Choosing Effective Projections for Fast and Accurate Anomaly Detection

Chen Almagor   Yedid Hoshen
School of Computer Science and Engineering
The Hebrew University of Jerusalem

ABSTRACT

Successful anomaly detection methods require making accurate assumptions on the statistics of normal and anomalous data. Tabular anomaly detection is particularly challenging due to the great diversity of distributions between different datasets. One class of methods, combines multiple classifiers, each first linearly projecting the sample features to a scalar and then estimating its probability density in 1D. Departing from previous methods, our focus in this work is to determine the optimal directions for projection. We first identify that multimodality of the distribution affects the optimal projection set. Our experiments show that for unimodal data, the principal component directions are an effective choice, while for multimodal data, the raw axes are better. We begin by proposing a simple baseline of choosing the projection directions based on whether the dataset is unimodal or multimodal. However, due to the unsupervised setting, this strategy requires a reliable unsupervised unimodality statistical test. To remove this requirement, we propose a more robust solution that does not require determining if the data is unimodal. Our method, S-Chimera, uses a consensus-based approach to select the best subset of projections from the combined set of principal and raw axes. Our method is evaluated on a large number of high-dimensional datasets and is shown to outperform top established methods, as well as recent deep learning methods, while being orders-of-magnitude faster. The results are demystified through ample analyses.

CCS CONCEPTS

• **Computing methodologies → Anomaly detection**.

KEYWORDS

anomaly detection, tabular data, naive bayes, ensemble method

ACM Reference Format:
Chen Almagor   Yedid Hoshen . 2021. Choosing Effective Projections for Fast and Accurate Anomaly Detection. In *Proceedings of Outlier Detection and Description Workshop (ODD'21).* ACM, New York, NY, USA, 11 pages.

1 INTRODUCTION

Anomaly detection methods aim to identify unusual patterns in data. As identifying what is 'usual' and 'unusual' is essentially a probabilistic task, a common approach is to estimate the probability density function (PDF) of normal data, and then score test samples as normal or anomalous by thresholding the probability of the sample. Samples that obtain low probability under the distribution of normal data are labelled as anomalous while those that are deemed likely are labelled normal. Estimating high-dimensional distributions is not an easy task, particularly under a limited number of training samples. Although non-parametric methods such as K-nearest neighbors or KDE make relatively few assumptions, they suffer from the "curse of dimensionality". Parametric methods such as Mahalnobis or mixture-of-gaussians have lower sample complexity, but make very strong assumptions on the distribution which may not be satisfied in practice. In this paper, we concentrate on a class of methods that tackles the more manageable task of estimating the PDF of the data after linear projection to a scalar along a set of pre-defined directions.

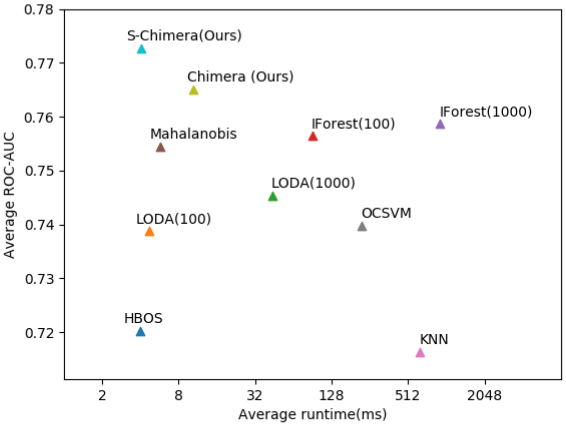

Figure 1: Average ROC-AUC vs. inference runtime: Our method S-Chimera achieves the top average ROC-AUC over the highest-dimensional ODDS datasets while having one of the fastest inference speeds.

We present a framework that generalizes projection-based anomaly detection methods. This framework consists of four main parts: (i) Composing a set of linear projection directions to 1D. (ii) Projecting the data. (iii) Estimating the 1D marginal probability distributions. (iv) Combining the marginal probabilities. We show that popular existing methods - HBOS, LODA and Mahalanobis correspond to special-cases of this framework.

Our focus in this work is to determine the optimal directions for projection. This differs from previous work such as LODA that uses random projection directions, which we show is suboptimal. We

ODD'21, August 2021, Virtual

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

identified that the number of modes in the distribution of normal data is critical to the choice of projections. One approach is to tailor a dedicated projection set for unimodal distributions and a dedicated projection set for multimodal distributions. In this paper we show that given the knowledge that the distribution is unimodal, principal component axes are an effective choice, while in case of multimodal data, the raw axes are better. However, since we are in an unsupervised setting, this approach requires a sufficiently accurate unimodality statistical test. We considered three different established tests, but they were not sufficiently accurate at detecting if distributions are unimodal or multimodal. We therefore present two new methods, which also follow this framework and obtain strong performance without requiring prior knowledge of the data distribution.

Our first method, Chimera[1], detects anomalies by probability estimation after projection to the principal and raw axes. It provides a fast, robust and accurate method for detecting anomalies in tabular data, and offers the advantage of making very few requirements on the data statistics. After projecting the data onto principal and original axes, the marginal distribution of each axis can be estimated using a non-parametric histogram estimator. Lastly, the estimated log probabilities from all projections are summed and the anomaly score is determined.

Although Chimera obtains strong results, some of the projections that it uses actually degrade the performance. We present our second method, S-Chimera, that select a subset of Chimera's set of projections (original and principal axes). S-Chimera uses the idea that good projections make similar predictions (and therefore agree with the consensus) while poor projections result in random prediction. S-Chimera requires having both normal and anomalous samples in the training set, but it does not require labels indicating which samples are normal and which are anomalous. At inference time, both Chimera and S-Chimera first project the sample on their projection set, and generate the anomaly score by summing the log-probability of the sample according to the individual projections. In our numerical experiments, it will be shown that for high-dimensional data, S-Chimera achieves superior results than more complex and much slower approaches such as Isolation Forest, KNN ,Mahalanobis and LODA, as well as recent deep learning-based approaches.

