Quantifying Statistical Significance of Deep Nearest Neighbor Anomaly Detection via Selective Inference

Mizuki Niihori[1] **Shuichi Nishino**[1,2] **Teruyuki Katsuoka**[1] **Tomohiro Shiraishi**[1,2]
Kouichi Taji[1] **Ichiro Takeuchi**[1,2]
[1]**Nagoya University** [2]**RIKEN**
niihori.mizuki.nagoyaml@gmail.com
takeuchi.ichiro.n6@f.mail.nagoya-u.ac.jp

Abstract

In real-world applications, anomaly detection (AD) often operates without access to anomalous data, necessitating semi-supervised methods that rely solely on normal data. Among these methods, deep k-nearest neighbor (deep kNN) AD stands out for its interpretability and flexibility, leveraging distance-based scoring in deep latent spaces. Despite its strong performance, deep kNN lacks a mechanism to quantify uncertaintyan essential feature for critical applications such as industrial inspection. To address this limitation, we propose a statistical framework that quantifies the significance of detected anomalies in the form of p-values, thereby enabling control over false positive rates at a user-specified significance level (e.g.,0.05). A central challenge lies in managing selection bias, which we tackle using Selective Inferencea principled method for conducting inference conditioned on data-driven selections. We evaluate our method on diverse datasets and demonstrate that it provides reliable AD well-suited for industrial use cases.

1 Introduction

In many practical anomaly detection (AD) problems, anomalous data is often unavailable in advance; therefore, AD algorithms must be developed using only normal dataa setting known as semi-supervised AD [1–3]. Among various semi-supervised AD methods, we focus on the k-nearest neighbor (kNN) approach [4]. The kNN approach is simple yet effective, offering flexibility, minimal assumptions about the data, and adaptability to different distance metrics. Especially, by applying kNN in a latent feature space identified by deep learning models, anomalies can be detected more flexibly using task-specific distance metricsan approach we refer to as *deep kNN*-based AD. Deep kNN-based AD combines the power of deep learning models with interpretable kNN scoring in latent space, making it a promising method for practical applications [5].

However, its dependence on complex deep learning-based detection procedures means that no established method currently exists for rigorously quantifying uncertainty or ensuring reliability. This limitation is particularly critical in high-stakes applications such as industrial inspection, where reliable uncertainty estimation is essential. Despite the interpretability of deep kNN-based scoring, it lacks a principled way to assess how confidently a test case deviates from normal cases. Therefore, developing a statistical framework that complements the high detection accuracy of deep kNN with uncertainty quantification is a crucial step toward practical and reliable deployment in real-world practical applications.

To address this issue, we propose a method that can quantify the statistical significance of the results obtained by deep kNN-based ADs. Specifically, we formulate the kNN-based AD as a statistical hypothesis testing problem and develop a method to quantify the statistical significance of detected anomalies in the form of p-values. The p-value of a detected anomaly represents the probability that the anomaly is a false positive. By making decisions based on anomalies with p-values below

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

a certain threshold (e.g., 0.05), we can ensure that the probability of erroneous decisions remains below the significance level of 5%. In this study, we refer to the proposed testing method as the *deep-kNN test*. From a practical perspective, the ability to explicitly control the false positive rate is highly beneficial in real-world anomaly detection. In safety-critical applications such as medical diagnosis or industrial inspection, false positives often lead to unnecessary interventions, costs, or workflow disruptions. By providing a statistically significant upper bound on the false positive rate, our framework enables practitioners to make anomaly detection decisions with a quantifiable level of reliability.

Accurately quantifying the statistical significance of anomalies identified by deep kNN is a non-trivial challenging task from the following two perspectives. The first challenge arises from the fact that both AD and testing are performed on the same dataset, leading to selection bias called *double-dipping* [6]. The second challenge is that when anomalies are detected based on the latent feature space of a deep learning model, it becomes necessary to account for the complex process of the deep learning model computation. The core idea of our deep kNN test is to address these two challenges by introducing a statistical testing framework known as *Selective Inference (SI)* [7, 8]. SI has recently gained attention as a method for statistically testing data-driven hypotheses, by conditioning statistical inference on the event that the hypothesis has been selected. Figure 1 demonstrates how the proposed deep kNN test can be applied to industrial visual inspection task.

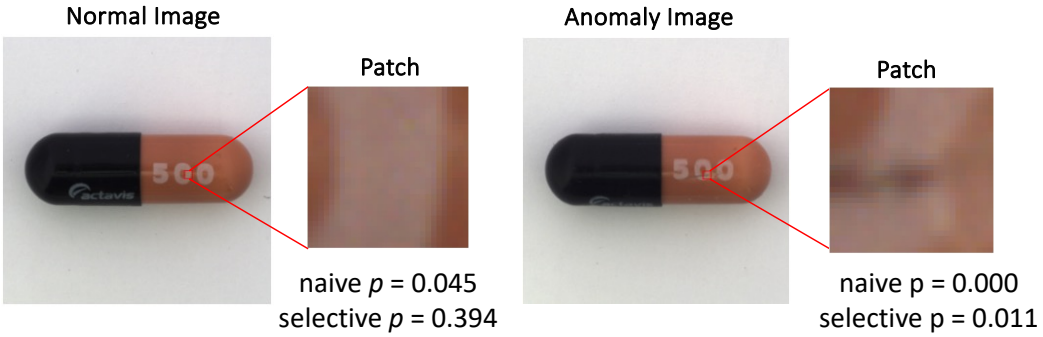


Figure 1: Examples of anomaly patches extracted from *Capsule* images using kNN-based AD are shown (see § 5 for detailed settings). For both the normal image (left) and the anomaly image (right), two types of p-values derived from different statistical tests are presented. The "naive p" represents the p-value obtained using a conventional method, while the "selective p" denotes the p-value computed using the method proposed in this study. At a significance level of $\alpha = 0.05$, the conventional naive p-value for the normal image (left) falls below the threshold, resulting in a false positive detection. In contrast, the proposed selective p-value correctly identifies it as a true negative. For the anomaly image (right), both p-values fall below the threshold, correctly identifying the patch as anomalous (true positive). In this study, we show that that conventional naive p-values are invalid as measures of statistical significance, whereas the proposed selective p-values serve as valid uncertainty measures for assessing the significance of anomalies detected by kNN-based ADs.

Related Works Semi-supervised AD algorithms can generally be divided into three main categories [4, 9, 10]. The first group comprises AD methods based on parametric probabilistic models. These methods assume that normal data follow a specific statistical distribution, such as a multivariate Gaussian. A common traditional technique involves computing the Mahalanobis distance and assigning p-values using the χ^2 distribution, allowing for statistical significance testing. These methods are interpretable and computationally efficient but may perform poorly with complex or non-Gaussian data distributions [11, 12]. The second group includes AD methods based on classical machine learning algorithms. Representative methods include One-Class SVM [13], Isolation Forest [14], and k-Nearest Neighbors (kNN) [15]. While these methods are more flexible than parametric probabilistic model-based methods, they often lack a principled mechanism to quantify statistical significance of the detected anomalies, making it difficult to assess confidence in AD. The third group comprises deep learning-based AD methods, which use neural networks to capture complex patterns in normal data. Representative approaches include autoencoders or variational autoencoders (VAEs) [16], GAN-based method [17], and diffusion-based method [18]. Deep SVDD [1] and deep

kNN-based AD [19] are also widely used as AD methods based on deep learning models. While these deep learning-based approaches often achieve high detection performance, most still lack a well-established framework for quantifying statistical significance, such as assigning p-values.

Selective Inference (SI) -also known as post-selection inference- provides valid statistical inference for data-driven hypotheses. By leveraging the conditional distribution of a test statistic given that a particular hypothesis was selected based on the data, SI corrects the selection bias introduced by *cherry-picking* findings. This ensures that the reported p-values -known as selective p-values- maintain valid false positive rates, even when the same data is used for both hypothesis selection and testing, which would otherwise inflate the error rates due to *double dipping* [7]. Early work in SI primarily focused on feature selection in linear regression. A seminal contribution by Lee et al [8] introduced an exact SI procedure for the Lasso, deriving valid selective p-values for selected regression coefficients by conditioning on the selection event induced by the Lasso solution. Since then, extensive research has extended SI to various other feature selection settings, including marginal screening [20], stepwise feature selection [21], generalized linear models [22], and many others [23–33]. Recent developments have also focused on improving the power of SI methods through new theoretical insights and algorithmic innovations [34–38]. In parallel, SI has been adapted to problems beyond feature selection, finding applications in a range of domains, such as clustering [39–42] and many others [43–47]. In the context of deep learning, SI has been applied to provide statistical inference for segmentation tasks [48], saliency maps such as CAM [49], and attention weights in Vision Transformers [50]. In recent years, this line of research has begun to explore the reliability of various deep learning model components from a statistical inference perspective [51–53]. In contrast, applications of SI in AD settings remain limited. One related work is the application of SI to change-point detection in time series data [54–58]. However, that line of research focuses on testing the significance of global changes in the entire sequence, which differs from our setting where the goal is to assess the abnormality of individual data points. Another related work is the application of SI to robust regression [59], where inference is performed on regression coefficients obtained after excluding outliers. However, the statistical significance of the detected outliers themselves is not addressed. In conclusion, no studies have yet explored the SI framework to quantify the statistical significance of kNN-based AD, and its deep kNN variants remain entirely unexplored, leaving a significant gap in the literature.

Our contributions In this paper, as a proof of concept, we address the problem of quantifying the statistical significance of results produced by a simple deep kNN-based anomaly detection (AD) approach, as studied in [19], where anomalies are identified by thresholding the kNN distance in the latent feature space of a CNN classifier. Unlike prior studies on deep kNN-based AD, our contribution is *not* the development of a new AD algorithm aimed at improving detection accuracy, but rather the introduction of a method to quantify the statistical uncertainty of detected anomalies. Our first contribution is to formulate kNN-based semi-supervised AD as a statistical test within the SI framework, enabling rigorous reliability assessment of detected anomalies. The second contribution is to develop a computational method to incorporate the distance measure derived from a deep learning model into the SI framework. Finally, we validate the effectiveness of the proposed deep-kNN test through experiments on various datasets and industrial inspection scenarios, demonstrating its practical utility and robustness. Furthermore, to facilitate reproducibility and further research, we release an open-source implementation of our proposed method, including code and experimental scripts, available at `https://github.com/Takeuchi-Lab-SI-Group/Quantifying_Statistical_Significance_of_Deep_Nearest_Neighbor_Anomaly_Detection_via_SI`.

