionosphere	JaB	.148	.167	.163	.172	.176	.179	.178	.177	.178	.179
Ionosphere	J+aB	.134	.156	.157	.167	.166	.170	.168	.167	.170	.172
Breast	JaB	.165	.175	.164	.173	.179	.176	.177	.173	.175	.176
Breast	J+aB	.161	.170	.159	.167	.174	.170	.171	.166	.169	.169
Pima	JaB	.021	.043	.088	.077	.239	.228	.230	.225	.238	.233
Pima	J+aB	.017	.041	.083	.073	.235	.225	.226	.221	.235	.230
Musk	JaB	.139	.161	.173	.175	.177	.177	.179	.178	.181	.183
Musk	J+aB	.137	.159	.171	.172	.174	.175	.176	.176	.178	.180
Annthyroid	JaB	.044	.081	.101	.120	.145	.143	.146	.146	.146	.154
Annthyroid	J+aB	.042	.080	.101	.119	.143	.141	.143	.144	.144	.152
Mammography	JaB	.122	.148	.153	.162	.174	.168	.168	.176	.170	.173
Mammography	J+aB	.121	.147	.151	.160	.173	.167	.167	.175	.169	.172
Shuttle	JaB	.157	.165	.169	.167	.172	.175	.175	.176	.177	.178
Shuttle	J+aB	.156	.163	.167	.165	.170	.172	.173	.174	.175	.176
Fraud	JaB	.117	.159	.168	.167	.173	.176	.178	.180	.180	.182
Fraud	J+aB	.116	.159	.168	.167	.173	.176	.177	.180	.180	.182
		(Mean) Statistical Power									
		$n_{\text{calib}} = 100$	200	300	400	500	600	700	800	900	1,000
Wine	JaB	.092	.089	.082	.083	.085	.084	.079	.084	.080	.081
Wine	J+aB	.088	.084	.078	.077	.080	.079	.075	.078	.075	.075
WBC	JaB	.471	.680	.683	.694	.713	.715	.744	.744	.743	.746
WBC	J+aB	.461	.679	.674	.684	.702	.707	.738	.735	.736	.744
Ionosphere	JaB	.338	.608	.604	.616	.607	.617	.612	.618	.613	.616
Ionosphere	J+aB	.330	.610	.602	.615	.608	.618	.614	.619	.613	.615
Breast	JaB	.583	.756	.780	.800	.802	.806	.799	.809	.806	.804
Breast	J+aB	.581	.756	.780	.803	.805	.809	.803	.814	.810	.808
Pima	JaB	.006	.018	.042	.037	.071	.068	.070	.066	.076	.073
Pima	J+aB	.005	.017	.041	.037	.070	.067	.068	.064	.075	.071
Musk	JaB	.861	.839	.827	.825	.823	.823	.821	.822	.819	.817
Musk	J+aB	.863	.841	.829	.828	.826	.825	.824	.824	.822	.820
Annthyroid	JaB	.059	.213	.345	.453	.594	.602	.653	.680	.704	.759
Annthyroid	J+aB	.056	.211	.343	.453	.593	.602	.651	.677	.702	.758
Mammography	JaB	.259	.432	.505	.565	.611	.607	.616	.647	.659	.670
Mammography	J+aB	.258	.433	.505	.564	.612	.607	.615	.647	.660	.671
Shuttle	JaB	.820	.825	.831	.833	.828	.825	.825	.824	.823	.822
Shuttle	J+aB	.818	.827	.833	.835	.830	.828	.827	.826	.825	.824
Fraud	JaB	.525	.794	.830	.824	.824	.824	.822	.820	.820	.818
Fraud	J+aB	.524	.791	.830	.825	.824	.824	.823	.820	.820	.818

SOFTWARE AND DATA

Conducted experiments are accessible at `github.com/OliverHennhoefer/resampling-conformal-cad` for exact reproduction (`Python`). Applied conformal methods are implemented in our publicly available package `unquad` (`Python`) for uncertainty-quantified anomaly detection.

