Proceedings of Machine Learning Research 266:1–19, 2025 Conformal and Probabilistic Prediction with Applications

# On the Integration of Cross-Conformal Prediction, Ensembles, and Sampling for Uncertainty Quantification in One-Class Anomaly Detection

**Ishan Garg**[*] IGARG@ZENON.AI
**Shayan Majumder**[†] SM3054@HW.AC.UK

**Editor:** Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

## Abstract

Given the increasing usage of black-box Machine Learning models in high-risk scenarios such as clinical trials and fraud detection, a need for safe, robust and trustworthy machine learning solutions with reliable outcomes becomes all the more paramount. Uncertainty quantification in anomaly detection applications helps the cause of trustworthiness in non-parametric models used in One-Class classification. While ensembles and the sampling approaches can quantify uncertainty by learning on varied distributions of data and aggregating multiple predictions on test data, making the results more robust, statistical guarantees for Type-I Errors are not provided by ensembling and sampling techniques. This is where conformal prediction comes into play, providing statistical guarantees for controlling Type-I errors (false positives) below a user-specified error threshold, whilst not compromising on the Type-II errors (false negatives). This work proposes $B_aKC+$, a novel approach for cross-conformal anomaly detection by combining K-fold cross-validation based cross-conformal prediction with ensembles and sampling techniques. $B_aKC+$ proves to be a model-agnostic, distribution-free uncertainty quantification technique for highly imbalanced datasets, providing conformal guarantees for Type-I errors whilst showcasing high statistical power. Without additional post-hoc operations for Type-I error control needed, $B_aKC+$ outperforms existing cross-conformal frameworks on benchmark anomaly detection datasets, and demonstrates itself to be a robust and reliable conformal anomaly detection framework, providing highly *certain* outcomes to the data analyst.

**Keywords:** Conformal Prediction, Distribution-free Uncertainty quantification, Outlier detection, Type-I Error control

## 1. Introduction

The One Class Classification (OCC) problem presents a new approach to solving the outlier detection problem, by learning exclusively from normal data, distinguishing outliers as variances from learned data. OCC methods include density estimation, boundary learning, and support vector data description, all aiming to define normal data patterns and identify outliers effectively. Support Vector Machines in particular implement OCC by constructing a decision boundary around normal data points, effectively distinguishing them from outliers. By maximizing the margin between normal data and the decision boundary, SVM helps detect anomalies reliably in high-risk applications.

[*] Zenon Analytics Pvt Ltd, BITS Pilani. Correspondence to: igarg2001@gmail.com
[†] Heriot Watt University

© 2025 I. Garg & S. Majumder.

However, due to the non-parametric nature of many of the Machine Learning models utilized in One Class Classification, interpretability of the outcomes of such models is a challenge, with a tendency to make highly confident but incorrect assumptions based on purely theoretical deductions. For example, Neural Network based Machine Learning models like MLP, BERT and GPT provide the predicted probability logits along with the class predictions, which are usually the softmax activated outputs of the underlying layers. Whilst these predicted probabilities act as confidence scores, the lack of calibration means that they carry an inherent level of uncertainty associated with them, that cannot be precisely quantified. This gap between the confidence scores and actual probability distributions in uncalibrated ML systems reduces the robustness and trustworthiness of the confidence scores, thereby reducing their interpretability and adoptability in highly critical systems. These limitations of OCC methodologies, combined with the highly imbalanced nature of datasets in outlier detection scenarios increase the significance of uncertainty estimation methods to quantify *epistemic* uncertainty, i.e, model uncertainty.

Another common issue prevalent among such OCC methodologies is the absence of statistical guarantees concerning their predictions. Consequently, the inability to quantify an estimator's uncertainty and subsequent error rates undermines its reliability. The need for uncertainty quantification in the field of outlier detection grows invariably as it finds its applications in high-risk and critical domains such as healthcare Bercea et al. (2024), banking Habibpour et al. (2021); Chaquet-Ulldemolins et al. (2022), insurance Garmdareh et al. (2023); Lu et al. (2020) and cybersecurity Phoha (2002); Chen et al. (2018), among others.

He and Jiang (2023) defines uncertainty quantification as "knowing what a deep neural network does not know". The main goal of uncertainty quantification is to accurately measure the underlying uncertainty associated with a model's predictions. The uncertainty can lie either at the model level (epistemic uncertainty) or at the data level (aleatoric uncertainty). The scope of this work is to discuss about methods than can be applied for quantifying the *epistemic uncertainty.*

Existing uncertainty quantification methods such as Bayesian neural networks and deep ensembles can be applied, however, they do not provide statistical guarantees for distribution free uncertainty quantification. Conformal prediction, on the other hand, do provide marginal guarantees in the form of conformal prediction sets, hence elevating the *certainty* of the outcomes supplemented by such *conformal guarantees.* Conformal prediction, therefore, aims to quantify the uncertainty associated with the confidence scores output by a non-parametric model.

As mentioned in Bates et al. (2023) and Angelopoulos and Bates (2021), the problem statement of conformal anomaly detection can be re-phrased as a hypothesis testing problem for finding out the points that are out-of-distribution (OOD) w.r.t the probability distribution of the seen data ($P(x)$). The null hypothesis can therefore be formalized as $H_{0,i} : X_i \sim P(x)$ for any $X_i \epsilon D_{test}$. For the provided hypothesis, statistically valid p-values can now be computed from the outputs of one-class classifiers.

This work introduces a novel conformal anomaly detection approach, combining K-fold cross-validation based cross-conformal prediction with ensembling and sampling methods and aims to achieve Type-I error control whilst uncompromising on statistical power, without post-hoc adjustments. This approach contributes to the existing set of conformal anomaly detection methods, enhancing non-parametric methods by providing marginal statistical guarantees on Type-I error control.