2 Problem setup: deep kNN-based anomaly detection

In this section, we describe the problem setting of a simple deep kNN-based AD, nearly identical to that in [19], as a proof of concept for our statistical testing framework. In semi-supervised AD problems, the available training dataset consists only of the set of normal instances. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ represent the set of d-dimensional input feature vectors for n normal training instances, where n is the number of instances. If an instance is an image, for example, d is the number of pixels in the image and $\boldsymbol{x}_i, i \in [n]$ is the d-dimensional vector of pixel values. We assume that a preterined CNN is available, and let us denote its feature representation as $\boldsymbol{\phi}(\boldsymbol{x}_i), i \in [n]$ using a feature extractor $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^D$, where D is the dimension size of the latent feature vectors. We measure the distance

between two input instances $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ in the latent feature space of a pre-trained CNN classifier as

$$\mathrm{dist}_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{x}') := \|\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}')\|_2, \tag{1}$$

where we adopt the L_2 distance within the latent feature space in this study for concreteness, although the choice of distance metric is flexible and discussed further in the §6.

Given a test instance $\boldsymbol{x}^{\mathrm{test}} \in \mathbb{R}^d$, the kNN-based AD is formulated as follows. Consider an order of indices by ascending distance with respect to a distance function $\mathrm{dist}_\phi(\cdot, \cdot)$ such that

$$\mathrm{dist}_\phi(\boldsymbol{x}^{\mathrm{test}}, \boldsymbol{x}_{o(1)}) \leq \mathrm{dist}_\phi(\boldsymbol{x}^{\mathrm{test}}, \boldsymbol{x}_{o(2)}) \leq \cdots \leq \mathrm{dist}_\phi(\boldsymbol{x}^{\mathrm{test}}, \boldsymbol{x}_{o(n)}). \tag{2}$$

Then, $\boldsymbol{x}_{o(k)}$ is called the k^{th} nearest neighbor instance of $\boldsymbol{x}^{\mathrm{test}}$. Since the choice of k affects the distance magnitude, we adopt the following well-known anomaly score [4, 60]:

$$a(\boldsymbol{x}^{\mathrm{test}}) = \log \mathrm{dist}\left(\boldsymbol{x}^{\mathrm{test}}, \boldsymbol{x}_{o(k)}\right) - \frac{\log k}{D}, \tag{3}$$

where the first term represents the log-scale distance, whereas the second term adjusts for the influence of k's selection[1]. In kNN-based AD, if Eq. (3) exceeds a certain threshold θ, the test instance $\boldsymbol{x}^{\mathrm{test}}$ is selected as an anomaly. The threshold θ is typically determined based on the empirical distribution of anomaly scores among normal instances. The choice of k greatly affects the results in kNN-based AD. Users can set k based on domain knowledge or experience. However, when domain knowledge is limited or data is complex, a systematic approach is desirable. In semi-supervised AD, unlike supervised learning such as kNN classification or regression, it is not possible to determine k through data splitting such as cross-validation. One commonly used heuristic to select k is to calculate the anomaly scores for various k values per test instance $\boldsymbol{x}^{\mathrm{test}}$, choosing the k that maximizes this score. Our proposed kNN-test is valid regardless of the method used to determine k.

3 Statistical testing framework for kNN-base AD

In this section, we formulate the kNN-based AD problem as a statistical hypothesis testing problem in order to quantify the statistical significance of the detected anomalies. First, we consider each input feature vector $\boldsymbol{x}_i, i \in [n]$ as a realization of a random vector $\boldsymbol{X}_i, i \in [n]$. The statistical model for the random vector $\boldsymbol{X}_i$ is written as

$$\boldsymbol{X}_i = \boldsymbol{s}_i + \boldsymbol{\varepsilon}_i, \quad i \in [n], \tag{4}$$

where $\boldsymbol{s}_i \in \mathbb{R}^d$ represents the signal component, and $\boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ represents the noise component. In this study, we adopt a *semi-parametric* assumption for the statistical model described in Eq.(4). Specifically, we place no restrictions on the distribution of the signal components $\boldsymbol{s}_i$ for $i \in [n]$, treating them in a completely non-parametric fashion. In contrast, we assume that the noise component follows a Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \sigma^2 I)$, where σ^2 is either known or can be estimated from an independent dataset. This setup differs from traditional AD approaches based on parametric probabilistic models, which assume that the signal components follow a specific parametric distribution. In our semi-parametric framework, the distribution of the signal components is entirely unknown and unrestricted, allowing the model to remain valid even when the signals exhibit complex or multimodal characteristics. For example, if the input feature vectors $\boldsymbol{x}_i$ represent images—each element corresponding to a pixel value—then our model in Eq.(4) allows a set of arbitrary original images each of which is contaminated by Gaussian noise. Our goal is to determine, within statistical hypothesis testing framework, whether an observed anomaly originates from the underlying signal or is merely a consequence of the noise.

Let the feature vector of a test instance be denoted as $\boldsymbol{x}^{\mathrm{test}}$ and its corresponding random version as $\boldsymbol{X}^{\mathrm{test}}$. We assume $\boldsymbol{X}^{\mathrm{test}} = \boldsymbol{s}^{\mathrm{test}} + \boldsymbol{\varepsilon}^{\mathrm{test}}$ in the same way as Eq.(4). In kNN-based AD, the k-nearest instance $\boldsymbol{x}_{o(k)}$ is selected from the n available training instances. Our interest lies in determining whether the signal of the test instance is statistically significantly different from the signal of its k-nearest instance. This problem can be formulated as a hypothesis testing problem with the following null hypothesis H_0 and alternative hypothesis H_1:

$$\mathrm{H}_0 : \boldsymbol{s}^{\mathrm{test}} = \boldsymbol{s}_{o(k)} \quad \text{vs.} \quad \mathrm{H}_1 : \boldsymbol{s}^{\mathrm{test}} \neq \boldsymbol{s}_{o(k)} \tag{5}$$

[1]The choice of Eq. (3) is based on certain assumptions and heuristics in the literature, but its details are beyond the scope of this paper. For further information, refer to [4].

The null hypothesis H_0 states that the true signal of the k^{th} nearest normal training instance equals the true signal of the test instance, while the alternative hypothesis H_1 asserts they are different. By performing a statistical test for these hypotheses, the false detection rate of an anomaly can be quantified using p-values.

As a reasonable test statistic, we consider

$$T\left(\boldsymbol{X}^{\mathrm{test}}, \{\boldsymbol{X}_i\}_{i\in[n]}\right) := \frac{1}{2}\left\|\boldsymbol{X}^{\mathrm{test}} - \boldsymbol{X}_{o(k)}\right\|_2. \tag{6}$$

The p-value is defined as the probability of observing a test statistic greater than or equal to the one actually observed under the null hypothesis H_0, i.e.,

$$p = \mathbb{P}_{\mathrm{H}_0}\left(T\left(\boldsymbol{X}^{\mathrm{test}}, \{\boldsymbol{X}_i\}_{i\in[n]}\right) \geq T\left(\boldsymbol{x}^{\mathrm{test}}, \{\boldsymbol{x}_i\}_{i\in[n]}\right)\right). \tag{7}$$

Unfortunately, the probability in Eq. (7) is computationally intractable, as it depends on the complex computation process of deep kNN. To address this issue, we introduce the SI framework and define a significance quantification measure called the *selective p-values*. In the following section, we present the concept of selective p-values and demonstrate that this measure can appropriately quantify the statistical significance of anomalies detected by kNN-based AD.

4 Selective Inference (SI) for kNN-base AD

In this section, we propose selective p-values, based on the framework of SI, as a measure of the statistical significance of anomalies detected by kNN-based AD. These selective p-values can be interpreted in the same way as conventional p-values. For example, if the significance level is set to $\alpha = 0.05$ and we consider anomalies with selective p-values less than 0.05, it is theoretically guaranteed that the proportion of falsely detected anomalies will remain below 0.05.

4.1 Alternative formulations of the test statistic in Eq.(6)

First, let us denote the $(1+n)d$-dimensional vector obtained by concatenating the test instance $\boldsymbol{x}^{\mathrm{test}}$ and n training instance $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, all of which are d-dimensional vectors, as

$$\boldsymbol{y} = \mathrm{vec}\left(\boldsymbol{x}^{\mathrm{test}}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right) \in \mathbb{R}^{(1+n)d}, \tag{8}$$

where vec is the operation that concatenates multiple vectors into a single column vector. Similarly, the $(1+n)d$-dimensional vector obtained by concatenating $1+n$ random vectors is denoted as

$$\boldsymbol{Y} = \mathrm{vec}\left(\boldsymbol{X}^{\mathrm{test}}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right) \in \mathbb{R}^{(1+n)d}. \tag{9}$$

With these notations, we can rewrite the test statistic in Eq.(6) as

$$T(\boldsymbol{Y}) = \|\boldsymbol{\eta}_{\boldsymbol{y}}^\top \boldsymbol{Y}\|_2, \tag{10}$$

where $\boldsymbol{\eta}_{\boldsymbol{y}}$ is a $(1+n)d$-dimensional vector defined as

$$\boldsymbol{\eta}_{\boldsymbol{y}} = \frac{1}{\sqrt{2}}\left(\underbrace{1,\ldots,1}_{1,\ldots,d}, \underbrace{0,\ldots,0}_{d+1,\ldots,2d}, \quad \ldots, \quad \underbrace{-1,\ldots,-1}_{o(k)d+1,\ldots,(1+o(k))d}, \quad \ldots, \quad \underbrace{0,\ldots,0}_{nd+1,\ldots,(1+n)d}\right). \tag{11}$$

Note that the vector $\boldsymbol{\eta}_{\boldsymbol{y}}$ depends on the data $\boldsymbol{y}$ through the selected neighborhood $o(k)$.

4.2 Naive p-values

Here, we discuss a measure referred to as the *naive p-value*. Although this measure is *invalid* as an indicator of statistical significance, it serves as a contrasting concept that helps introduce the notion of selective p-values. The naive p-value is defined as

$$p_{\mathrm{naive}} := \mathbb{P}_{\mathrm{H}_0}\left(\|\boldsymbol{\eta}_{\boldsymbol{y}}^\top \boldsymbol{Y}\|_2 \geq \|\boldsymbol{\eta}_{\boldsymbol{y}}^\top \boldsymbol{y}\|_2\right), \tag{12}$$

where we emphasize the distinction between the random vector $\boldsymbol{Y}$ and the observed vector $\boldsymbol{y}$. Under the statistical model in Eq.(4), the random vector $\boldsymbol{Y}$ follows a multivariate normal distribution. Therefore, the statistic $\tilde{T}(\boldsymbol{Y})$ follows a χ distribution with $(1+n)d$ degrees of freedom. Consequently, Eq.(12) can be easily computed as the tail probability of $\chi((1+n)d)$ distribution. Unfortunately, this easily computable naive p-value is invalid in the sense that it does not account for the fact that the k-nearest neighbor instance $\boldsymbol{x}_{o(k)}$ is selected based on the same observed data. If naive p-values are used for decision-making in the same way as ordinary p-values, the false detection rate cannot be properly controlled as intended.