REFERENCES

[1] M. Evangelou and N. M. Adams, "An anomaly detection framework for cyber-security data," *Computers & Security*, vol. 97, p. 101941, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404820302170

[2] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: a review of anomaly detection techniques and recent advances," *Expert Systems with Applications*, vol. 193, p. 116429, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421017164

[3] J. Carrasco, D. López, I. Aguilera-Martos, D. García-Gil, I. Markova, M. García-Barzana, M. Arias-Rodil, J. Luengo, and F. Herrera, "Anomaly detection in predictive maintenance: a new evaluation framework for temporal unsupervised anomaly detection algorithms," *Neurocomputing*, vol. 462, pp. 440–452, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2021.07.095

[4] H. Choi, D. Kim, J. Kim, J. Kim, and P. Kang, "Explainable anomaly detection framework for predictive maintenance in manufacturing systems," *Applied Soft Computing*, vol. 125, pp. 109 – 147, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494622004069

[5] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection – a survey," *ACM Comput. Surv.*, vol. 54, no. 7, jul 2021. [Online]. Available: https://doi.org/10.1145/3464423

[6] S. Keogh, E. A. Lonardi, C. Ratanamahatana, L. Wei, S.-H. Lee, and J. Handley, "Compression-based data mining of sequential data," *Data Mining and Knowledge Discovery*, vol. 14, no. 1, pp. 99–129, 2007, pp. 1–3.

[7] R. Laxhammar and G. Falkman, "Conformal prediction for ddistribution-independent anomaly detection in streaming vessel data," in *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, ser. StreamKDD '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 47–55. [Online]. Available: https://doi.org/10.1145/1833280.1833287

[8] R. Laxhammar, "Conformal anomaly detection: detecting abnormal trajectories in surveillance applications," Ph.D. dissertation, University of Skövde, Sweden, 2014, pp. 45 – 58. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-8762

[9] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, *Inductive confidence machines for regression*. Springer Berlin Heidelberg, 2002, p. 345–356. [Online]. Available: http://dx.doi.org/10.1007/3-540-36755-1_29

[10] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Berlin, Heidelberg: Springer-Verlag, 2005.

[11] J. Lei and L. Wasserman, "Distribution-free prediction bands for non-parametric regression," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 76, no. 1, pp. 71–96, 07 2013. [Online]. Available: https://doi.org/10.1111/rssb.12021

[12] S. Bates, E. Candès, L. Lei, Y. Romano, and M. Sesia, "Testing for outliers with conformal p-values," *The Annals of Statistics*, vol. 51, no. 1, pp. 149 – 178, 2023. [Online]. Available: https://doi.org/10.1214/22-AOS2244

[13] Z. Liang, M. Sesia, and W. Sun, "Integrative conformal p-values for out-of-distribution testing with labelled outliers," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, p. qkad138, 01 2024. [Online]. Available: https://doi.org/10.1093/jrsssb/qkad138

[14] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, Dec. 2008. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2008.17

[15] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *ACM SIGMOD Record*, vol. 29, no. 2, p. 93–104, May 2000. [Online]. Available: http://dx.doi.org/10.1145/335191.335388

[16] K. P. F.R.S., "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[17] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, "Adbench: anomaly detection benchmark," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 142–32 159, 2022.

[18] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: http://dx.doi.org/10.2307/2346101

[19] A. Gammerman, V. Vovk, and V. Vapnik, "Learning by transduction," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, p. 148–155.

[20] V. Vovk, "Transductive conformal predictors," in *9th Artificial Intelligence Applications and Innovations (AIAI)*, Paphos, Greece, Sep. 2013, pp. 348–360. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01459630

[21] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, Aug. 2001, p. 1168. [Online]. Available: http://dx.doi.org/10.1214/aos/1013699998

[22] V. Vovk, "Cross-conformal predictors," *Annals of Mathematics and Artificial Intelligence*, vol. 74, no. 1–2, p. 9–28, Jul. 2013, p. 1, Appendix C. [Online]. Available: http://dx.doi.org/10.1007/s10472-013-9368-4

[23] L. Steinberger and H. Leeb, "Leave-one-out prediction intervals in linear regression models with many variables," *arXiv: Statistics Theory*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:88514378

[24] ——, "Conditional predictive inference for stable algorithms," *The Annals of Statistics*, vol. 51, no. 1, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1214/22-AOS2250

[25] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *The Annals of Statistics*, vol. 49, no. 1, Feb. 2021. [Online]. Available: http://dx.doi.org/10.1214/20-AOS1965

[26] B. Kim, C. Xu, and R. F. Barber, "Predictive inference is free with the jackknife+-after-bootstrap," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020, p. 3.