## 2. Literature Review

There have been numerous approaches to quantify uncertainty, both aleatoric and epistemic in the field of Machine Learning. Gawlikowski et al. (2023) describes various approaches such as dropout, Bayesian methods, ensemble methods and test time augmentation for measuring epistemic uncertainty. On the other hand, modelling anomaly detection as a one class classification problem has been surveyed extensively in Seliya et al. (2021). Methods such as OC-NN Chalapathy et al. (2018), Deep SVDD Ruff et al. (2018) and AnoGAN Schlegl et al. (2017) have been proposed in the context of novelty detection by modelling the problem as an OCC problem.

The application of Support Vector Machines in the field of novelty detection was first proposed in Schölkopf et al. (1999), in the form of One-Class SVM (OC-SVM). Existing applications of Conformal Prediction on top of SVMs exist in Balasubramanian et al. (2009), which calculates the non-conformity score as a function of the distance from the hyperplane, and then applies inductive conformal prediction on top of it.

Ensembles have been extensively researched to measure epistemic uncertainty. Lakshminarayanan et al. (2017) and Ovadia et al. (2019) propose frameworks of Deep NN Ensembles and evaluate the out-of-distribution (OOD) data detection on image classification tasks. In Scemama and Kapusta (2023), the non-conformity score is calculated as the Bayesian Model Average of each ensemble member's probability logits, and then split conformal prediction is applied on the same.

Cross conformal prediction was first introduced in Vovk (2015) and discussed in the survey Angelopoulos and Bates (2021). Various methods for performing cross-conformal prediction have been suggested, for example, the Jackknife Steinberger and Leeb (2023), and cross-validation based approaches in Barber et al. (2021). However, little work has been done in the application of cross conformal prediction in the field of anomaly detection. Hennhöfer and Preisach (2024) makes the choice of an Isolation Model as a fitter for the cross conformal prediction approaches, and where it mentions the Jackknife-after-Bootstrap approach Kim et al. (2020), it leaves its implementation for future work. Kim et al. (2020) introduces the Jackknife+-after-Bootstrap approach, marrying conformal prediction with ensembles and applies a bootstrapping operation, similar to what this work aims to accomplish. However, the aforementioned work applies Leave-one-out-validation (LOOV), as its cross-conformal framework, and the overall approach is only evaluated in a *regression* setting.

### 2.1. Research Gaps

Based on the literature review undertaken, the following research gaps have been identified:

- There is an absence of work that combines ensemble learning with cross-conformal prediction for solving the problem of anomaly detection. The Jackknife+-after-Bootstrap approach is evaluated in a regression setting in Kim et al. (2020), and while mentioned in Hennhöfer and Preisach (2024), the implementation and evaluation is left for future scope.

- Existing cross-conformal approaches, for example, the Jackknife+ and CV+ Barber et al. (2021) do not make the use of ensemble learning to calculate the calibration scores for a given random variable. Kim et al. (2020) combines ensemble learning with the Jackknife+ to efficiently compute residual values for the LOO variables. The application of ensemble learning to improve the robustness of the calibration scores themselves has not been explored in current research.

## 3. Methodology

The aim of this work is to present a combination of OC-SVM ensembles and cross-conformal predictors and evaluate the proposed framework on a set of anomaly detection tasks. Given a dataset $D = \{x_1, x_2, \ldots, x_n\} = [X_i]_{i=1}^{n}$ divided into two parts $D_{train} = [X_i]_{i=1}^{n_{train}}$ and $D_{test} = [X_{n_{train}+i}]_{i=1}^{n_{test}}$ consists of $n$ observations, each observation being of dimension $f$, where $D_{train} \cap D_{test} = \emptyset$.

It is considered that all observations in $D_{train}$ are non-anomalous, and $D_{test}$ exclusively consists of anomalies.

For the underlying estimator, one class SVM, the non-conformity score is defined as follows:

$$\Lambda(x) = sigmoid(-\lambda) \tag{1}$$

where, $\lambda = \delta_h(X) - \phi$; $\phi \equiv$ offset of hyperplane from the origin.

As proposed in Barber et al. (2021), the training data is divided into equal and disjoint subsets (or folds) $F_1, F_2, F_3, \ldots, F_K$. CV+ for K-fold cross-validation approach proposed in Barber et al. (2021) fits a regression function on $-F_k$, where $-F_k$ is the training data with the $k^{th}$ fold removed. This paper follows the approach in Barber et al. (2021) with a sampling approach, replacing the regression function with a one-class classifier function. The novel approach performs out-of-bag sampling (or *partitioning*) by creating $M$ equal and disjoint sets out of each out-of-fold set of observations, leaving $-B_{-k,1}, -B_{-k,2}, \ldots -B_{-k,M}$ out-of-bag observations. Thus, $M$ scoring functions can be fit on each out-of-bag observation set and can be validated on both the in-fold and in-bag observations for enhanced calibration sets.

An ensemble of OC-SVM models is utilized for training, with the number of members being equal to the number of bags created during the sampling process, i.e, $M$. Each $M_m$