Our methods are simpler and more robust than previous methods, and achieve stronger results with nearly the fastest runtime. We provide extensive analysis, explaining the differences between Chimera, S-Chimera, and other top methods.

.

1.1 Related Work

In this section, we give a high-level overview of tabular anomaly detection methods. A deeper explanation of the most relevant methods, is provided in Sec. 2.

Non-parametric methods: A large class of methods attempts to estimate the probability distribution of data without making parametric assumptions. One dominant line of work is based on K nearest-neighbor (kNN) e.g. [3, 12, 18]. As kNN can be slow, speeding it up by subsampling the training set was proposed by [26]. In some datasets, the anomalies do not lie in regions that are of low-density in relation to all data, but only in relation to their local neighborhoods. This motivates local kNN methods such as LOF [7]. In extensive comparisons by Goldstein and Uchida [10] it was found that kNN outperforms LOF on most datasets, but there are cases where LOF helps. Kernel Density Estimators (KDE) [21] are related to kNN but results in explicitly probabilistic outputs. KDE notoriously suffers from the curse of dimensionality, and therefore has issues for high-dimensional data.

Histogram-based methods: Histograms are another way to estimate probability distributions. As histograms scale exponentially with the data dimension, they can only be used in low dimensions. HBOS [9] proposed to independently learn a histogram estimators for the marginal distribution when projected to every axis. LODA [17] further extended this idea to learn the probability density for a set of random projection directions. Histogram-based methods are efficient and achieve surprisingly good results. Our method significantly improves over such methods. We will give a more in-depth introduction to histogram-based methods in the Sec. 2.

PCA-based methods: Principal component analysis is a commonly used for learning the main directions of variation. A line of anomaly detectors utilize this technique, and use the Mahalanobis distance, which can be seen as a measure of the Euclidean distance between samples after projection to the principal axes. Shyu et al. [14] identify the contribution of each group of components, and further propose to employ two functions of principal component scores - from the major and from the minor components, such that the major component are utilized to detect extreme observations with large values on some original features, while the minor components help to detect the observations that do not conform to the normal correlation structure. PCA, however, has a major limitation, by assuming that the mean and the variance entirely describe the probability distribution. This requires the probability distribution of the data to be a Gaussian.

Tree-based methods: Random forests [6] and boosted trees [28] are very successful for classifying tabular data. Isolation forest (IF) [15] is a leading method for anomaly detection. The method consists of a set of isolation trees, each tree classifies the sample into a progressive set of nodes. Some nodes have high density while others have low density. Samples that are classified into low density nodes that occurs after only a few splits are considered anomalous. IF is an ensemble method aggregating the results over many randomly generated trees.

Classification-based methods: Instead of attempting to estimate the probability density of normal data, classification methods attempt to directly learn a classifier that discriminates between normal and anomalous data. This is hard as we have no labeled examples. One class classification (OCC) tackles this task by learning a manifold which compactly contains the data. Methods include One-Class SVM [24] and SVDD [27]. Such methods have been extended to learning deep classifiers by Deep SVDD [22]. GOAD [5] was recently proposed to extend the method of Golan and El-Yaniv [8] to tabular data. Deep learning methods are much more computationally expensive and contain many hyper-parameters in comparison to the previously mentioned methods.

[1] A chimera is a single organism that's made up of cells from two or more "individuals"

Consensus methods: When several anomaly classifiers are available, it is often possible to achieve better classification performance than naively averaging their anomaly scores. Ensemble methods attempt to optimize the combination of classifiers [2]. Rayana et al. [20] presented a consensus-based method for combining classifiers. The final step of our method also uses consensus, and scores the classifiers in a similar manner of [20], but it has a different selection rule and outperforms it.

2 BACKGROUND - PROJECTION-BASED ANOMALY DETECTION

In this section, we present a framework that generalizes projection-based anomaly detection methods. We show that popular existing methods correspond to special-cases of this framework. In Sec. 3, we will present effective new methods that obtain strong performance without requiring prior knowledge of the data distribution.

Let our dataset $\mathcal{X}_{train}$ consist of N training samples. We denote individual samples x, each sample is a D dimensional vector $x \in \mathbb{R}^D$. Most samples in our training set are normal while significantly fewer samples are anomalous. We define the probability density function of data as $p(x) = p(x^1, x^2...x^D)$ (x^d denotes the d^{th} dimension of x). The label of each sample (normal or anomalous) is given by labelling function $L(x)$, however this function is **not** known to us, as the setting is unsupervised. At test time, we are presented with a new sample x_{test}, our objective is to predict the correct label $L(x_{test})$.

Although it is difficult to estimate $p(x)$ directly, projection-based methods estimate 1D marginals of $p(x)$. For projections by unit vectors, having a small marginal probability of a sample x implies that $p(x)$ is also small, although the converse is not necessarily true. We identify a framework which generalizes many popular projection-based methods. This framework consists of four main parts:

(I) Composing a set of linear projections to 1D: The first step is to select a set of projection directions (parametrized by unit vectors) $w_1, w_2..w_P$. In this paper we show that the choice of projections directions is the key. By understanding the contribution of each projection, we can leverage and optimize the selected projection set. For example, if the set of projections is generated randomly - accuracy can be improved by increasing the number of projections (as suggested by LODA) since each projection is a very weak learner. In another example, in the case where the data distribution is multimodal, projecting onto the principal components is may not capture the most discriminative directions.

(II) Projecting the data: The projection vectors $w_1, w_2..w_P$ project the original features x onto their respective projected scalars $z_1, z_2..z_P$.

$$z_i = w_i \cdot x \tag{1}$$

(III) Estimating the 1D marginal probability distribution:

The third part, estimates the probability distribution of each 1D marginal z_i. One approach is to employ a one-dimensional histograms as non-parametric estimators for each probability density function p_i. The bins of the histogram can be equispaced or quantile-based. Another approach is to assume a parametric distribution e.g. estimating the marginal distribution by a 1D Gaussian.