[27] J. Smith, I. Nouretdinov, R. Craddock, C. R. Offer, and A. Gammerman, "Conformal anomaly detection of trajectories with a multi-class hierarchy," in *Statistical Learning and Data Sciences*, A. Gammerman, V. Vovk, and H. Papadopoulos, Eds. Cham: Springer International Publishing, 2015, pp. 281–290.

[28] L. Guan and R. Tibshirani, "Prediction and outlier detection in classification problems," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, no. 2, pp. 524–546, Feb. 2022. [Online]. Available: http://dx.doi.org/10.1111/rssb.12443

[29] M. Haroush, T. Frostig, R. Heller, and D. Soudry, "A statistical framework for efficient out-of-distribution detection in deep neural networks," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=Oy9WeuZD51

[30] F. Cai and X. Koutsoukos, "Real-time out-of-distribution detection in learning-enabled cyber-physical systems," in *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, 2020, pp. 174–183.

[31] V. Vovk, "Testing randomness online," *Statistical Science*, vol. 36, no. 4, pp. 595–611, 2021. [Online]. Available: https://doi.org/10.1214/20-STS817

[32] V. Vovk, I. Petej, I. Nouretdinov, E. Ahlberg, L. Carlsson, and A. Gammerman, "Retrain or not retrain: conformal test martingales for change-point detection," in *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, ser. Proceedings of Machine Learning Research, L. Carlsson, Z. Luo, G. Cherubin, and K. An Nguyen, Eds., vol. 152. PMLR, 08–10 Sep 2021, pp. 191–210. [Online]. Available: https://proceedings.mlr.press/v152/vovk21b.html

[33] J. Smith, "The efficiency of conformal predictors for anomaly detection," Ph.D. dissertation, Royal Holloway, University of London, 2016.

[34] V. Ishimtsev, A. Bernstein, E. Burnaev, and I. Nazarov, "Conformal k-NN anomaly detector for univariate data streams," in *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, ser. Proceedings of Machine Learning Research, A. Gammerman, V. Vovk, Z. Luo, and H. Papadopoulos, Eds., vol. 60. PMLR, 13–16 Jun 2017, pp. 213–227. [Online]. Available: https://proceedings.mlr.press/v60/ishimtsev17a.html

[35] C. Xu and Y. Xie, "Conformal prediction interval for dynamic time-series," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 11 559–11 569. [Online]. Available: https://proceedings.mlr.press/v139/xu21h.html

[36] ——, "Conformal anomaly detection on spatio-temporal observations with missing data," 2021.

[37] M. H. Quenouille, "Approximate tests of correlation in time-series," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 45, no. 3, p. 483–484, 1949. [Online]. Available: http://dx.doi.org/10.1017/s0305004100025123

[38] ——, "Notes on bias in estimation," *Biometrika*, vol. 43, no. 3/4, p. 353, 1956. [Online]. Available: http://dx.doi.org/10.2307/2332914

[39] J. Tukey, "Bias and confidence in not quite large samples," *Annals of Mathematical Statistics*, vol. 29, p. 614, 1958.

[40] J. Shao and D. Tu, *The Jackknife and Bootstrap*. Springer New York, 1995, p. 414. [Online]. Available: http://dx.doi.org/10.1007/978-1-4612-0795-5

[41] B. Efron, "Jackknife-after-bootstrap standard errors and influence functions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 54, no. 1, pp. 83–127, 1992. [Online]. Available: http://www.jstor.org/stable/2345949

[42] J. W. Tukey, "The problem of multiple comparisons," 1953, unpublished manuscript. See Braun (1994), pp. 1-300.

[43] Y. Benjamini, R. Heller, and D. Yekutieli, "Selective inference in complex research," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4255–4271, Nov. 2009, pp. 4257–4259. [Online]. Available: http://dx.doi.org/10.1098/rsta.2009.0127

[44] Y. Zhao, Z. Nasrullah, and Z. Li, "Pyod: A python toolbox for scalable outlier detection," *Journal of Machine Learning Research*, vol. 20, no. 96, pp. 1–7, 2019. [Online]. Available: http://jmlr.org/papers/v20/19-011.html

[45] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *CoRR*, vol. abs/2107.07511, 2021, pp. 14–15. [Online]. Available: https://arxiv.org/abs/2107.07511