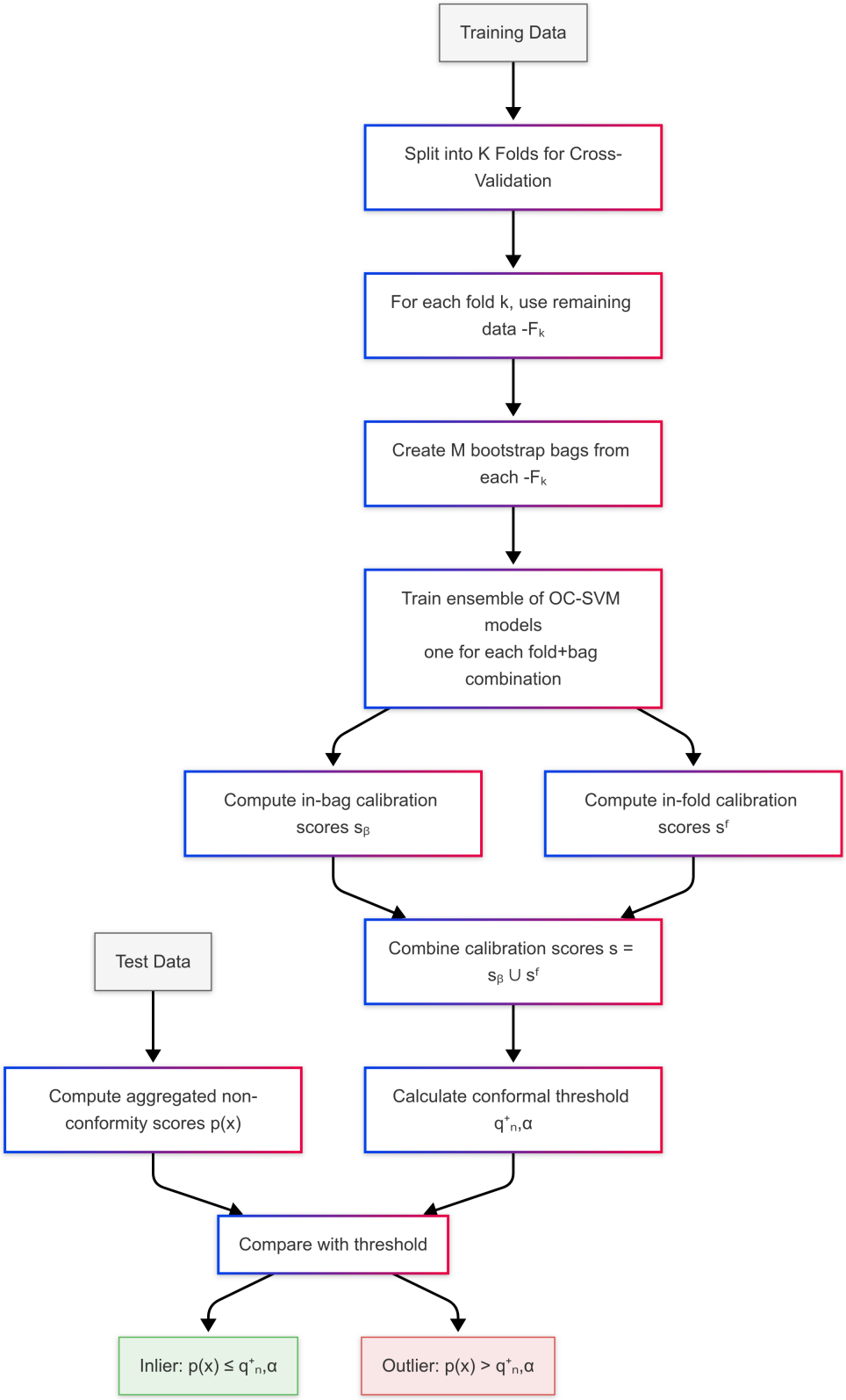


Figure 1: Overview of the BaKC+ methodology, showing the integration of K-fold cross-validation, sampling, and ensemble learning for conformal prediction in anomaly detection. The workflow demonstrates how calibration scores from both in-fold and in-bag sets combine to calculate the conformal threshold for determining inliers versus outliers.

member can then be trained on the out-of-fold subset $-F_k$, while leaving out exactly one bag $B_{-k,m}$ from training. Therefore, we end up with $m * k$ scoring functions as follows:

$$\hat{\mu}(-F_k, -B_{-k,m}) = \Lambda(X_i \colon i \in \{1, 2, \ldots, n\} \mid (F_k \cup B_{-k,m})) \tag{2}$$

, where $\hat{\mu}_{-F_k,\ -B_{-k,\ m}}$ denotes the scoring function fit on to the training data with the $k^{th}$ fold and the $m^{th}$ bag removed. Calibration scores are thus computed on the held-out subsets of training data, i.e., the in-fold and in-bag sets.

The calibration scores for the in-bag sets are computed as follows:

$$\widehat{s_b} = \left| \hat{\mu}_{-F_k,\ -B_{-k,\ m(i)}} (X_i) \right| \tag{3}$$

where, $i \in B_{-k,m}, k \in \{1, \ldots K\},\ m(i) \in \{1, \ldots M\},$

where, $m(i)$ denotes the value of $m$ for which the out-of-bag subset $-B_{-k,m}$ does not contain the point $X_i$,

whereas the calibration scores for the in-fold sets are aggregated over each ensemble members score, and the resultant score is as follows:

$$\widehat{s_f} = \left| Agg \left( \hat{\mu}_{-F_{k(i)},\ -B_{-k(i),m}} (X_i) \right)_{m=0}^{M} \right|, \tag{4}$$

where $i \in F_k$, $k \in \{1, \ldots, K\}$ and $k\,(i)\ \in \{1, \ldots,\ K\}$ denotes the in-fold subset containing the data point $i$. As defined in Angelopoulos and Bates (2021), the adjusted quantile is computed as:

$$q^{+}_{n,\ \alpha}\{\hat{s}\}$$

where for a user-chosen error rate $\alpha$, $q^{+}_{n,\ \alpha}$ is defined as the $\lceil \frac{(n+1)(1-\alpha)}{n} \rceil^{th}$quantile of the calibration score set $\hat{s} = \widehat{s_b} \cup \widehat{s_f}$.