4.3 Selective p-values

The basic idea of SI, pioneered by the seminal work by Lee et al. [8], is to address the problem of the naive p-values by employing the framework of conditional testing, based on the key insight that the sampling distribution of a test statistic can be tractable if the statistic is conditioned on.

In order to define selective p-value, let us represent an event that the k-nearest neighbor index $o(k)$ is selected based on the random vector $\boldsymbol{Y}$ as "$\mathcal{E}_{\boldsymbol{Y}} = o(k)$". Then, selective p-value is defined as

$$p_{\text{selective}} = \mathbb{P}_{\text{H}_0}\left(\|\boldsymbol{\eta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}\|_2 \geq \|\boldsymbol{\eta}_{\boldsymbol{y}}^\top \boldsymbol{y}\|_2 \;\middle|\; \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\right), \tag{13}$$

where the first condition "$\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$" indicates the event that the k-nearest neighbor index $o(k)$ obtained from the random data vector $\boldsymbol{Y}$ is the same as that obtained from the observed data vector $\boldsymbol{y}$. The second condition "$\mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}$" indicates that the sufficient statistic of the nuisance parameter defined as

$$\mathcal{Q}_{\boldsymbol{Y}} = \left(\frac{P\boldsymbol{Y}}{\|P\boldsymbol{Y}\|}, \left(I_{(n+1)d} - P\right)\boldsymbol{Y}\right), \tag{14}$$

where $P = \boldsymbol{\eta}_{\boldsymbol{y}}\boldsymbol{\eta}_{\boldsymbol{y}}^\top \in \mathbb{R}^{(1+n)d\times(1+n)d}$, is the same for both the random data vector $\boldsymbol{Y}$ and the observed data vector $\boldsymbol{y}$. Here, the key idea is, by the first conditioning on the k-nearest neighbor index $o(k)$, $\boldsymbol{\eta}_{\boldsymbol{Y}}$ is fixed as $\boldsymbol{\eta}_{\boldsymbol{y}}$, making the computation of the probability in Eq.(13) tractable with the use of χ^2 distribution as with the case of the naive p-value. Due to space limitations, we omit the mathematical details and statitical rationale of SI, including the role of the event related to the nuisance parameter [2]. For further details, we refer the reader to the literature such as [7, 8, 34].

The computation of selective p-value in Eq.(13) is reduced to a tail probability computation of *truncated χ^2 distribution* as formally stated in the following theorem.

Theorem 4.1. *The following conditional test statistic*

$$\|\boldsymbol{\eta}_{\boldsymbol{Y}}^\top \boldsymbol{Y}\|_2 \;\middle|\; \{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\} \tag{15}$$

follows a truncated χ distribution with $(1+n)d$ degrees of freedom, where the truncation is determined by the constraint "$\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$" and the domain of the distribution is on the one-dimensional subspace defined by $\{\boldsymbol{Y} \mid \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\}$.

The proof of Theorem 4.1 is deferred to Appendix A.2. Furthermore, the selective p-value in Eq.(13) is valid p-value in the sense that, for any significance level $\alpha \in (0, 1)$, the false detection rate of the anomalies with $p_{\text{selective}} < \alpha$ is exactly α as formally stated in the following theorem.

Theorem 4.2. *The selective p-values defined in Eq.*(13) *satisfies*

$$\mathbb{P}_{\text{H}_0}\left(p_{\text{selective}} \leq \alpha\right) = \alpha, \; \forall \alpha \in (0, 1). \tag{16}$$

The proof of Theorem 4.2 is deferred to Appendix A.2. The statement of Theorem 4.2 indicates that the selective p-value in Eq.(13) can be used as a measure to quantify the statistical significance of the detected anomalies.

[2] In this context, the nuisance parameter refers to a parameter that is included in the null distribution but is not of direct inferential interest. To characterize the null distribution, it is necessary to eliminate the influence of the nuisance parameter. In our approach, this is achieved by conditioning on its sufficient statistic. This is a standard technique in the field of selective inference [7, 8, 34].

4.4 Selection event characterization

The main technical challenge in computing selective p-values lies in characterizing the selection event "$\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$". Conditioning on this selection event means identifying the set of vectors $\boldsymbol{Y} \in \mathbb{R}^{(1+n)d}$ such that the same normal instance $\boldsymbol{x}_{o(k)}$, which was selected as the k-nearest neighbor based on the observed data vector $\boldsymbol{y}$, is also selected as the k-nearest neighbor. To realize this, one must appropriately characterize and account for several factors, including the computation of latent feature vectors in the trained CNN model and the comparison of distances between the test instance and normal instances in the latent feature space. Due to space limitations, the detailed characterization of the selection event is provided in Appendix B, where we adopt the parametric programmingbased approach proposed in [38] as the core computational framework.

5 Numerical Experiments

In this section, we demonstrate that the proposed method exhibits high power (true positive rate) while controlling the Type I error rate (false positive rate) below the significance level compared to other methods. First, experiments are conducted on synthetic datasets, followed by similar experiments on two types of real datasets. All experiments are conducted with a significance level $\alpha = 0.05$. We examined the case where ϕ, a mapping to a latent feature space, is the identity function in simple settings for the synthetic and tabular data experiments, while for the image data experiments, we considered the case where ϕ is defined by a DNN model. We executed all experiments on AMD EPYC 9474F processor, 48-core 3.6GHz CPU and 768GB memory.

5.1 Baseline Methods

In the experiments on synthetic datasets and tabular datasets, we compared the proposed method (`proposed`) with four other baseline methods: `Hotelling_t2`, `w/o-pp`, `naive`, and `Bonferroni`. Subsequently, in the experiments on image datasets, we additionally compare two further ablation studies: `OpA1` and `OpA2`.

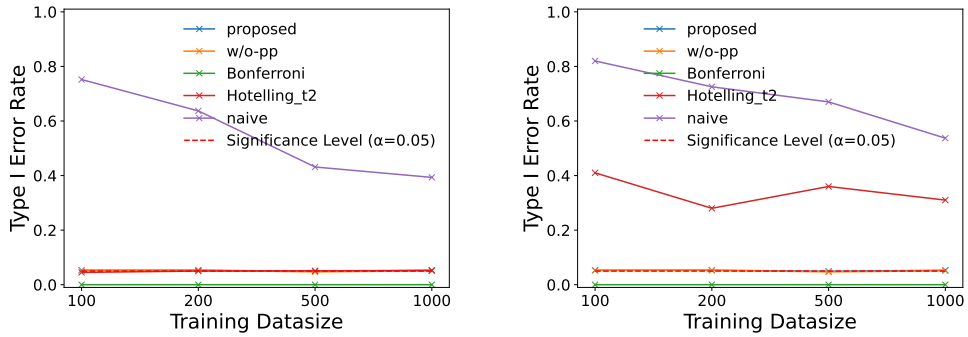
- `Hotelling_t2`: This method is not based on kNN but instead performs classical anomaly detection by computing a p-value using Hotellings T^2 test. Unlike `proposed`, which assumes semi-parametric conditions on the training data as shown in Eq.(4), this method makes a stricter parametric assumption that all training samples are i.i.d. from a single Gaussian distribution. It uses the squared Mahalanobis distance $T^2(\boldsymbol{X}^{\mathrm{test}})$ as the test statistic and computes the p-value as $p_{\mathrm{hotelling}} = \mathbb{P}_{\mathrm{H}_0}(T^2(\boldsymbol{X}^{\mathrm{test}}) \geq T^2(\boldsymbol{x}^{\mathrm{test}}))$, where the null distribution is assumed to follow the χ^2 distribution.
- `Bonferroni`: We chose the `Bonferroni` correction as the most basic multiple comparison method[3]. This is a method to control the Type I error rate by using the Bonferroni correction. There are $\binom{n}{k}$ ways to choose the neighbors $\mathcal{N}$, then we compute the Bonferroni corrected p-value as $p_{\mathrm{bonferroni}} = \min(1, \binom{n}{k} \cdot p_{\mathrm{naive}})$.
- `naive`: This is a statistical test that uses the naive p-value defined in Eq.(12).
- `w/o-pp`: An ablation study that excludes the parametric programming technique described in Appendix B.
- `OpA1`: Another ablation study that excludes the selection events for kNN (i.e., $\mathcal{N}_{\boldsymbol{Y}}$, $\mathcal{K}_{\boldsymbol{Y}}$, and $\mathcal{S}_{\boldsymbol{Y}}$ in Appendix B).
- `OpA2`: Another ablation study that excludes the selection events for DNN (i.e., $\mathcal{D}_{\boldsymbol{Y}}$ in Appendix B).

5.2 Synthetic Datasets

To evaluate the Type I error rate, we varied the training dataset size $n \in \{100, 200, 500, 1000\}$ and set the data dimension $d = 5$. The number of neighbors k was adaptively selected in a data-driven manner from $\{1, 2, 5, 10\}$. See Appendix C.2 for results when d and k are varied. For each

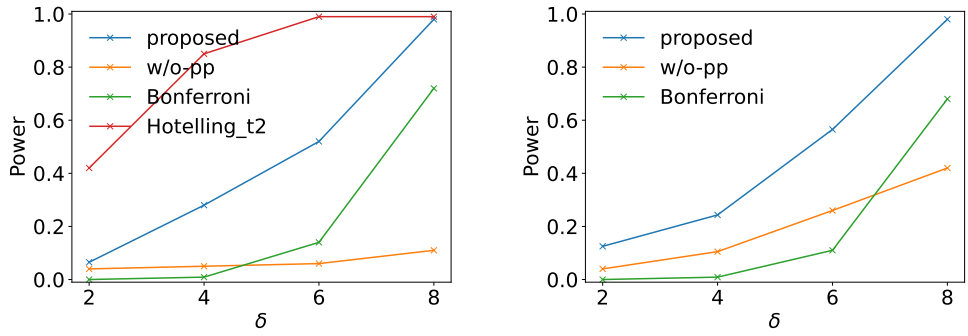
[3]Other multiple testing procedures, such as Holm's method, require computing nominal p-values for all possible hypotheses, which is an intractable task due to the combinatorial explosion.

configuration, we conducted 1,000 independent experiments. In each iteration, we generated a test instance $\boldsymbol{x}^{\text{test}}$ and training instances $\boldsymbol{x}_i$ for $i \in [n]$, sampled from the same distribution under one of the following two settings. In the **parametric** setting, all instances were drawn from a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. In the **semi-parametric** setting, each instance was drawn from $\mathcal{N}(\boldsymbol{s}_i, \boldsymbol{I}_d)$, where $\boldsymbol{s}_i$ is a randomly generated mean vector. For more details about data generation, see Appendix C.1. To evaluate the power, we generated data in the same way, except that a signal $\delta \in \{2, 4, 6, 8\}$ was added to a randomly selected coordinate of the test instance $\boldsymbol{x}^{\text{test}}$. We set $d = 5$, $n = 100$ and k was adaptively selected in the same way. See Appendix C.3 for results when d, k, and n are varied. The results of Type I error rate are shown in Figure 2. The `proposed`, `w/o-pp`, and `Bonferroni` successfully controlled the Type I error rate under the significance level, whereas the `naive` could not. Since `naive` failed to control the Type I error rate, we excluded it from the power experiment. `Hotelling_t2` also fails in the semi-parametric setting (where s_i follows a non-Gaussian distribution), because it assumes that all samples are i.i.d. from a single Gaussian distribution. So, it is likewise excluded from the power experiment in the semi-parametric setting. The results of power are shown in Figure 3. Among the methods that controlled the Type I error rate, the `proposed` has the highest power.