(IV) Combining the marginal probabilities: The probabilities estimated in part III are integrated together to a unified anomaly score, where higher score indicates a more anomalous sample. A common approach is to select all the projections and average their contributions with equal weighting. Under restrictive assumptions, their product recovers $p(x)$. It has been shown that even when the variables are not perfectly independent, the naive-Bayes combination can still be effective [29]. Therefore, the anomaly score function is defined in the following way:

$$score = -\sum_i \log(p_i(z_i)) \tag{2}$$

In practice, however, not all projection directions are equally informative for anomaly detection. Therefore, a more general approach gives different weights to different projections reflecting their relative importance. For example, [14] suggest a weighting function for principal components projections, such that the eigenvectors with small eigenvalues are assigned greater importance.

In this work we propose an additional approach. We hypothesise that an optimal selection of the projection directions may significantly boost anomaly detection performance over using all directions. Moreover, it can dramatically reduce the runtime of projection-based methods. We can interpret this approach as a hard weighting ($w_i \in \{0, 1\}$). We will investigate this hypothesis and propose an unsupervised method for projection selection in Sec. 3.3.

We finally summarize the entire framework. We denote the individual projection vectors $w_1, w_2..w_P$, and the weights of the projections $\alpha_1, ...\alpha_P$. The anomaly score is therefore:

$$score(x) = -\sum_i \alpha_i \log(p_i(w_i \cdot x)) \tag{3}$$

2.1 Framework Generalizes Popular Methods

We show that several popular anomaly detection methods can be described as special cases of this framework:

HBOS: When the projection directions are along the standard basis vectors $e_1, e_2..e_N$, the marginal distributions are estimated by 1D histograms, and all $\alpha_i = 1$, Eq. 3 recovers HBOS. Although HBOS achieves surprisingly strong results for such a simple method, it typically does not outperform the state-of-the-art. Also, as the projection directions are not adaptive to the data, they do not model particular axes of variation, reducing the potential discriminative ability. We will show in Sec. 4.3, that standard basis vector projections are however effective when the normal data has a multimodal distribution. Such projections are used as a part of our method.

LODA: When the projection directions are along randomly sampled directions from distribution $\mathbb{N}(0, I)$ s.t. $w \sim \mathbb{N}(0, I)$, all $\alpha_i = 1$ and the marginal distributions are estimated by 1D histograms, LODA is recovered. The authors justify choosing random projections by stating that at the limit of very high-dimensions, random projections approach PCA projections. In practice, many projections are required for achieving strong performance, with correspondingly slow runtime. Furthermore, the main conceptual issue with LODA is that at a finite number of dimensions, different random projections may not in fact be independent and more importantly, they may not correspond to those of particularly significant variation, meaning the directions may not be as informative as the directions of the principal components.

Mahalanobis: Let W be the matrix computed by PCA, whose column vectors are the eigenvectors of the covariance of the training data. An anomaly detector by the Mahalanobis distance is obtained from Eq. 3 by choosing the columns of W as the projections, and estimating the marginal distribution by a univariate Gaussian distribution for each projection direction. As mentioned previously, there exist versions of the Mahalanobis method in which each component is weighted according the proportion of explained variance (but we did not observe significant differences in performance for the unweighted method). The Mahalanobis method makes the assumption the data is distributed as a multivariate Gaussian. In practice, we observed that relying only on the principal components is not sufficient when the normal data follows a multimodal distribution.

3 METHOD

Our methods, **Chimera** and **S-Chimera** follow the framework described in Sec. 2, and provide a generic solution for the common case where the normal and the anomaly distributions are not known apriori.

This section is organized as follows. We first claim that the number of modes in the distribution of the normal data affects the optimal projection set. We further propose a strategy for determining a strong set of projections given the knowledge of whether a distribution is unimodal or multimodal. Next, we relax this assumption by presenting a new method, Chimera, which obtains strong performance without requiring prior knowledge of the data distribution. Lastly, we present S-Chimera, which extends Chimera with a consensus-based selection procedure, resulting in an even faster and more accurate method.

3.1 The Choice of Optimal Projections Depends on the Data Multimodality

In this section, we argue that the multimodality of the distribution of normal data is critical to the choice of projections. Let us first consider the case where the data follows a unimodal distribution. In the absence of knowledge of the distribution of anomalies, and assuming the data are formed linearly by a set of independent factors, projecting the data to these components would form a very strong prior for the directions along which anomalies would be easiest to discriminate. On the other hand such linear models would be insufficient in the case where the data distribution has multiple modes. A weaker, but none-the-less powerful prior is projecting onto the raw axes. It is often the case that anomalies are created when the value of a particular feature is anomalous while the other feature values are all normal. This prior in not data-driven and may be used for multimodal data.

We therefore propose a simple, but surprisingly effective baseline. We first determine if our dataset is unimodal or multimodal. Note that making this determination automatically is not trivial - please see Sec. 3.2 for our solution that avoids this unrealistic requirement. In case it is unimodal, we use PCA projections while if it is multimodal, we project it onto the raw axes. We then estimate the probability density function of each marginal, and compute their unweighted sum as in Eq. 3. Results can be seen in Sec. 4.3, showing that this baseline outperforms many popular methods.

Implementation: To avoid making further parametric assumptions on the distribution of the data, the one-dimensional marginals are estimated by histograms. In practice, we implement the probability density estimators by histograms with 10 fixed-width bins. Our hyper-parameter settings for the histograms are identical to those of HBOS in the PyOD library [30]. It is possible to run a normality test on each marginal, and in the case it indeed follows a normal distribution, replace the estimator by a 1D Gaussian estimator. We do not show this in our experiments, as it did not yield better results.