As discussed in Angelopoulos and Bates (2021), when a test point is encountered, it can be classified as follows:

$$C(x) = \begin{cases} \text{inlier} & p(x) \leq q^{+}_{n,\ \alpha} \\ \text{outlier} & p(x) > q^{+}_{n,\ \alpha} \end{cases} \tag{5}$$

where, $p\,(x) = Agg \left( Agg \left( \hat{\mu}_{-F_k,\ -B_{-k,m}} (X_i) \right)_{m=1}^{M} \right)_{k=1}^{K}$, $\qquad X\,(i)\ \in D_{test}$

For a concise definition of the terms used in this section, see Appendix A.

## 4. Results and Evaluation

This section covers the evaluation of our approach and compares it to existing anomaly detection methods, both regular and conformal-based approaches. This work utilizes `scikit-learn's` implementation of the OneClassSVM model with its default parameter configuration Pedregosa et al. (2011), and the value of $nu = 0.05$ as its base estimator.

### 4.1. Experiment Details

The setup of the experiment was followed loosely on the setup proposed in Hennhöfer and Preisach (2024) and Bates et al. (2023). $D_{train}$ is divided into $T$ non-disjoint datasets $[D_t]_{t=1}^{T}$. Each $D_t$ is further divided into 2 sets $D_{train\_cal}$ and $D_{train\_test}$. $D_{train\_cal}$ is then utilized for the training and calibration procedure as discussed in Section 3, with $D_{train\_test}$ held out for testing. Therefore, $T$ training and calibration cycles are performed as part of the experiment. Furthermore, each training cycle $t \in [1, \ T]$ is accompanied by $L$ disjoint test sets $\left[D_{t,l}^{eval}\right]_{l=1}^{L}$, each of size $\frac{D_{eval}}{L}$, where,

$$D_{eval} = D_{test} \cup D_{train\_test} \tag{6}$$

As in Hennhöfer and Preisach (2024), this work evaluates the proposed approach by covering the two kinds of scenarios that can occur in an outlier detection system. A typical observation (inlier) has the potential to be mistakenly identified as an anomaly (leading to a false positive), or inversely, an anomaly might go unnoticed (resulting in a false negative). Keeping that in mind, two metrics were chosen to be calculated, the false discovery rate (FDR) and the statistical power. For any given $t \in T, l \in L$, the power is calculated as:

$$P_w\left(D_{t,l}^{test} \mid D_t\right) = \frac{TP}{TP + FN} \tag{7}$$

Where $TP$ represents the outliers correctly classified as outliers, and $FN$ represents the outliers incorrectly classified as inliers. The power is averaged over each test cycle $l \in L$, and then again over each train cycle $t \in T$, to provide the statistical power as calculated below:

$$P_{stat}\left(D_{test}\right) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L} P_w\left(D_{t,l}^{test} \mid D_t\right) \tag{8}$$

For any given $t \in T, l \in L$, the false discovery proportion is calculated as:

$$FDP\left(D_{t,l}^{test} \mid D_t\right) = \frac{FP}{TP + \ FP} \tag{9}$$

Where $TP$ represents the inliers correctly classified as inliers, and $FP$ represents the inliers incorrectly classified as outliers. The false discovery rate (FDR) for each train cycle $t \in T$ is calculated by averaging the false discovery proportion over $L$ test cycles. The FDR for an experiment is calculated as the 90th quantile of all false discovery rates obtained for each training cycle. Therefore, the FDR for an experiment comes out to be:

$$FDR\left(D_{test}\right) = Q_{90}\left[\left(\frac{1}{L}\sum_{l}^{L} FDP\left(D_{t,l}^{test} \mid D_t\right)\right)\right]_{t=1}^{T} \tag{10}$$

For each training cycle $t \in T$, $|D_{train\_cal}| = |D_{train\_test}| = \frac{|D_t|}{2}$. For performing K-fold cross validation, the size of each fold is maintained equal at $|F_k| = \min\left(2000, \ \frac{|D_{train\_cal}|}{3}\right)$. Therefore, the value of $K$ is calculated as $\frac{|D_{train\_cal}|}{\min\left(2000, \frac{|D_{train\_cal}|}{3}\right)}$. For very large datasets with

$|D_{train\_seen}| \geq 20000$, the value of $K$ is fixed at 20, keeping in mind compute costs.

For each of the datasets, $T = 10$ and $L = 20$ is chosen. The user-chosen error rate is $\alpha = 0.1$

### 4.2. Comparison Details

For the purposes of comparing with existing methods, a vanilla One Class SVM with the same configuration as that of the novel approach is evaluated as a baseline. $D_{train}$ is divided into two disjoint datasets $D_{train\_seen}$ and $D_{train\_unseen}$. The SVM model is then fit on $D_{train\_seen}$ and evaluated on $D_{eval}$ where $D_{eval} = D_{test} \cup D_{train\_unseen}$. The evaluation function used here is $\lambda$, where $\lambda$ represents the `decision_function` API of the `OneClassSVM` library provided by `scikit-learn`. As per Schölkopf et al. (1999), $\lambda \geq 0$ is considered as an inlier, while $\lambda < 0$ is considered as an outlier.

Existing conformal prediction-based approaches like split conformal prediction, Jackknife and CV-based approaches have also been considered for evaluation, however, due to the constraints in time, such approaches could not be applied on top of OneClassSVM as a base estimator. However, the performance of such methods is evaluated in Hennhöfer and Preisach (2024) with IsolationForest as a base estimator, with Power and FDR being the metrics chosen for evaluation. This paper presents the metrics (Average Power and 90th quantile FDR) for the CV+ approach presented in Hennhöfer and Preisach (2024) alongside the above discussed approach and baseline approach for reference. The evaluation of existing conformal methods on top of OneClassSVM is left for future scope.