(a) Parametric (b) Semi-Parametric

Figure 2: Results of Type I error rate when changing the dataset size n. `proposed`, `w/o-pp`, and `Bonferroni` successfully control the Type I error rate across all settings. Their lines are almost overlapping. `naive` fails and the results of `Bonferroni` are almost zero, because it is too conservative. `Hotelling_t2` also fails in the semi-parametric setting.

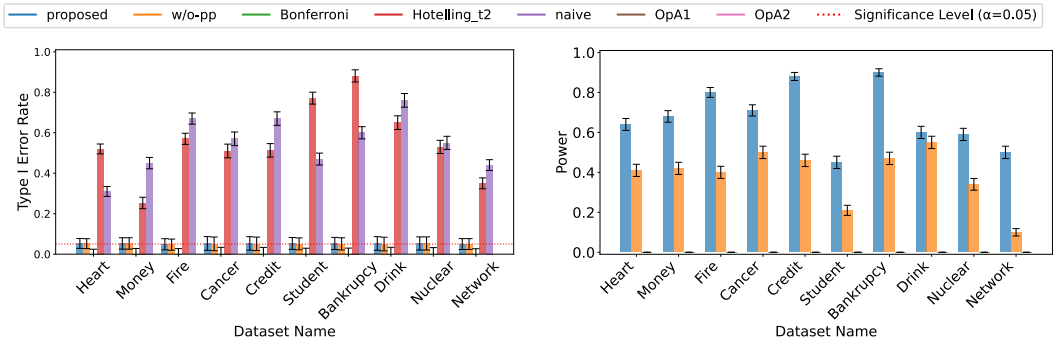


(a) Parametric (b) Semi-Parametric

Figure 3: Power when varying signal strength δ. `proposed` and `Hotelling_t2` outperformed other methods. However, `Hotelling_t2` failed to control the Type I error rate in the semi-parametric setting so it is not shown in the power results.

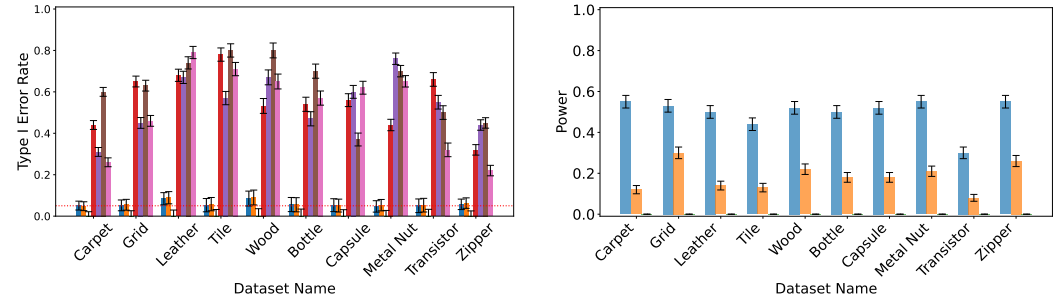
5.3 Real Datasets I: Tabular Data

We conducted evaluations using 10 tabular real-world datasets. These datasets reflect various real-world problems from different domains. The datasets used in our experiments are listed in



Type I Error Rate Results on Tabular Data

Power Results on Tabular Data



Type I Error Rate Results on Image Data

Power Results on Image Data

Figure 4: Results on real datasets. The proposed method (`proposed`) outperformed the other methods in terms of power, while controlling the Type I error rate below the significance level across all datasets. The Type I error rate, power, and error bars of the `Bonferroni` are almost zero, because it is too conservative.

Appendix C.4. Only numerical features from each dataset were used in the experiments. The datasets vary in dimensionality, ranging from 4 to 10 dimensions. The number of neighbors k was adaptively selected in a data-driven manner from $\{1, 2, 5, 10\}$. Before conducting the experiments, All datasets are standarized with each feature having mean 0 and variance 1. The results of the Type I error rate and power are shown in Figure 4. The `proposed` method outperformed the other methods in terms of power, while controlling the Type I error rate.

5.4 Real Datasets II: Image Data

In this experiment, we used the MVTec AD dataset [61, 62]. The dataset consists of 15 classes, and we chose 10 classes for the experiments which seem to follow a normal distribution. The datasets used in our experiments are listed in Appendix C.5. Before conducting the experiments, All datasets are standarized with each feature having mean 0 and variance 1. Following the conventional Deep kNN approach [19], we employed a ResNet model pre-trained on the ImageNet dataset as a feature extractor in this experiment. As a preprocessing step, the original image, which has a size of 900×900 or 1024×1024, was divided into 30×30 patches, and the patch was used as the test instance. For the training instances, we used 100 patches from the same position as the test instance. We set the number of neighbors $k = 3$. The results of the Type I error rate and power are shown in Figure 4. The `proposed` method outperformed the other methods in terms of power, while controlling the Type I error rate below the significance level. Some examples of the experimental results are shown in Figure 5 and Figure 15.

6 Scope, limitations and conclusions

In this study, we proposed a method for quantifying uncertainty in kNN-based AD by assessing statistical significance. Uncertainty quantification in the outputs of deep learning models remains a major challenge in machine learning community, and our work contributes toward addressing this gap. This is particularly important for AD because it is often used in high-stakes applications

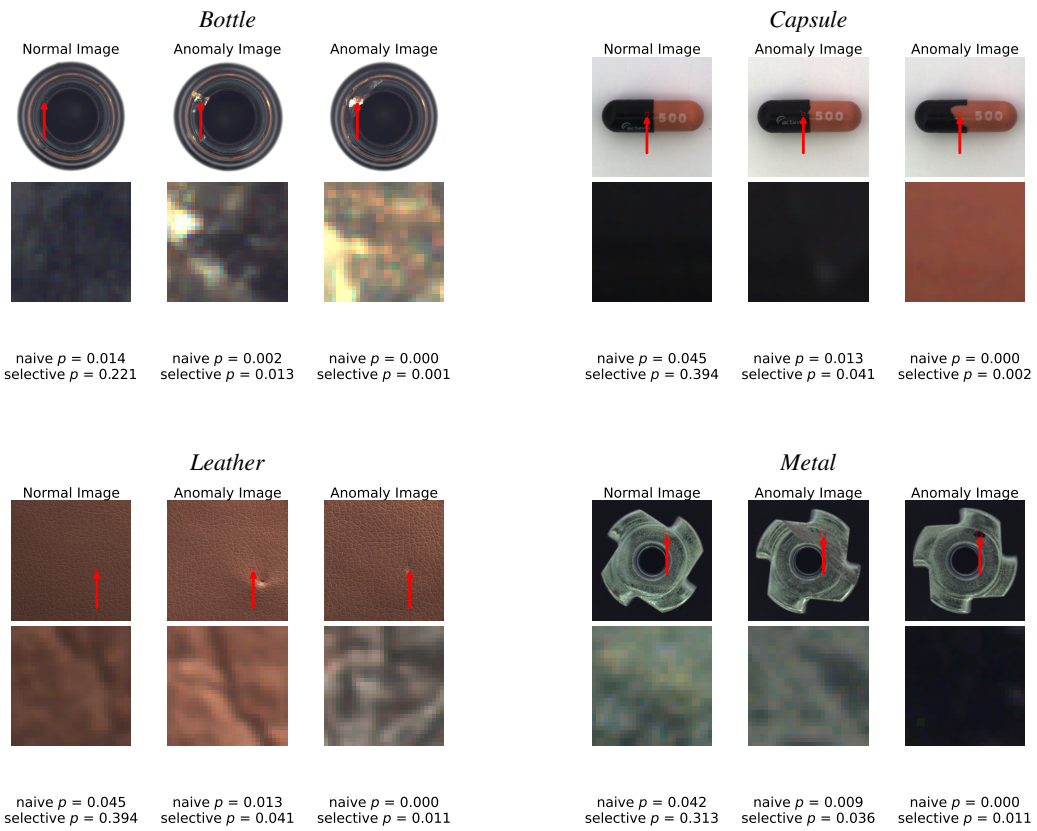


Figure 5: Experimental results of 4 datasets from MVTec AD dataset. For each dataset, one normal example (left) and two anomaly examples (center, right) are showed. For each example, the top row displays the original image used for testing along with the patch location (marked in red), while the bottom row presents the extracted patch image. For all normal examples, the naive p-value is below the significance level $\alpha = 0.05$ (false positive), whereas the proposed selective p-value correctly results in a true negative. For all anomaly examples, the selective p-value successfully detects anomalies.

such as medical diagnosis and industrial inspection, where evaluating statistical significance is crucial for practical reliability. Therefore, our proposed method, which enables the quantification of statistical significance of the detected anomalies, has substantial practical importance. Our SI-based fundamental idea is not limited to the specific kNN-based AD algorithm presented as a proof of concept in this study; it is also applicable to various other variants. For SI of deep kNN-based AD, it is essential to characterize both the computations by the deep learning model and the selection event of the kNN instances. The former is applicable to many CNN-type networks, while the latter applies to distance metrics such as L_1, L_2, and L_∞. It is, in principle, extendable to other deep AD methods, as long as the selection event (i.e., $\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$ in Eq.(13)) can be characterized in a tractable form.