3.2 Dealing with unknown numbers of modes

The direct consequence of Sec. 3.1 is that given knowledge of whether the data was unimodal or multimodal, an effective deterministic choice of projections can be made directly. In practice the number of modes is not known apriori. A potentially simple mitigation is using established statistical tests for the unimodality of data. Unfortunately, in our experiments, the results of using statistical tests for selecting where to use projections appropriate for unimodal or multimodal distributions have significantly lagged behind doing so using the groundtruth. We first present a simple approach for dealing with this uncertainty, which we name Chimera (the name suggestive of the two headed nature of our approach). Chimera simply uses the concatenation of the projections along the raw and principal axes.

$$score(x) = -\sum_i \log(p_i^{Raw}(x)) - \sum_i \log(p_i^{PCA}(w_i^{PCA} \cdot x)) \quad (4)$$

3.3 Projection Selection by Consensus

The method proposed in Sec. 3.2, has two major drawbacks: i) it increases the number of projections by a factor of 2, increasing runtime by the same amount ii) for each modality, Chimera includes a set of projections that are most suitable (e.g. PCA for unimodal) and another set that is less suitable (e.g. raw axes for unimodal), which has a negative effect on accuracy. We propose to solve this by adding a selection stage that selects, in an unsupervised data-driven way, a subset from the set of projections used by Chimera that is best performing. This is equivalent to the weighting part in Sec. 2 with binary weights.

To investigate whether such weighting holds potential, we first evaluate the maximum performance by selecting the optimal subset of Chimera projections given full supervision. Please note that we only use the supervision here to understand the expressivity of the examined projections, we **do not** use supervision in our main methods or experiments. The supervised procedure is detailed in the SM. We perform this experiment on LODA and on our Chimera projection set, and reached an average ROC-AUC boost of +11% and +10.2% respectively over no selection. This confirms our hypothesis, that selecting a suitable subset of projections, can obtain significant accuracy gains as well as runtime improvements.

Having concluded that careful selection of projections can significantly boost performance, we propose a method for doing so in an **unsupervised** way. Our method is consensus-based, we choose the projections that accord most closely to the consensus. The intuition is that accurate classifiers have a high degree of agreement on which samples are anomalous while noisy classifiers have a strong degree of disagreement over their results. Furthermore, we consider agreement on which samples are anomalous as more significant than agreement on the precise ranking of normal samples, as the true ordering between two anomalies is easier to determine than the ordering between two normal samples.

Our method first computes the consensus anomaly score of each sample in the training set, using Eq. 4. This gives the measure of how anomalous each sample is considered by the average of scores of all projections. We proceed to compute the weighted Pearson correlation between the individual scores according to each projection and those of the consensus. We sort the samples by their score from most anomalous to most normal and record their rank. We denote the rank according to the consensus score as $r_c(x)$. The weighting function gives weight to samples as an inverse function of their rank (according to the consensus scores) $\beta(x) = \frac{1}{r_c(x)}$ - the motivation is giving more importance to consistent scores on anomalous samples rather than on normal samples. See the algorithms for more details.

Our method selects the t% projections that are most correlated to the consensus - and uses the sum of their scores as the final score function. We denote this method S-Chimera. In practice, we found that using $t = 40\%$ achieved the best result while using *fewer* projections than HBOS and Mahalanobis. This can be interpreted as choosing the best of both worlds i.e. raw and principal axes. We reiterate that the difference between Chimera and S-Chimera is that Chimera uses all projections while S-Chimera uses subset of only the t% projections with the highest correlation to the consensus.

Algorithm 1: Chimera: Training Algorithm

```
Input: data samples $\mathcal{X} = \{x_i \in \mathbb{R}^D\}_{i=1}^N$
Output:
histograms $\{h_{e_1}, .., h_{e_D}, h_{pc_1}, ...h_{pc_M}\}$
projection set $\{e_1, , ..e_D, pc_1.., pc_M\}$

1. Perform PCA on $\mathcal{X}$, and extract the $M = min(D, N)$
   principal components:
       $pc_1, ...pc_M = PCA(X)$
2. Estimate histograms:
       $\forall j \in [1..M]\ h_{pc_j} = histogram(X \cdot pc_j)$
       $\forall j \in [1..D]\ h_{e_j} = histogram(X_j^T)$
3. Return $\{h_{e_1}, .., h_{e_D}, h_{pc_1}, ...h_{pc_M}\}$ and
   $\{e_1, , ..e_D, pc_1.., pc_M\}$
```

4 EXPERIMENTS

We conducted an extensive evaluation, comparing our method to well-established state-of-the-art methods for anomaly detection on tabular data. We also conducted ablations, validating the importance of the different components of our method.

Algorithm 2: S-Chimera: Training Algorithm

```
Input:
data samples $X = \{x_i \in \mathbb{R}^D\}_{i=1}^N$
the ratio of selected projections $t$
Output:
histograms $\{h_{s_1}, ...h_{s_g}\} \subset \{h_{e_1}, .., h_{e_D}, h_{pc_1}, ...h_{pc_M}\}$
projection set $\{w_{s_1}, ...w_{s_g}\} \subset \{e_1, , ..e_D, pc_1.., pc_M\}$

1. Train Chimera and obtain the learned histograms and
   projections:
       $\{h_k\}_{k=1}^{D+M}, \{w_k\}_{k=1}^{D+M} = Chimera().train(X)$
2. Calculate $c$, the consensus score per sample:
       $\forall i \in [1..N]\ c_i = -\frac{1}{D+M} \sum_{k=1}^{D+M} log(p_{h_k}(x_i))$
3. Rank the samples, such that $r_1$ is the most anomalous
   point by the consensus, and $r_N$ is the most normal,
   $r_1 < r_N$:
       $r = rank(c)$
4. Calculate each classifier score, $corr_k$, by a weighted
   Pearson correlation with the consensus, $weights = \frac{1}{r}$:
       $\forall k \in [1..D+M]\ corr_k =$
         $WPearson((-log(p_{h_k}(X \cdot w_k)), c), \frac{1}{r})$
5. Sort the scores in descending order:
       $corr_{s_1}, ...corr_{s_{D+M}} = sort(corr_1, ...corr_{D+M})$
6. Select the first $g = t \cdot (D + M)$ of the sorted projections.
7. Return $\{w_{s_1}, ...w_{s_g}\}$ and their corresponded histograms
   $\{h_{s_1}, ...h_{s_g}\}$
```