The proposed approach is also compared with multidimensional models like autoencoders, which are suitable for one-class classification tasks. Autoencoders can perform representation learning on the input data and utilize an encoder-decoder architecture to capture non-linear relationships in multidimensional datasets. This is a powerful approach that can be combined with reconstruction loss metrics like Mean Squared Error (MSE) to measure how well the test data matches the data seen during training. This reconstruction loss technique can be utilized to detect outliers in the test data. While autoencoders are extremely powerful, they do not inherently quantify the uncertainty in the predictions, nor do they provide any statistical guarantees.

For the sake of interpretation, the following labels will be used for all the approaches used for comparison:

- $B_aKC+$ : The proposed novel approach applying cross-conformal prediction using K-fold cross-validation onto an ensemble of SVM's trained with the sampling approach.
- $SVM$ : The baseline OneClassSVM estimator trained on 50% inliers and tested on the remaining inliers and the entire outliers dataset.
- $CV+$ : The cross-validation based conformal prediction approach proposed in Hennhöfer and Preisach (2024).

Table 1: Outcome of the statistical power evaluation performed on the novel approach with $\alpha = 0.1$, and the corresponding comparison with baseline, CV+ ($\alpha = 0.2$) and autoencoder (AE) approaches. *Higher is better*

| **Statistical Power** | | | | |
|---|---|---|---|---|
| Dataset | $B_aKC+$ | SVM | CV+ | AE |
| Shuttle | **.993** | .991 | .982 | .978 |
| Mammography | .523 | .534 | .111 | **.584** |
| Cardio | **.984** | .952 | .460 | .903 |
| Gamma | **.477** | .414 | .181 | .323 |
| Musk | **1.0** | **1.0** | .982 | .887 |
| Fraud | .880 | .847 | .684 | **.910** |

- *AE*: The multi-dimensional autoencoder model trained on on 50% inliers and tested on the remaining inliers and the entire outliers dataset. For training details, see Appendix B.

### 4.3. Dataset Details

For the purposes of evaluating an outlier detection solution in a multitude of environments and scenarios, a diverse assembly for datasets, encompassing multiple domains, dimensions and outlier proportions is essential. This was the primary reason for selecting the anomaly detection benchmark ADBench Han et al. (2022). ADBench is a compilation of datasets catering multiple dimensions and domains and is diverse in terms of outlier proportions in its datasets. For the scope of this paper, 6 datasets covering multiple disciplines like Finance, Healthcare and Astronomy. The choice of these datasets was made in the view of covering wide-ranging dataset sizes, diverse outlier %age, and differing dimension ranges.

### 4.4. Results

Table 1 demonstrates the superior statistical power of $B_aKC+$ in comparison with a vanilla SVM model, an autoencoder model and existing cross-conformal approaches applied on top of Isolation Forest estimators. Both the novel approach and the vanilla SVM model significantly outperform the Isolation Forest-based CV+ approach as presented in Hennhöfer and Preisach (2024). $B_aKC+$ outperforms the vanilla SVM based approach on all datasets but 1, the mammography dataset. In case of the mammography dataset, vanilla SVM reports better statistical power than $B_aKC+$, however is outperformed on FDR. $B_aKC+$ also outperforms the autoencoder on 6 datasets, falling behind in the mammography and fraud datasets. However, in both these datasets, $B_aKC+$ exhibits better FDR than AE. While the statistical power of $B_aKC+$ can be observed to be better with smaller datasets, achieving perfect power for the musk dataset, it also performs well with very large datasets like shuttle ($\sim 50,000$ rows) and fraud ($\sim 280,000$ rows).

The novel approach $B_aKC+$ also significantly outperforms both the vanilla SVM-based

Table 2: Outcome of the false discovery rate evaluation performed on the novel approach with $\alpha = 0.1$, and the corresponding comparison with baseline, CV+ ($\alpha = 0.2$) and autoencoder (AE) approaches. *Lower is better*

| **False Discovery Rate** | | | | |
|---|---|---|---|---|
| Dataset | $B_aKC+$ | SVM | CV+ | AE |
| Shuttle | **.086** | .096 | .206 | .096 |
| Mammography | **.032** | .067 | .218 | .095 |
| Cardio | **.076** | .102 | .287 | .119 |
| Gamma | **.077** | .330 | .226 | .100 |
| Musk | **.054** | .075 | .109 | .099 |
| Fraud | .096 | **.070** | .175 | .101 |

approach and the Isolation Forest-based CV+ approach, as well as the autoencoder, with regard to the False Discovery Rate (FDR), as can be seen in Table 2. At the same time, $B_aKC+$ also maintains the conformal coverage guarantee for the inlier class, maintaining the false discovery rates less than $\alpha$ for all datasets. This is empirically proven by Table 2. On the other hand, $B_aKC+$ is outperformed on the fraud dataset by its vanilla counterpart in respect to FDR, however reports superior statistical power on the same dataset.

Overall, the proposed approach $B_aKC+$ exhibits improved statistical power outcomes compared to its counterparts while maintaining the False Discovery Rates below the user-chosen error tolerance. An important point to keep in mind regarding conformal approaches is the chosen base estimator and its capability to distinguish outliers from in-distribution data. Furthermore, as described in Angelopoulos and Bates (2021), the choice of the scoring function is one of the primary factors impacting the effectiveness of calculated calibration scores and test scores.

## 5. Discussion

This section covers some points of observations over the results seen in Section 4.

### 5.1. Discussion on FDR control

The evaluation metrics displayed in Table 2 empirically prove that the proposed approach, $B_aKC+$ can control the False Discovery Rate below the chosen value of $\alpha$, regardless of the dataset chosen. Following theoretical evidence that sampling techniques (like bootstrapping) on ensembles coupled with conformal prediction achieve FDR control intrinsically, methods such as $B_aKC+$ possess a notable advantage over existing cross-conformal anomaly detection methods Hennhöfer and Preisach (2024), which make the use of post-hoc methods such as Benjamini and Hochberg (1995) to solve the multiple testing Tukey (1953) problem.