One limitation of the proposed kNN test lies in the form of the semi-parametric model discussed in Section 3. In this model, the signal components are entirely non-parametric, offering significantly greater flexibility than conventional statistical AD. However, the need to assume a distribution for the noise component remains a limitation. Additional experimental results on robustness when the noise distribution deviates from the normal distribution are presented in Appendix C, in which we observe that when the deviation is small, the false detection rate can be maintained at approximately the desired level. Another limitation arises when the selection event becomes more complex, rendering the current framework inapplicable in its existing form. This issue, for instance, occurs when a Transformer is used as the deep learning model for identifying the latent feature space. This remains an important challenge for our future work.

Acknowledgments and Disclosure of Funding

This work was partially supported by JST CREST (JPMJCR21D3, JPMJCR22N2), JST Moonshot R&D (JPMJMS2033-05), RIKEN Center for Advanced Intelligence Project, and JSPS KAKENHI Grant Number JP24K15080.

References

[1] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[2] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3379–3388, 2018.

[3] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018.

[4] Kishan G Mehrotra, Chilukuri K Mohan, HuaMing Huang, Kishan G Mehrotra, Chilukuri K Mohan, and HuaMing Huang. *Anomaly detection*. Springer, 2017.

[5] Shuai Lyu, Dongmei Mo, and Wai keung Wong. Reb: Reducing biases in representation for industrial anomaly detection. *Knowledge-Based Systems*, 290:111563, 2024.

[6] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540, 2009.

[7] Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

[8] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 2016.

[9] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438, 2000.

[10] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[11] Bo Du and Liangpei Zhang. A discriminative metric learning based anomaly detection method. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):6844–6857, 2014.

[12] Thibaud Ehret, Axel Davy, Jean-Michel Morel, and Mauricio Delbracio. Image anomalies: A review and synthesis of detection methods. *Journal of Mathematical Imaging and Vision*, 61:710–743, 2019.

[13] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[14] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[15] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *European conference on principles of data mining and knowledge discovery*, pages 15–27. Springer, 2002.

[16] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18, 2015.

[17] Mikael Sabuhi, Ming Zhou, Cor-Paul Bezemer, and Petr Musilek. Applications of generative adversarial networks in anomaly detection: A systematic literature review. *Ieee Access*, 9:161003–161029, 2021.

[18] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022.

[19] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020.

[20] Jason D Lee and Jonathan E Taylor. Exact post model selection inference for marginal screening. *Advances in neural information processing systems*, 27, 2014.

[21] Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

[22] Jonathan Taylor and Robert Tibshirani. Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018.

[23] Joshua R Loftus. Selective inference after cross-validation. *arXiv preprint arXiv:1511.08866*, 2015.

[24] Ali Charkhi and Gerda Claeskens. Asymptotic post-selection inference for the akaike information criterion. *Biometrika*, 105(3):645–664, 2018.

[25] Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pages 2469–2477, 2016.

[26] Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, and Ichiro Takeuchi. Selective inference for sparse high-order interaction models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3338–3347. JMLR. org, 2017.

[27] Sangwon Hyun, Max Gsell, and Ryan J Tibshirani. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097, 2018.

[28] Kazuya Sugiyama, Vo Nguyen Le Duy, and Ichiro Takeuchi. More powerful and general selective inference for stepwise feature selection using the homotopy continuation approach. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[29] David Rügamer and Sonja Greven. Inference for l 2-boosting. *Statistics and computing*, 30(2):279–289, 2020.

[30] Diptesh Das, Vo Nguyen Le Duy, Hiroyuki Hanada, Koji Tsuda, and Ichiro Takeuchi. Fast and more powerful selective inference for sparse high-order interaction model. *arXiv preprint arXiv:2106.04929*, 2021.

[31] David Rügamer, Philipp FM Baumann, and Sonja Greven. Selective inference for additive and linear mixed models. *Computational Statistics & Data Analysis*, 167:107350, 2022.

[32] Snigdha Panigrahi, Peter W MacDonald, and Daniel Kessler. Approximate post-selective inference for regression with the group lasso. *Journal of machine learning research*, 24(79):1–49, 2023.

[33] Thang Loi Nguyen, Loc Duong, and Vo Nguyen Le Duy. Statistical inference for feature selection after optimal transport-based domain adaptation. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025.

[34] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

[35] Yoshikazu Terada and Hidetoshi Shimodaira. Selective inference for the problem of regions via multiscale bootstrap. *arXiv preprint arXiv:1711.00949*, 2017.

[36] Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

[37] Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein. Bayesian post-selection inference in the linear model. *arXiv preprint arXiv:1605.08824*, 28, 2016.

[38] Vo Nguyen Le Duy and Ichiro Takeuchi. More powerful conditional selective inference for generalized lasso by parametric programming. *The Journal of Machine Learning Research*, 23(1):13544–13580, 2022.

[39] Jason D Lee, Yuekai Sun, and Jonathan E Taylor. Evaluating the statistical significance of biclusters. *Advances in neural information processing systems*, 28, 2015.

[40] Chihiro Watanabe and Taiji Suzuki. Selective inference for latent block models. electron. *J. Stat*, 15(1), 2021.

[41] Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pages 1–11, 2022.

[42] Yiqun T Chen and Daniela M Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

[43] Kosuke Tanizaki, Noriaki Hashimoto, Yu Inatsu, Hidekata Hontani, and Ichiro Takeuchi. Computing valid p-values for image segmentation by selective inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9553–9562, 2020.

[44] Vo Nguyen Le Duy and Ichiro Takeuchi. Exact statistical inference for time series similarity using dynamic time warping by selective inference. *arXiv preprint arXiv:2202.06593*, 2022.

[45] Toshiaki Tsukurimichi, Yu Inatsu, Vo Nguyen Le Duy, and Ichiro Takeuchi. Conditional selective inference for robust regression and outlier detection using piecewise-linear homotopy continuation. *Annals of the Institute of Statistical Mathematics*, 74(6):1197–1228, 2022.

[46] Vo Nguyen Le Duy, Hsuan-Tien Lin, and Ichiro Takeuchi. Cad-da: Controllable anomaly detection after domain adaptation by statistical inference. In *International Conference on Artificial Intelligence and Statistics*, pages 1828–1836. PMLR, 2024.

[47] Tatsuya Matsukawa, Tomohiro Shiraishi, Shuichi Nishino, Teruyuki Katsuoka, and Ichiro Takeuchi. Statistical test for auto feature engineering by selective inference. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025.

[48] Vo Nguyen Le Duy, Shogo Iwazaki, and Ichiro Takeuchi. Quantifying statistical significance of neural network-based image segmentation by selective inference. *Advances in Neural Information Processing Systems*, 2022.

[49] Miwa Daiki, Vo Nguyen Le Duy, and Ichiro Takeuchi. Valid p-value for deep learning-driven salient region. In *Proceedings of The International Conference on Learning Representations*, 2023.

[50] Tomohiro Shiraishi, Daiki Miwa, Teruyuki Katsuoka, Vo Nguyen Le Duy, Kouichi Taji, and Ichiro Takeuchi. Statistical test for attention maps in vision transformers. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2024.

[51] Teruyuki Katsuoka, Tomohiro Shiraishi, Daiki Miwa, Vo Nguyen Le Duy, and Ichiro Takeuchi. Statistical test for generated hypotheses by diffusion models. *arXiv preprint arXiv:2402.11789*, 2024.

[52] Daiki Miwa, Tomohiro Shiraishi, Vo Nguyen Le Duy, Teruyuki Katsuoka, and Ichiro Takeuchi. Statistical test for anomaly detections by variational auto-encoders. *arXiv preprint arXiv:2402.03724*, 2024.

[53] Teruyuki Katsuoka, Tomohiro Shiraishi, Daiki Miwa, Shuichi Nishino, and Ichiro Takeuchi. si4onnx: A python package for selective inference in deep learning models. *arXiv preprint arXiv:2501.17415*, 2025.

[54] Sangwon Hyun, Max G'Sell, and Ryan J Tibshirani. Exact post-selection inference for change-point detection and other generalized lasso problems. *arXiv preprint arXiv:1606.03552*, 2016.

[55] Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, and Ichiro Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. In *Advances in Neural Information Processing Systems*, pages 11356–11367, 2020.

[56] Sean Jewell, Paul Fearnhead, and Daniela Witten. Testing for a change in mean after change-point detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1082–1104, 2022.

[57] Tomohiro Shiraishi, Daiki Miwa, Vo Nguyen Le Duy, and Ichiro Takeuchi. Selective inference for change point detection by recurrent neural network. *Neural Computation*, 37(1):160–192, 2024.

[58] Akifumi Yamada, Tomohiro Shiraishi, Shuichi Nishino, Teruyuki Katsuoka, Kouichi Taji, and Ichiro Takeuchi. Change point detection in the frequency domain with statistical reliability. *Transactions on Machine Learning Research*, 2025.

[59] Shuxiao Chen and Jacob Bien. Valid inference corrected for outlier removal, 2019.

[60] Don O Loftsgaarden and Charles P Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

[61] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[62] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (12 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our paper includes the discussion of limitations in §6

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provided complete proofs of all the theoretical claims in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the experiment settings are fully described.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We submitted the code as supplementary materials. In the experiments, we only used datasets that are publicly available or can be synthetically generated.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides all necessary training and test details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: This study is a paper on a new statistical testing method and properly discusses statistical significance of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We described the information on computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All the studies are conducted under NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work, which focuses on statistical tests for kNN-based anomaly detection, aims to enhance the reliability of AI and has the potential to broadly influence the machine learning community. On the other hand, it does not present significant ethical concerns or foreseeable societal consequences because this work is theoretical and, as of now, has no direct applications that might impact society or raise ethical considerations.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: This study is on the statistical reliability of AI and poses no risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We appropriately credited all the code, data, and models used in this study.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets. (Details of the experiments performed by the code for reproducibility provided in the supplementary material are given in the paper.)