4.1 Experimental Settings

Datasets: Since we focus on high-dimensional data, we evaluated our method using the high-dimensional real world datasets from the ODDS library[2][19]. These datasets have more than 20 dimensions, and cover a broad spectrum of sizes, dimensionality, and anomaly ratios (see Tab. 1), as well as equal numbers of unimodal and multimodal datasets.

Baseline methods: We compared the performance of our methods with a wide selection of state-of-the-art baseline methods. The methods were selected due to their strong performance, popularity and relatedness to our method. The baseline methods that we compared against are: HBOS [9], LODA [17] (using 100 or 1000 random projections), Mahalanobis [14], Isolation Forest [15] (using 100 or 1000 isolation trees), KNN [3, 18] (with $k = 5$), and OCSVM[23]. All the methods were described in Sec. 1.1 and Sec. 2. We used the implementation by the PyOD library [30] for all these methods, except from Mahalanobis that was implemented by scikit-learn [16] Empirical Covariance implementation, and LODA that was implement as part of our projections pipeline. For the evaluation process, we used the PyOD benchmark evaluation procedure [3].

Evaluation metrics: Following the standard practice in the field, we evaluated the different methods according to: i) the area

[2] http://odds.cs.stonybrook.edu/

[3] https://github.com/yzhao062/pyod/blob/master/notebooks/benchmark.py

under the receiver operating characteristic curve (ROC-AUC) ii) the precision@N_a measures for a test set with N_a anomalies, the percentage of anomalies found within the N_a test samples ranked by the method as most anomalous iii) the inference time. All results were averaged over 10 runs. All metrics were implemented using the PyOD library.

4.2 Comparison with State-of-the-Art

ROC-AUC: We compare the average ROC-AUC of the evaluated methods in Tab. 2. The detailed results of each dataset are presented in the SM. We observe that HBOS and KNN achieve the weakest average performance, while Isolation Forest and Mahalanobis are the strongest baselines. Our method Chimera beats all baselines, while combination with selection (S-Chimera) achieves even better performance with results that are 2% better than the best baseline. In Tab. 2, we also present the average ROC-AUC rank of each method. S-Chimera is the top ranked method, and Chimera is the second best. The difference in Mahalanobis performance across both measures is due to its sensitivity to the multimodality of the data, see Sec. 4.3 for further analysis. We conclude that S-Chimera is the most accurate method across both measures.

Precision@N_a: In Tab. 2 we present results for the precision metric, with OCSVM being the strongest baseline and being competitive with Chimera. S-Chimera significantly boosts performance to 3.4% over the best baseline.

Runtime: The average inference times of each of the evaluated methods is presented in Tab. 2. We can observe that Isolation Forest, one of the strongest baselines, is orders of magnitude slower than our proposed method while achieving lower average accuracy. LODA with 100 projections is faster than Chimera and similar runtime as S-Chimera, but it achieves weaker performance. Additionally, using more random projections for LODA and more random trees for Isolation Forest increases their average ROC-AUC score by around 1% while increasing their runtime by a factor of ten, further increases do not improve performance. The Mahalanobis method has faster inference time than Chimera, and comparable inference time as S-Chimera, while being less accurate than both. S-Chimera has 40% of the number of projections of Chimera (thus faster runtime), while typically being of higher accuracy. Although S-Chimera has fewer projections than HBOS, it is a little slower as some of the components require projection while HBOS uses the raw samples - however its accuracy is much higher. To conclude, our final method, S-Chimera, is much faster than the most accurate methods - while having slightly faster or comparable speed to the faster but much less accurate methods.

4.3 Analysis by Multimodality

Comparison between projection-based methods: We observe in Fig. 2 that HBOS outperforms for multimodal datasets, while Mahalnobis outperforms for unimodal datasets. However, both methods are low ranked for datasets of the wrong modality. On the other hand, Chimera obtains high performance and S-Chimera reaches best or near-best performance for unimodal and multimodal data. This demonstrates the robustness of Chimera and S-Chimera.

Unimodality oracle (Sec. 3.1): When selecting HBOS for multimodal datasets and Mahalanobis for unimodal datasets, according to a unimodality oracle, the obtained average ROC-AUC score is 0.781. S-Chimera, without any supervision or further assumptions - achieves 0.776, and Chimera obtains 0.765. This shows that the gains achievable by perfect unimodality tests over our unsupervised method are limited.