Intrinsic FDR control can provide computational benefits to the testing process of a cross-

conformal prediction framework wrapped on to an ensemble of estimators, as compared to post-hoc analyses like Benjamini and Hochberg (1995), where p-values for each hypothesis would typically have to be computed for each ensemble member. Furthermore, it makes the validation part of the framework less complicated and reduce development efforts, at the cost of a comparatively more expensive training process. However, as mentioned above, this deduction is based purely on empirical evidence, and furnishing theoretical evidence for error control guarantees is left for future work.

### 5.2. Important note on the value of $K$ for cross-validation

It is clear from the results that cross-conformal approaches perform exceptionally well with smaller datasets in context of both FDR and Statistical Power. This was expected because of the larger values of $\frac{|D_{train\_cal}|}{|D_{train}|}$. For larger datasets, the value of $K$ was fixed at 20, which meant that the base estimators were trained with a *smaller* training corpus, as compared to the vanilla approaches. This directly translates into increased *generalization* and reduced *sensitivity* [1] to the splitting approach across datasets of varying dimensions and outlier proportions, while making the outcomes more reliable and actionable for high-risk scenarios.

### 5.3. Important note on the chosen value of $\alpha$

A point to be noted with respect to the evaluation setting of this work is the chosen value of $\alpha$. It is commonly acknowledged that a trade-off exists between statistical power and the False Discovery Rate (FDR). An increase in FDR is usually accompanied by a reduction to statistical power. $\alpha$ value of 0.1 was considered to be an acceptable balance between the two metrics. Decreasing the value of $\alpha$ to 0.05 would maintain the FDR below 0.05, however more outliers would miss detection, thus bringing down the statistical power. Whereas, increasing the value of $\alpha$ to 0.2, would categorise more observations as outliers, while increasing the False Discovery Rates, but controlling them below 0.2. The advantage presented by conformal prediction is the controlled false discovery rates, which directly affects the statistical powers, hence imparting more *certainty* into the outcomes.

## 6. Practical Implications

Besides exploring new avenues in conformal prediction, the results of this research have major practical implications for several relevant domains, such as Finance, Healthcare, Astrophysics, Energy and Manufacturing, to name a few Conformal prediction methods, be it inductive Laxhammar and Falkman (2010), cross Vovk (2015); Barber et al. (2021); Hennhöfer and Preisach (2024); Kim et al. (2020) or full Shafer and Vovk (2008), impart a greater degree of certainty to non-deterministic Machine Learning model outcomes by providing statistical guarantees over controlling the false alarms and at the same time, not allowing too many missed detections.

1. Not to be confused with the sensitivity metric (also known as recall) to assess model performance. Here, sensitivity implies a change in the model behaviour and/or outcomes on changing one or more aspects of the model training process.

Angelopoulos and Bates (2021) has a complete section on some of the applications of conformal prediction, such as multi-label classification, tumor segmentation and outlier detection in toxic comments. In general, conformal prediction methods can be applied to all applications of other uncertainty quantification methods such as those outlined in Gawlikowski et al. (2023); He and Jiang (2023). Balasubramanian et al. (2014) cites works applying cross-conformal prediction methods in multidisciplinary domains like bio-metrics and face recognition, diagnostic and prognostic applications in the biomedical space and demand forecasting in network traffic.

Conformal anomaly detection methods which do not require the use of multiple testing Tukey (1953) to control the False Discovery Rate, ($B_aKC+$ falls under the umbrella of such methods) can also be applied to online anomaly detection problems and can impart better certainty to outcomes generated in real-time applications such as telemetry, stream processing and Big Data.

## 7. Conclusion and Future Work

This work presents a novel approach to domain-agnostic, distribution-free and uncertainty-aware outlier detection by combining existing approaches solving both anomaly detection and uncertainty quantification, namely, ensembles, sampling, cross-validation and conformal prediction. Next, this work evaluates the proposed approach on benchmark datasets for anomaly detection which are also used in comparative studies, and compares the results with a baseline approach and existing cross-conformal approaches. The evaluations performed demonstrate superior overall performance of the proposed approach over the compared methods, as well as superior individual performance on both statistical power and false discovery rate aspects.

Both research gaps underlined in Subsection 2.1 have been addressed in this work. The proposed approach combines ensemble learning with cross-conformal prediction and applies it to the problem of anomaly detection. The proposed methodology reaps the benefits of ensemble learning in the computation of the calibration scores, enhancing their robustness and trustworthiness.

Technical advantages of the proposed approach include higher statistical powers than corresponding counterparts, whilst controlling the FDR below a chosen threshold. This is achieved without the use of post-hoc multiple testing methods like the Benjamini-Hochberg procedure. Technical advantages of conformal prediction based approaches are also demonstrated in this work, such as, 1) the statistical guarantees associated with the metrics, and 2) the finite-sample validity demonstrated by the smaller datasets like Cardio and Musk Angelopoulos and Bates (2021). As discussed in Hennhöfer and Preisach (2024), the model-agnostic nature of all conformal prediction methods applies to the proposed approach of this work as well, wherein the proposed approach can elegantly be fit onto existing anomaly detection methods and provide statistical guarantees on Type-1 Errors. Aside from the exchangability assumption, conformal prediction makes no assumptions on the underlying distribution of data, which results in the framework being applicable towards any kind of

domain or disciplinary dataset.