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This study does not involve research with human subjects, including crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our research does not involve studies on human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We did not use LLMs in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Proof of Theorems

A.1 Proof of Theorem 4.1

According to the condition on $\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$ and $\mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}$, i.e., $\mathcal{U}(\boldsymbol{Y}) = \mathcal{U}(\boldsymbol{y})$ and $\mathcal{V}(\boldsymbol{Y}) = \mathcal{V}(\boldsymbol{y})$, where

$$\mathcal{V}(\boldsymbol{Y}) = \frac{P\boldsymbol{Y}}{\|P\boldsymbol{Y}\|_2} \in \mathbb{R}^{(n+1)d}, \quad \mathcal{U}(\boldsymbol{Y}) = (I_{(n+1)d} - P)\boldsymbol{Y} \in \mathbb{R}^{(n+1)d}.$$

we have

$$\begin{aligned} & \mathcal{U}(\boldsymbol{Y}) = \mathcal{U}(\boldsymbol{y}) \\ \Leftrightarrow\ & (I_{(n+1)d} - P)\boldsymbol{Y} = \mathcal{U}(\boldsymbol{y}) \\ \Leftrightarrow\ & \boldsymbol{Y} = \mathcal{U}(\boldsymbol{y}) + \mathcal{V}(\boldsymbol{Y})z \\ \Leftrightarrow\ & \boldsymbol{Y} = \mathcal{U}(\boldsymbol{y}) + \mathcal{V}(\boldsymbol{y})z \quad (\because \mathcal{V}(\boldsymbol{Y}) = \mathcal{V}(\boldsymbol{y})) \\ \Leftrightarrow\ & \boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}z, \end{aligned}$$

where $\boldsymbol{a} = \mathcal{U}(\boldsymbol{y}), \boldsymbol{b} = \mathcal{V}(\boldsymbol{y})$, and $z = T(\boldsymbol{Y}) = ||\boldsymbol{\eta}_{\boldsymbol{y}}^T \boldsymbol{Y}||_2 = ||P\boldsymbol{Y}||_2$.

Then, we have

$$\begin{aligned} & \{\boldsymbol{Y} \in \mathbb{R}^{(1+n)d} \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})\} \\ =& \{\boldsymbol{Y} \in \mathbb{R}^{(1+n)d} \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}z, z \in \mathbb{R}\} \\ =& \{\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}z \in \mathbb{R}^{(1+n)d} \,|\, \mathcal{E}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{E}_{\boldsymbol{y}}, z \in \mathbb{R}\} \\ =& \{\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}z \in \mathbb{R}^{(1+n)d} \,|\, z \in \mathcal{Z}\}, \end{aligned}$$

where $\mathcal{Z}$ is the truncation region defined as

$$\mathcal{Z} = \{z \in \mathbb{R} \,|\, \mathcal{E}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{E}_{\boldsymbol{y}}\}.$$

Therefore, by noting that $\|\eta_{\boldsymbol{y}}^\top \boldsymbol{s}\|_2$ is zero, we obtain

$$T(\boldsymbol{Y}) \mid \{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})\} \sim \mathrm{TC}(\mathrm{tr}(P), \mathcal{Z}),$$

where $\mathrm{TC}(\mathrm{tr}(\mathrm{P}), \mathcal{Z})$ is a truncated χ-distribution with the degrees of freedom $(1+n)d$, whose domain is the truncation region $\mathcal{Z}$.

A.2 Proof of Theorem 4.2

The sampling distribution of the test statistic conditional on $\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}$ and $\mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})$ denoted by

$$T(\boldsymbol{Y}) \,|\, \{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})\}$$

is a truncated χ-distribution with the degrees of freedom $(1+n)d$ and the truncation region $\mathcal{Z}$ defined in Theorem 4.1. Thus, by applying the probability integral transform, under the null hypothesis,

$$p_{\text{selective}} \,|\, \{\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})\} \sim \mathrm{Unif}(0, 1),$$

which leads to

$$\mathbb{P}_{\mathrm{H}_0}\left(p_{\text{selective}} \leq \alpha \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y})\right) = \alpha, \forall \alpha \in (0, 1).$$

Next, for any $\alpha \in (0, 1)$, we have

$$\begin{aligned} & \mathbb{P}_{\mathrm{H}_0}\left(p_{\text{selective}} \leq \alpha \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}\right) \\ &= \int \mathbb{P}_{\mathrm{H}_0}\left(p_{\text{selective}} \leq \alpha \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}, \mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{Y})\right) \mathbb{P}_{\mathrm{H}_0}\left(\mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{Y}) \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}\right) d\mathcal{Q}(\boldsymbol{y}) \\ &= \alpha \int \mathbb{P}_{\mathrm{H}_0}(\mathcal{Q}(\boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{y}) \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}) d\mathcal{Q}(\boldsymbol{y}) \\ &= \alpha. \end{aligned}$$

Therefore, we obtain the result in Theorem 4.2 as follows:

$$\begin{aligned} \mathbb{P}_{\mathrm{H}_0}\left(p_{\text{selective}} \leq \alpha\right) &= \sum_{\mathcal{E}_{\boldsymbol{y}}} \mathbb{P}_{\mathrm{H}_0}\left(p_{\text{selective}} \leq \alpha \,|\, \mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}\right) \mathbb{P}_{\mathrm{H}_0}(\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}) \\ &= \alpha \sum_{\mathcal{E}_{\boldsymbol{y}}} \mathbb{P}_{\mathrm{H}_0}(\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}_{\boldsymbol{y}}) \\ &= \alpha. \end{aligned}$$

B Selection Event Characterization

In this section, we characterize the selection events $\mathcal{E}_{\boldsymbol{Y}} = \mathcal{E}\boldsymbol{y}$ of deep kNN-based anomaly detection (AD). The selection event of deep kNN-based AD consists of two components: the selection event related to the kNN-based AD, and the selection event related to the deep learning models that perform the transformation into latent features. The former is described in Appendix B.1, and the latter in Appendix B.2. Finally, in Appendix B.3, we describe how to identify the data space that satisfies the selection event and how to compute the selective p-values.

B.1 Selection Event for kNN Anomaly Detection

In the selection events of kNN-AD, it is necessary to consider events such as selecting the k nearest instances, the anomaly score exceeding a threshold, and determining k based on the data. In the following, we describe these events one by one. It is worth noting that all the events described below can be collectively represented by a set of linear inequalities, which facilitates the computation of truncation regions for the truncated normal distribution used in selective p-value calculations.

Selection event for k^{th} nearest neighbor The test statistic in Eq. (6) depends on the selection of k^{th} nearest neighbor instance of the test instance $\boldsymbol{X}^{\text{test}}$. Therefore, the condition on the k^{th} nearest neighbor instance is required. Specifically, by conditioning on

$$\text{dist}(\boldsymbol{X}^{\text{test}}, \boldsymbol{X}_{o(k)}) \geq \text{dist}(\boldsymbol{X}^{\text{test}}, \boldsymbol{X}_{o(k')}) \tag{17}$$

for $k' = 1, \ldots, k-1$, and

$$\text{dist}(\boldsymbol{X}^{\text{test}}, \boldsymbol{X}_{o(k)}) \leq \text{dist}(\boldsymbol{X}^{\text{test}}, \boldsymbol{X}_{o(k')}) \tag{18}$$

for $k' = k+1, \ldots, n$, we can consider only cases where the k-the nearest neighbor is the same as the observed case. Hereafter, the conditions in Eq.(17) and Eq.(18) are collectively represented as $\mathcal{N}_{\boldsymbol{Y}} = \mathcal{N}_{\boldsymbol{y}}$.

Selection event for anomaly score Since the statistical test is performed only on test instances selected in the AD, it is essential to consider the selection events associated with it. A test instance is selected and if its anomaly score, as defined in Eq. (3), exceeds a threshold θ. The condition for the anomaly score is written as

$$\log \text{dist}\left(\boldsymbol{X}^{\text{test}}, \boldsymbol{X}_{o(k)}\right) - \frac{\log k}{d} \geq \theta. \tag{19}$$

With the conditions in Eq.(19), we can characterize the selection event that the test case $\boldsymbol{X}^{\text{test}}$ is selected in AD. Hereafter, the condition in Eq.(19) is represented as $\mathcal{K}_{\boldsymbol{Y}} = \mathcal{K}_{\boldsymbol{y}}$.

Selection event for data-driven selection of k In the case of the data-driven option for determining the number of neighbors k, its effect must also be appropriately considered as a selection event. For example, consider the scenario where $k_1, \ldots, k_K$ are candidate values for k, and the candidate that maximizes the anomaly score in Eq. (3) is selected. Let the selected $k \in \{k_1, \ldots, k_K\}$ be denoted as k^*. Then, the selection event is simply given by $\log \text{dist}(\boldsymbol{x}^{\text{test}}, \boldsymbol{x}_{o(k^*)}) - \frac{\log k^*}{d} \geq \log \text{dist}(\boldsymbol{x}^{\text{test}}, \boldsymbol{x}_{o(k_t)}) - \frac{\log k_t}{d}, \forall t \in [K]$. In the case of data-driven option to determine k, in addition to the four selection events mentioned above, this event must also be incorporated as an additional condition. Hereafter, we denote this selection event as $\mathcal{S}_{\boldsymbol{Y}} = \mathcal{S}_{\boldsymbol{y}}$.

B.2 Selection Event for Deep Learning Models

When using k-nearest neighbors AD with feature representations from a pre-trained deep learning model, the influence of the model should be considered as a selection event. SI for deep learning has been discussed in prior studies, and tools like the software facilitate the analysis of selection events in these models. In this study, we employ methods from earlier research to calculate selective p-values, taking into account selection events related to deep learning models. The basic idea in these methods involves decomposing the model into components and representing each as a piecewise linear function. For example, operations in a CNN such as convolution, ReLU activation, max

pooling, and up-sampling are represented as piecewise linear functions. In the experiment, we utilize the feature representation of a CNN model pre-trained on the ImageNet database. This model is represented precisely as a composition of piecewise linear functions. We explain the selection events regarding the deep learning model that transforms an image instance $\boldsymbol{x}_i \in \mathbb{R}^d$ to a latent feature vector $\boldsymbol{z}_i \in \mathbb{R}^{\tilde{d}}$. We consider a deep learning model that consists of sequential piecewise-linear functions (e.g., convolution, ReLU activation, max pooling, and up-sampling). Obviously, the composite function of those piecewise-linear functions maintains its piecewise-linear nature. Thus, within a specific real space in $\mathbb{R}^d$, the deep learning model simplifies to a linear function, which can be expressed as:

$$\phi_{\mathrm{DL}}(\boldsymbol{x}_i) = \boldsymbol{B} + \boldsymbol{W}\boldsymbol{x}_i \quad \text{if } \boldsymbol{x}_i \in \mathcal{P}, \tag{20}$$

where $\boldsymbol{B} \in \mathbb{R}^{\tilde{d}}$ and $\boldsymbol{W} \in \mathbb{R}^{\tilde{d} \times d}$ represent the bias and weight matrices, and $\mathcal{P} \subseteq \mathbb{R}^d$ is a polytope where ϕ_{DL} acts as a linear function. The polytope can be characterized by a set of linear inequalities. For details on computing these linear inequalities, see [53]. Let us denote the set of polytopes for all instances in $\boldsymbol{Y}$ as:

$$\mathcal{D}_{\boldsymbol{Y}} := \{\mathcal{P} \mid \boldsymbol{X}_i \in \boldsymbol{Y}, \boldsymbol{X}_i \in \mathcal{P}\}. \tag{21}$$

Hereafter, we denote the selection event as $\mathcal{D}_{\boldsymbol{Y}} = \mathcal{D}_{\boldsymbol{y}}$.