Table 1: Dataset properties

Dataset	#Samples	#Dimensions	% Outlier
speech	3686	400	1.6549
arrhythmia	452	274	14.6018
musk	3062	166	3.1679
mnist	7603	100	9.2069
optdigits	5216	64	2.8758
heart	267	44	20.5993
satellite	6435	36	31.6395
satimage-2	5803	36	1.2235
ionosphere	351	33	35.8974
letter	1600	32	6.25
wbc	378	30	5.5556
cardio	1831	21	9.6122

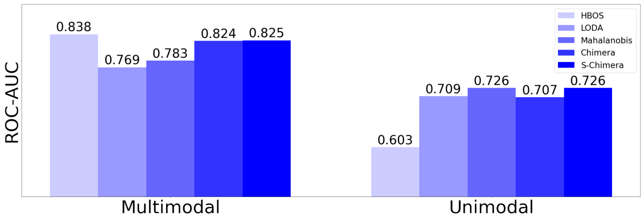

Figure 2: Comparison between projection-based methods - by unimodality and multimodality

Unsupervised unimodality tests: We evaluated the efficacy of unsupervised statistical tests for determining if datasets are unimodal or multimodal. We examined three established methods: the Dip test on all pair-distances and on PCA [1, 13] and the Folding test [25]. We tested the unimodality classification accuracy on our datasets of Tab. 1. The dip-PCA test reached the highest accuracy of 83.3%, while the two other two tests scores had 58.3% accuracy. Although using PCA/HBOS projection given the groundtruth multimodality of the data reaches an average 0.781 ROC-AUC over our datasets, using the best unsupervised test obtained a much lower 0.746. This is significantly lower than our method S-Chimera. This suggests that our method is more robust than relying on statistical unimodality tests.

4.4 Analysis of One-Dimensional Estimators

We compared the effect of estimating all the 1D marginals by: i) Gaussians ii) histograms iii) a hybrid approach in where principal component marginals are estimated by Gaussians and raw axes marginals by histograms. The results are reported in Table 3. We can observe that the hybrid and the histogram estimators outperform the Gaussian estimators, while the histogram estimators are a bit

Table 2: Performance comparison on ODDS datasets - S-Chimera is fast and accurate.

	HBOS	LODA(100/1000)	IForest(100/1000)	Mahalanobis	KNN	OCSVM	Chimera	S-Chimera
ROC-AUC score	0.720	0.739/0.745	0.756/0.759	0.754	0.716	0.740	*0.765*	**0.776**
ROC-AUC rank	5.00	4.58	4.42	4.83	5.50	4.58	3.83	**3.17**
Precision@N_a	0.383	0.418/0.432	0.416/0.422	0.381	0.340	0.436	*0.437*	**0.470**
Runtime(ms)	**3.95**	*4.65*/43.44	89.79/906.34	5.70	633.56	219.90	10.39	*4.07*

Table 3: 1D Estimator Comparison: Histogram estimators are better than Gaussian estimators

Gaussians	Hybrid	Histograms (Chimera)
0.752	0.763	**0.765**

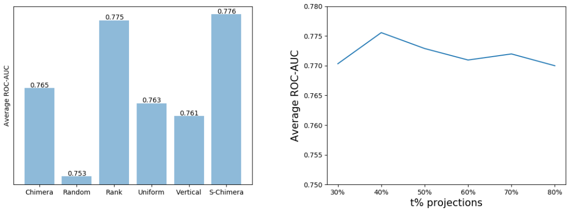

Figure 3: (left) Ablation of our selection method (right) Comparison of accuracy as a percentage of the number of selected classifiers ($t\%$).

better than hybrid. This confirms that histograms are a robust choice for estimation of the 1D marginals.

4.5 Analysis of Our Selection Method

In Sec. 3.3, we presented a new method for selecting projections. Although our method is simpler than previously proposed consensus-based selection methods, it performs better.

Comparison to random selection: We first test if our selection method improves over random selection of Chimera projections. The results are presented in Fig. 3. It is clear that random selection achieves worse results than no selection at all, while our selection method improves over all other methods.

Comparison to vertical selection. Rayana et al. [20] proposed a method for selecting a subset of an ensemble of anomaly detectors. We evaluate their "vertical selection" approach, which is similar to ours but has a much more complex scoring and selection rule (please see their paper for more details). We evaluate the average ROC-AUC of their greedy-selection approach against ours in Fig. 3 (left - "vertical" bar). Their approach is slightly less accurate than not selecting, while our method is better than both selecting greedily and not selecting.

S-Chimera scoring function ablations: We ablated the effect of assigning higher weights to anomalies in the correlation score and found that it significantly improves ROC-AUC over uniformly weighting all samples (see Fig. 3). Using non-linear rank correlation does not improve the performance over Pearson correlation.

Determining the ratio of selected projections ($t\%$): We explored different ratios for determining the percentage of projections most similar to the consensus that are selected by S-Chimera. We present results for a range of retained component percentage values in Fig. 3. It can be observed that 40% is optimal but deviations from this value do not significantly reduce performance.

4.6 Comparison to Deep Learning Methods

Over the last several years, deep learning methods have been advanced as a promising direction for detecting anomalies in tabular data. Therefore, we also compared our performance to the current state-of-the-art deep learning methods - GOAD [5] and DROCC [11]. These methods were mostly evaluated in the semi-supervised setting, in which the training set contains only normal data points, while the test set contains both normal and anomalous examples. We therefore evaluated our methods with respect to these methods in both of the settings - unsupervised and semi-supervised.

Unsupervised settings We compared the performance against the main evaluation protocol used in this paper (unsupervised setting, 12 highest-dimensional ODDS datasets). For the implementation details please see SM. The results are reported in Tab 4. We can see that our simple and fast method outperforms the state-of-the-art deep methods, both in term of the average score and ranking.

Semi-supervised settings We also evaluated the performance using the exact protocol used in GOAD (4 datasets, no anomalies in training dataset). Please see the implementation details in the SM. Since the selection process assumes anomalies in the training data, S-Chimera is not relevant in this setting that only contains normal data and we therefore used Chimera (without selection). The ROC-AUC and F1 scores are reported in Tab. 4, our method outperforms the state-of-the-art deep methods in this setting. Note that the F1 scores of DROCC vary from those reported in the paper as they computed F1 differently from the standard practice. Their protocol reverses the labels of normal and anomalous data, significantly affecting F1 numbers (but not ROC-AUC). We evaluated all methods with exactly the same protocol.