Practical applications of the proposed approach can be in various anomaly detection problems not explored in this work, like text classification (toxic comments detection, false news detection) and image classification. Conformal prediction itself can find future applications in various multidisciplinary domains within the Machine Learning landscape. Conformal prediction methods can be applied to time-series anomaly detection problems, and some of the basis for this has been proposed in Xu and Xie (2023). Conformal prediction methods can also be applied to generative models like in Wang et al. (2022); Quach et al. (2023) as an approach to reduce noise and hallucinations, and generate output tokens with statistical guarantees. Conformal prediction methods can also be applied to the field of graph neural networks to quantify node classification uncertainty for Natural Language Processing (NLP) classification as well as regression tasks.

Limitations of this work include the increasing training costs of a ensemble setup paired with sampling, wrapped with a $K$-fold cross-validation framework and added calibration steps for both in-fold and in-bag subsets. Based on empirical evidence, the computational complexity for training increases exponentially as $|D_{train}|$ increases, with training times of 14s for smaller datasets like Musk and Gamma, ($|D_{train}| < 5000$) rising to 8 minutes for Shuttle ($|D_{train}| \sim 50000$) and 6 hours for the Fraud dataset. ($|D_{train}| \sim 280000$).[2]

Future work may involve seeking ways to reduce the training times of the proposed methodology by utilizing methods similar to Kim et al. (2020), where out-of-bag trained models are reused to compute the leave-one-out models, similarly, a method by which out-of-fold estimates can be obtained from the out-of-bag estimates and do not require re-fitting the base estimator on the out-of-fold subsets. This would make the overall complexity of the training operation $O(M)$, instead of $O(K \cdot M)$, significantly reducing the training time for larger datasets like Shuttle and Fraud.

On another aspect, future work may include formalizing the theoretical proofs for the conformal guarantees provided by the proposed approach, to complement the empirical proofs provided by this work.

To conclude, the proposed approach provides a new dimension to a promising and well-researched field in machine learning and statistics and extends it to real-world application scenarios, and re-iterate the belief stated in Angelopoulos et al. (2021) about the expanding reputation of statistical procedures and their ability to provide measurable interpretations on the performance of non-parametric models.

2. Mentioned training times are for a slightly modified experiment setting with $T = 4$ and run on a CPU-only Jupyter notebook with 4 CPUs utilized for parallelization.

## References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

V. Balasubramanian, R. Gouripeddi, S. Panchanathan, J. Vermillion, A. Bhaskaran, and R. Siegel. Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure. In *2009 36th Annual Computers in Cardiology Conference (CinC)*, pages 5–8, 2009. URL https://ieeexplore.ieee.org/document/5445485.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2014. ISBN 0123985374.

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021. URL https://arxiv.org/abs/1905.02928.

Stephen Bates et al. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Towards universal unsupervised anomaly detection in medical imaging. *arXiv preprint arXiv:2401.10637*, 2024.

Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018. URL https://arxiv.org/abs/1802.06360.

Jacobo Chaquet-Ulldemolins, Francisco-Javier Gimeno-Blanes, Santiago Moral-Rubio, Sergio Muñoz-Romero, and José-Luis Rojo-Álvarez. On the black-box challenge for fraud detection using machine learning (ii): Nonlinear analysis through interpretable autoencoders. *Applied Sciences*, 12(8), 2022. ISSN 2076-3417. doi: 10.3390/app12083856. URL https://www.mdpi.com/2076-3417/12/8/3856.

Li Chen, Salmin Sultana, and Ravi Sahita. Henet: A deep learning approach on intel r processor trace for effective exploit detection. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 109–115, 2018.

François Chollet et al. Keras. https://keras.io, 2015.

Mahdi Sharifi Garmdareh, Behzad Soleimani Neysiani, Mohammad Zahiri Nogorani, and Mehdi Bahramizadegan. A machine learning-based approach for medical insurance anomaly detection by predicting indirect outpatients' claim price. In *2023 9th International Conference on Web Research (ICWR)*, pages 129–134, 2023. doi: 10.1109/ICWR57742.2023.10139290.

Jakob Gawlikowski et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023. URL https://arxiv.org/abs/2107.03342.

Maryam Habibpour, Hassan Gharoun, Mohammadreza Mehdipour, AmirReza Tajally, Hamzeh Asgharnezhad, Afshar Shamsi, Abbas Khosravi, Miadreza Shafie-Khah, Saeid Nahavandi, and Joao PS Catalao. Uncertainty-aware credit card fraud detection using deep learning. *arXiv preprint arXiv:2107.13508*, 2021.

Songqiao Han et al. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 35:32142–32159, 2022. URL https://arxiv.org/abs/2206.09426.

Wenchong He and Zhe Jiang. A survey on uncertainty quantification methods for deep neural networks: An uncertainty source perspective. *arXiv preprint arXiv:2302.13425*, 2023.

Oliver Hennhöfer and Christine Preisach. Uncertainty quantification in anomaly detection with cross-conformal $p$-values. *arXiv preprint arXiv:2402.16388*, 2024. URL https://arxiv.org/abs/2402.16388.

Byol Kim, Chen Xu, and Rina Barber. Predictive inference is free with the jackknife+-after-bootstrap. *Advances in Neural Information Processing Systems*, 33:4138–4149, 2020. URL https://arxiv.org/abs/2002.09025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://arxiv.org/abs/1612.01474.

Rikard Laxhammar and Göran Falkman. Conformal prediction for distribution-independent anomaly detection in streaming vessel data. In *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, StreamKDD '10, page 47–55, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450302265. doi: 10.1145/1833280.1833287. URL https://doi.org/10.1145/1833280.1833287.

Jiaqi Lu, Benjamin C. M. Fung, and William K. Cheung. Embedding for anomaly detection on health insurance claims. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 459–468, 2020. doi: 10.1109/DSAA49011.2020.00060.

Yaniv Ovadia et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://arxiv.org/abs/1906.02530.

Pedregosa et al. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830, 2011.

Vir V. Phoha, editor. *Internet Security Dictionary*. Springer New York, New York, NY, 2002.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ruff18a.html.