B.3 Computing Selective p-values

Based on the discussions in Appendix B.1 and B.2, selective p-values in (13) can be rewritten as follows:

$$p_{\text{selective}} := \mathbb{P}_{\mathrm{H}_0}\left(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}) | \mathcal{N}_{\boldsymbol{Y}} = \mathcal{N}_{\boldsymbol{y}}, \mathcal{K}_{\boldsymbol{Y}} = \mathcal{K}_{\boldsymbol{y}}, \mathcal{S}_{\boldsymbol{Y}} = \mathcal{S}_{\boldsymbol{y}}, \mathcal{D}_{\boldsymbol{Y}} = \mathcal{D}_{\boldsymbol{y}}, \mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}\right). \tag{22}$$

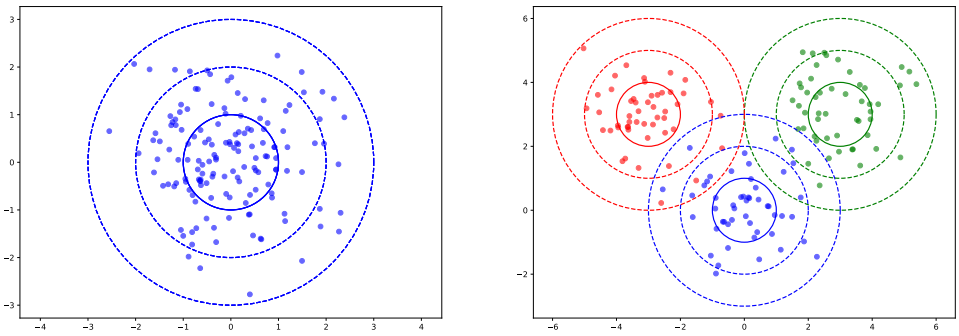
Calculating this selective p-values is complex, but we effectively use methods from existing SI research. We specifically use the parametric programming (pp)-based method from previous studies [38]. In SI, statistical inference is based on the probability measure within the subspace $\mathcal{Z}$ of the data space $\mathbb{R}^{(1+n)d}$ where selection event conditions are met. By conditioning on the selection event for the nuisance component, $\mathcal{Q}_{\boldsymbol{Y}} = \mathcal{Q}_{\boldsymbol{y}}$, $\mathcal{Z}$ reduces to a one-dimensional subspace (see Theorem 4.1 and its proof in Appendix A.1). The selection events are formulated as unions of intersections of linear or quadratic inequalities, suitable when using L_1 or L_2 distances for k-nearest neighbors. $\mathcal{Z}$ consists of finite number of intervals along a line in the $(1+n)d$-dimensional space, and the pp-based method systematically enumerates all intervals that meet these conditions.

Since the noise is Gaussian, the test statistic $T(\boldsymbol{Y})$ under the null hypothesis H_0 follows a one-dimensional truncated Gaussian distribution within the subspace $\mathcal{Z}$, comprising finite intervals along a line. The selective p-value is calculated as the tail probability of this truncated distribution. Early SI research often simplified calculations by assuming $\mathcal{Z}$ as a single interval under additional conditions, which still controls the false detection probability but reduces detection power. In our problem, a similar simplification can be considered by enforcing $\mathcal{Z}$ to be a single interval. In the experiments in §5, we conduct an ablation study comparing this simple approach (denoted as `w/o-pp`) as one of the baselines.

C Details of the Experiments

C.1 Details of Synthetic Data Generation

This subection provides additional details regarding the generation of synthetic datasets used in Section 5.2. We describe both the parametric and semi-parametric settings. To illustrate the two data-generation settings, we present in Figure 6 the distributions of the training samples in a two-dimensional example ($d = 2$). In the parametric setting, all samples are centered around the origin. In contrast, in the semi-parametric setting, the samples are distributed around different mean vectors $\boldsymbol{s}_i$, producing a mixture of Gaussian clusters. This visualization clarifies the structural difference between the two settings and the increased heterogeneity in the semi-parametric case.



(a) Parametric setting

(b) Semi-parametric setting

Figure 6: Visualization of the data-generation process in the parametric and semi-parametric settings for $d = 2$. In the parametric case, all training samples are drawn from a single Gaussian distribution centered at the origin. In the semi-parametric case, each training sample is drawn from a Gaussian distribution with a randomly shifted mean vector $\boldsymbol{s}_i$, resulting in a heterogeneous distribution.

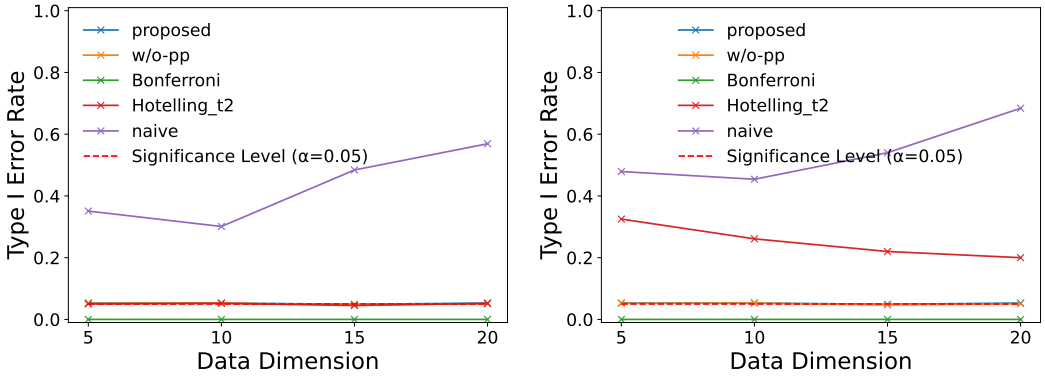
C.2 Additional Type I Error Rate Results

We also conducted experiments to investigate the Type I error rate when the data dimension n, d and the number of neighbors k were varied in the parametric and semi-parametric setting. Specifically, we varied $d \in \{5, 10, 15, 20\}$, $k \in \{1, 2, 5, 10\}$ and $n \in \{100, 200, 500, 1000\}$, while setting the default parameters as $d = 5$, $k = 3$ and $n = 100$. In all cases, we generated the datasets in the same way as in the experiments on synthetic datasets (§5.2). The results are shown in Figures 7, 8 and 9.

To further assess the robustness of our method, we conducted experiments on datasets that deviate from the normal distribution. Specifically, data are sampled from the exponentially modified Gaussian (EMG), generalized normal distribution (GND), skew normal distribution (SND), and Student's t-distribution. The degree of deviation from the normal distribution is quantified using the Wasserstein distance l, and we evaluate the Type I error rate for each case by varying $l \in \{0.01, 0.02, 0.03, 0.04\}$. The results are shown in Figure 10.

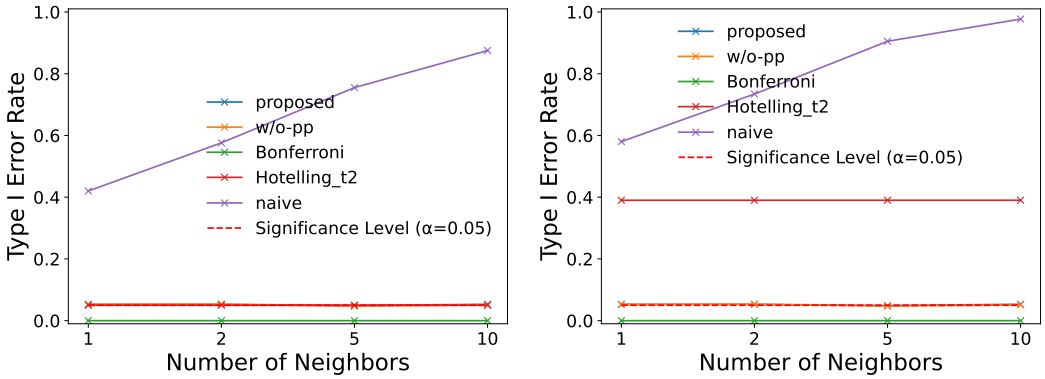
C.3 Additional Power Results

We also conducted experiments to investigate the power when the number of training data n, the data dimension d and the number of neighbors k are varied in the parametric and semi-parametric setting. We varied $n \in \{100, 200, 500, 1000\}$, $d \in \{5, 10, 15, 20\}$ and $k \in \{1, 2, 5, 10\}$ while setting the default parameters as $n = 100$, $d = 5$, $k = 3$ and signal strength $\delta = 5$. Furthermore, we conducted additional experiments where n and d was varied, considering the case where k was adaptively selected from $\in \{1, 2, 5, 10\}$ in a data-driven manner. In all cases, we generated the datasets in the same way as in the experiments on synthetic datasets (§5.2). The results are shown in Figures 11, 12, 13, and 14.



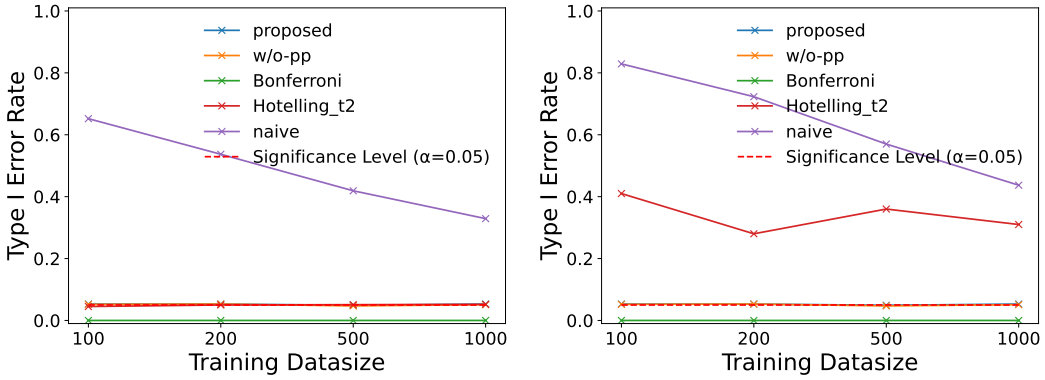
(a) Parametric (b) Semi-Parametric

Figure 7: Results of Type I error rate when varying the date dimension d. `proposed`, `w/o-pp`, and `Bonferroni` successfully control the Type I error rate across all settings. `naive` fails and the results of `Bonferroni` are almost zero, because it is too conservative. Since `Hotelling_t2` does not involve a parameter k, its value remains unchanged. `Hotelling_t2` also fails in the semi-parametric setting.