4.7 Combining Random Projections

We explored the effect of enriching the projection set with random rotations. Fig 4 presents the results. We can see that enriching the raw axes with random rotations enhances its performance, but by a lower margin than our method of combining the raw with the PCA projections (our method scores 2% better than raw+random). Combining random rotations with the principal component projections improves by 1% over projecting only on the PCs, however, Chimera

Table 4: Comparison to deep learning methods: i) on the 4 commonly reported datasets in the tabular deep learning literature (with normal-only training data) ii) on the main (unsupervised) evaluation protocol used in this paper.

	ROC-AUC / F1 Scores		
Dataset	**GOAD**	**DROCC**	**Chimera**
arrhythmia	0.763 / 0.521	0.564 / 0.409	**0.802** / **0.591**
thyroid	0.958 / 0.744	0.975 / 0.720	**0.990** / **0.785**
kddrev	0.994 / 0.984	0.980 / 0.955	**0.996** / **0.989**
kdd	**0.996** / **0.989**	0.718 / 0.807	**0.996** / 0.983

	ROC-AUC Score		
Dataset	**GOAD**	**DROCC**	**S-Chimera**
Average score	0.731	0.623	**0.776**
Average rank	2.167	2.167	**1.667**

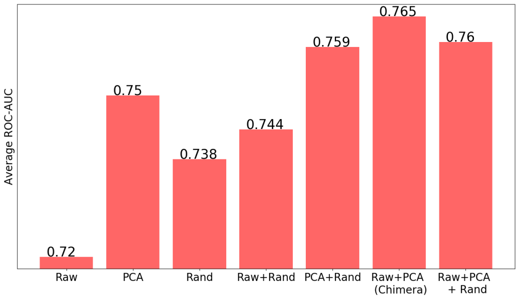

Figure 4: Comparison between projection sets

Table 5: Effect of different projections on top baselines

IF			LODA	
Original	**PCA**	**Raw+PCA**	**Original**	**LODA+Chimera**
0.756	0.749	0.759	0.739	0.755

boosts it by 1.5%. In addition, combining random rotations with Chimera reduces the performance.

4.8 Effect of Raw and PCA axes on other methods

In Tab. 5, we evaluate replacing the original inputs in our top baseline Isolation Forest by its projections to the principal axes as well as its concatenation of raw and PCA projected features. We can see that the combination of raw and PCA improves the average ROC-AUC. Furthermore, we see that adding Chimera to the LODA projections improves LODA, but does not reach the performance of Chimera. We conclude that enriching baseline methods with information from more axes has a potential to improve the original methods, but does not outperform Chimera.

5 DISCUSSION

In this section we discuss some of the wider consequences of this work and explain its limitations.

Optimal projections by distribution unimodality: We demonstrated that given the knowledge of whether a distribution is unimodal or multimodal, an effective strategy can be devised for determining a strong set of projections. Unfortunately, the statistical unimodality tests that we evaluated did not achieve the required level of accuracy. S-Chimera presented another approach, selecting the best projections from the combined set. Future research should consider developing more accurate selection criteria. On the other hand, research on the robustness of statistical unimodality tests is promising.

Deep and shallow methods: Over the past several years, deep learning methods have been advanced as a promising direction for detecting anomalies in tabular data. The disadvantage of deep learning methods is their heavy computational requirements and slow runtime (both for training and inference), as well as the need for large training datasets. In this paper, we investigated an alternative method which is simple, computationally cheap and easy to interpret. We found that our method convincingly outperformed deep learning methods on commonly reported high-dimensional benchmarks. Categorizing the types of tabular data that are most suitable for shallow learning and those that are most suitable for deep learning is in our mind an exciting avenue for future research.

Low-dimensional data: Isolation Forest captures non-linear dependencies between variables, which is one of our methods limitation. As [12] stated, Isolation Forest is efficient especially for low-dimensional data, since in high dimensions there is a high probability that a large number of features are neglected in the process. In addition, the relatively weak and low-rank dependencies between variables in tabular data might become more dominant in low-dimensional data. We evaluated the performance of Isolation Forest on some low-dimensional ODDs datasets, and indeed, it provided more accurate results than Chimera. The success of Isolation Forest on low-dimensional data indicates that focusing research of complex non-linear methods (e.g. trees and deep neural networks) on low-dimensional datasets is likely to yield promising results.

Semi-supervised vs. unsupervised settings: The semi-supervised setting is characterized by a training set that consists of only normal data, while the unsupervised setting assumes that the data contains an unknown rate of anomalous points. kNN is the leading method in the semi-supervised settings, but was not among the top methods when anomalies were present in the training set. In addition, we found that the sensitivity of Mahalanobis to the multimodality of the normal data is reduced in the semi-supervised settings.

The potential of selection methods: In Sec. 3.3, we showed that when selecting the best projection axes using supervised data can boost the performance dramatically. Unsupervised selection by our method improved the score over no selection and closed some of the gap. The large performance gap remaining between unsupervised and supervised projection selection, demonstrates the potential of improving unsupervised selection techniques. In addition, a similar selection procedure can improve other ensemble methods such as Isolation Forest.

Limitations of our methods: Although our methods achieve state-of-the-art results on high-dimensional tabular anomaly detection, there are several limitations: i) our selection method relies on the availability of anomalies in the training set (although we *do not* assume that we know their labels) - we do not expect it to select good projections in the case where no anomalies are available in the training set. In such case we suggest defaulting to Chimera without selection. ii) low-dimensional data can reduce the need-for and effectiveness-of projection selection. iii) our methods assume that the dependence between variables are linear - for data with strong high-order correlations (e.g. images, text or audio), we do not expect it to work well. In order to detect anomalies in such data, it would first be necessary to extract strong features e.g. by a pre-trained deep neural network and then run our method on the deep features (see DN2 [4] for a similar idea). iv) our methods are not designed for categorical data - finding sensible projection directions for categorical data would depend on the precise encoding method and would require more care - this is left for future work.