Paul Scemama and Ariel Kapusta. On the out-of-distribution coverage of combining split conformal prediction and bayesian deep learning. *arXiv preprint arXiv:2311.12688*, 2023. URL https://arxiv.org/abs/2311.12688.

Thomas Schlegl et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *International Conference on Information Processing in Medical Imaging*, 2017. URL https://arxiv.org/abs/1703.05921.

Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99)*, pages 582–588, Cambridge, MA, USA, 1999. MIT Press. URL https://dl.acm.org/doi/10.5555/3009657.3009740.

N. Seliya, A. Abdollah Zadeh, and T.M. Khoshgoftaar. A literature review on one-class classification and its potential applications in big data. *J Big Data*, 8:122, 2021. URL https://doi.org/10.1186/s40537-021-00514-x.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, jun 2008. ISSN 1532-4435.

Lukas Steinberger and Hannes Leeb. Conditional predictive inference for stable algorithms. *The Annals of Statistics*, 51(1):290–311, 2023. URL https://arxiv.org/abs/1809.01412.

J. W. Tukey. The problem of multiple comparisons. Unpublished manuscript. See Braun (1994), pp. 1-300., 1953.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015. URL https://arxiv.org/abs/1208.0806.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David M Blei. Probabilistic conformal prediction using conditional random samples. *arXiv preprint arXiv:2206.06584*, 2022.

Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

## Appendix A. Glossary of Terms

$X_i$**:** Random variable corresponding to a point in multidimensional space. In the context of the evaluation in this work, it represents the $i^{th}$ data point in the dataset, with the target feature removed.

$Y_i$**:** The target class for the $i^{th}$ data point. In the context of this work, it is label-encoded to integer values for the computation of calibration scores.

$\alpha$**:** Denotes a user-chosen error rate. As specified in Angelopoulos and Bates (2021), the probability of the true class to be part in the prediction set is statistically *guaranteed* to be *atleast* $1 - \alpha$ .

$C(x)$**:** Denotes the conformal prediction set for the point $x$. The *conformal prediction sets* contains the set of possible target classes that the random variable $x$ can hold, with the probability of each target class in the set required to be greater than $1 - \alpha$. In the context of this work, the size of the conformal prediction set is 1 for every data point, due to the nature of the problem being one-class.

$\delta_h(x)$**:** In the context of OneClassSVM, denotes the signed distance of the point $x$ from the origin.

$\Lambda(x)$**:** Denotes the non-conformity score that would be used to create the calibration sets.

$\lambda(x)$**:** Denotes the signed distance of the point $x$ from the hyperplane.

$\hat{\mu}(-A, -B)$**:** Denotes a scoring function that is fit onto some dataset $D$ with the subset $A \in D$ and $B \in (D - A)$ removed from $D$.

$Agg(f(x))_{x=1}^{X}$**:** Denotes a generic aggregation function applied over the function $f(x)$ with the given bounds of the variable $x$. In the context of this work, the aggregation function chosen for the evaluation in Section 4 is *median*.

$F_k$**:** Denotes the $k^{th}$ fold of the dataset split using the K-fold cross validation technique.

$B_{-k,m}$**:** Denotes the $m^{th}$ bag drawn from the $k^{th}$ *out-of-fold* subset of the dataset.

$\hat{s}(x)$**:** Denotes the calibration score (also referred to as *residual score* in the regression setting in some works) for the point $x$. A lower calibration score denotes more *conformity* of the data point $x$ to the seen data.

$q^{+}_{n,\ \alpha}(\hat{s})$**:** Denotes the adjusted quantile of the calibration score sets in accordance with the user-chosen error rate $\alpha$. For a user-chosen error rate $\alpha$, $q^{+}_{n,\ \alpha}$ is defined as the $\lceil \frac{(n+1)(1-\alpha)}{n} \rceil^{th}$ quantile (higher) of the calibration score set $\hat{s}$ .

$p(x)$**:** Denotes the aggregated non-conformity score of a test point, which is compared to the threshold non-conformity score, in order to classify the point $x$ as an inlier or outlier.

## Appendix B. Training Details

### B.1. OneClassSVM

The OneClassSVM implementation applied in this work is the one provided by `scikit-learn`. The value of *nu* is tuned to 0.05. The kernel is chosen as *RBF*, due to its capability of capturing non-linear representations in the training data. The other parameter values have been left as default.

### B.2. AutoEncoder

The autoencoder implementation applied in this work is the one provided by `Keras` Chollet et al. (2015). Following are the implementation details regarding the Model Architecture and Training details:

#### B.2.1. Model Architecture

1. Input Layer: The input layer is kept as standard, i.e, the set of training data vectors is passed to the input layer.
2. Encoder Layer: The encoder layer is implemented as a fully connected layer, with the encoding dimension set to 14. The activation function used for the encoding layer is ReLU.
3. Decoder Layer: The decoder layer is also implemented as a fully connected layer, wherein the activation function is chosen as Sigmoid, in order to aid reconstruction tasks.

For the purposes of model compilation, Adam is chosen as the optimizer with a learning rate of .001. The loss function is chosen as Mean Squared Error (MSE), which is a type of reconstruction loss function, and is a measure of how well the reconstruction has taken place.

#### B.2.2. Training and Evaluation Details

For the purposes of training, the number of epochs was set to 50 and the batch size set to 32. The training process includes shuffling the data before each epoch.

The MSE metric is responsible for categorizing a given data point as an anomaly. Higher MSE loss values indicate lesser agreement between the data point and the input data. However, in order to predict anomalies deterministically, a threshold needs to be set, above which, MSE loss values will categorize the point as an anomaly. This work computes the threshold as the $90\%ile$ of the reconstruction error. This ensures that, depending on the percentile value, only the highest MSE losses are considered as anomalies.