6 CONCLUSIONS

We presented two new methods, Chimera and S-Chimera, for detecting anomalies based on the estimation of the marginal probability distributions of the raw and principal axes of the data. S-Chimera selects the axes that are most beneficial for anomaly detection - both in terms of accuracy and efficiency. Extensive experiments showed that our method is both faster and more accurate than the state-of-the-art.

7 ACKNOWLEDGMENTS

This work was partly supported by the Federmann Cyber Security Research Center in conjunction with the Israel National Cyber Directorate.

REFERENCES

[1] Andreas Adolfsson, Margareta Ackerman, and Naomi C. Brownstein. 2019. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognit.* 88 (2019), 13–26. https://doi.org/10.1016/j.patcog.2018.10.026

[2] Charu C. Aggarwal and Saket Sathe. 2015. Theoretical Foundations and Algorithms for Outlier Ensembles. *SIGKDD Explor.* 17, 1 (2015), 24–47. https://doi.org/10.1145/2830544.2830549

[3] Fabrizio Angiulli and Clara Pizzuti. 2002. Fast Outlier Detection in High Dimensional Spaces. In *Principles of Data Mining and Knowledge Discovery, 6th European Conference, PKDD 2002, Helsinki, Finland, August 19-23, 2002, Proceedings (Lecture Notes in Computer Science, Vol. 2431)*, Tapio Elomaa, Heikki Mannila, and Hannu Toivonen (Eds.). Springer, 15–26. https://doi.org/10.1007/3-540-45681-3_2

[4] Liron Bergman, Niv Cohen, and Yedid Hoshen. 2020. Deep Nearest Neighbor Anomaly Detection. *CoRR* abs/2002.10445 (2020). arXiv:2002.10445 https://arxiv.org/abs/2002.10445

[5] Liron Bergman and Yedid Hoshen. 2020. Classification-Based Anomaly Detection for General Data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net. https://openreview.net/forum?id=H1lK_lBtvS

[6] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[7] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *ACM sigmod record*, Vol. 29. ACM, 93–104.

[8] Izhak Golan and Ran El-Yaniv. 2018. Deep Anomaly Detection Using Geometric Transformations. In *NeurIPS*.

[9] Markus Goldstein and Andreas Dengel. [n.d.]. Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm.

[10] Markus Goldstein and Seiichi Uchida. 2016. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS One* 11, 4 (April 2016). https://doi.org/10.1371/journal.pone.0152173

[11] Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. 2020. DROCC: Deep Robust One-Class Classification. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 3711–3721. http://proceedings.mlr.press/v119/goyal20c.html

[12] Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. 2019. Statistical Analysis of Nearest Neighbor Methods for Anomaly Detection. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 10921–10931. https://proceedings.neurips.cc/paper/2019/hash/805163a0f0f128e473726ccda5f91bac-Abstract.html

[13] J. A. Hartigan and P. M. Hartigan. 1985. The Dip Test of Unimodality. *Ann. Statist.* 13, 1 (03 1985), 70–84. https://doi.org/10.1214/aos/1176346577

[14] Mei ling Shyu, Shu ching Chen, Kanoksri Sarinnapakorn, and Liwu Chang. 2003. A novel anomaly detection scheme based on principal component classifier. In *in Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM'03*. 172–179.

[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy.* IEEE Computer Society, 413–422. https://doi.org/10.1109/ICDM.2008.17

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[17] Tomás Pevný. 2016. Loda: Lightweight on-line detector of anomalies. *Mach. Learn.* 102, 2 (2016), 275–304. https://doi.org/10.1007/s10994-015-5521-0

[18] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. 2000. Efficient Algorithms for Mining Outliers from Large Data Sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein (Eds.). ACM, 427–438. https://doi.org/10.1145/342009.335437

[19] Shebuti Rayana. 2016. ODDS Library. http://odds.cs.stonybrook.edu

[20] Shebuti Rayana and Leman Akoglu. 2016. Less is More: Building Selective Anomaly Ensembles. *ACM Trans. Knowl. Discov. Data* 10, 4 (2016), 42:1–42:33. https://doi.org/10.1145/2890508

[21] Murray Rosenblatt. 1956. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics* (1956), 832–837.

[22] Lukas Ruff, Nico Gornitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *ICML*.

[23] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. 2001. Estimating the Support of a High-Dimensional Distribution. *Neural Comput.* 13, 7 (2001), 1443–1471. https://doi.org/10.1162/089976601750264965

[24] Bernhard Scholkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. 2000. Support vector method for novelty detection. In *NIPS*.

[25] Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. 2018. Are your data gathered? The Folding Test of Unimodality. In *KDD 2018 - 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Minin.* London, United Kingdom, 2210–2218. https://doi.org/10.1145/3219819.3219994

[26] Mahito Sugiyama and Karsten Borgwardt. 2013. Rapid distance-based outlier detection via sampling. *Advances in Neural Information Processing Systems* 26 (2013), 467–475.

[27] David MJ Tax and Robert PW Duin. 2004. Support vector data description. *Machine learning* 54, 1 (2004), 45–66.

[28] Terry Windeatt and Gholamreza Ardeshir. 2002. Boosted Tree Ensembles for Solving Multiclass Problems. In *Multiple Classifier Systems, Third International Workshop, MCS 2002, Cagliari, Italy, June 24-26, 2002, Proceedings (Lecture Notes in Computer Science, Vol. 2364)*, Fabio Roli and Josef Kittler (Eds.). Springer, 42–51. https://doi.org/10.1007/3-540-45428-4_4

[29] Harry Zhang. 2004. The Optimality of Naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA*, Valerie Barr and Zdravko Markov (Eds.). AAAI Press, 562–567. http://www.aaai.org/Library/FLAIRS/2004/flairs04-097.php

[30] Yue Zhao, Zain Nasrullah, and Zheng Li. 2019. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research* 20, 96 (2019), 1–7. http://jmlr.org/papers/v20/19-011.html