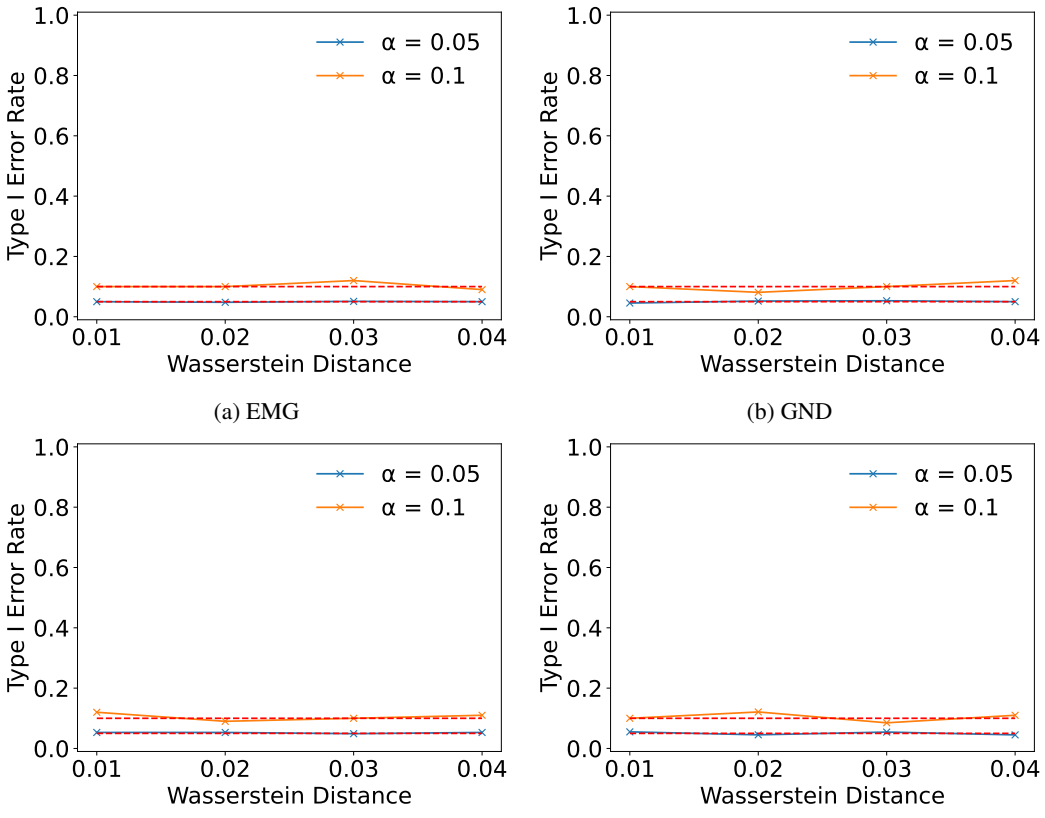
(a) Parametric (b) Semi-Parametric

Figure 8: Results of Type I error rate when varying the number of neighbors k. `proposed`, `w/o-pp`, and `Bonferroni` successfully control the Type I error rate across all settings. `naive` fails and the results of `Bonferroni` are almost zero, because it is too conservative. Since `Hotelling_t2` does not involve a parameter k, its value remains unchanged in the both settings. In the semi-parametric setting, `Hotelling_t2` fails to control the Type I error rate.



(a) Parametric (b) Semi-Parametric

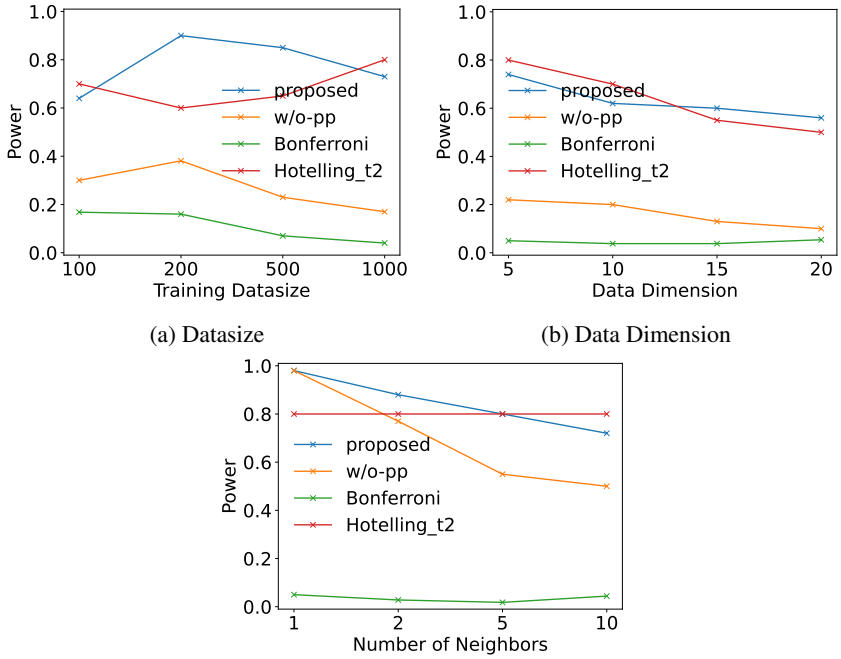
Figure 9: Results of Type I error rate when varying the number of datasize n. `proposed`, `w/o-pp`, and `Bonferroni` successfully control the Type I error rate across all settings. `naive` fails and the results of `Bonferroni` are almost zero, because it is too conservative. `Hotelling_t2` also fails in the semi-parametric setting.



(a) EMG (b) GND

(c) SND (d) Student's t-distribution

Figure 10: Results of Type I error rate when varying the Wasserstein distance l. `proposed` successfully control the Type I error rate in both significance levels $\alpha\{0.05, 0.1\}$.

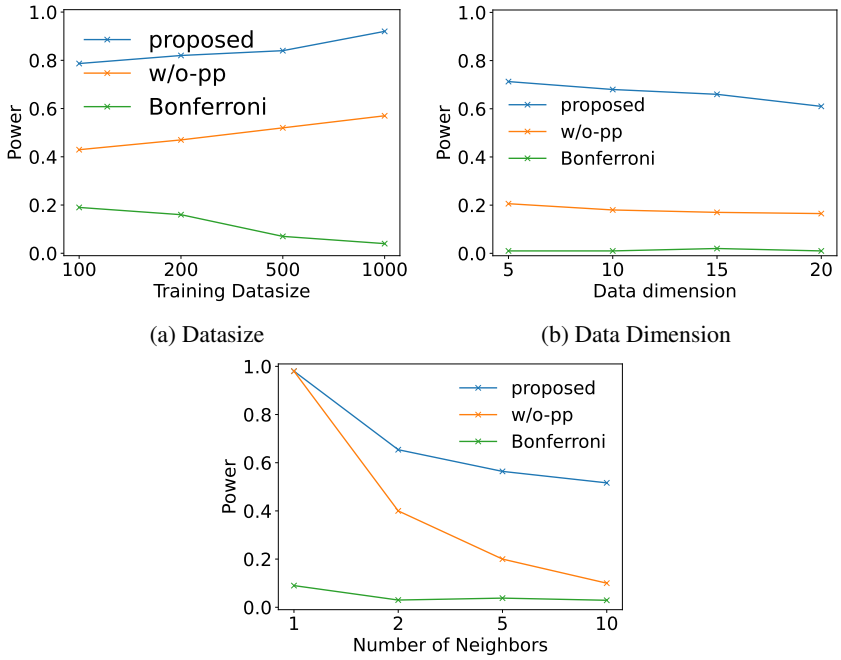


(a) Datasize

(b) Data Dimension

(c) Number of Neighbors

Figure 11: These are in the parametric setting. Power for a fixed number of neighbors k . The results show the effect of varying the training dataset size n, the data dimension d, and k. Our proposed method (`proposed`) and `Hotelling_t2` outperformed other methods across all settings.

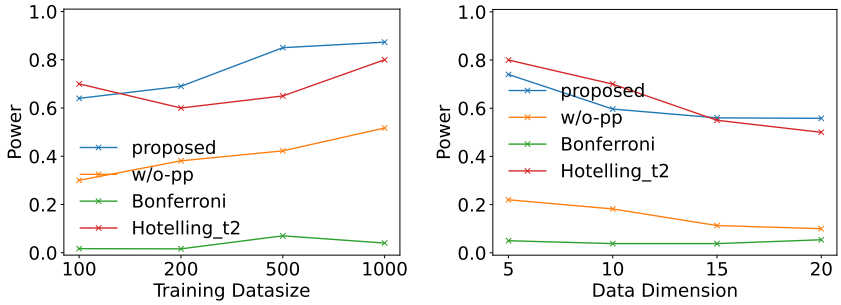


(a) Datasize

(b) Data Dimension

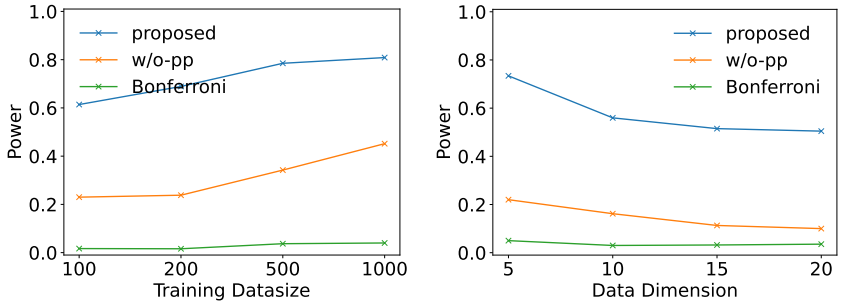
(c) Number of Neighbors

Figure 12: These are in the semi-parametric setting. Power for a fixed number of neighbors k. The results show the effect of varying the training dataset size n, the data dimension d, and k. Our proposed method (`proposed`) outperformed other methods across all settings.



(a) Datasize (b) Data Dimension

Figure 13: These are in the parametric setting. Power for an adaptively selected number of neighbors k. The results show the effect of varying the training dataset size n and the data dimension d. Our proposed method (`proposed`) and `Hotelling_t2` outperformed other methods across all settings.



(a) Datasize (b) Data Dimension

Figure 14: These are in the semi-parametric setting. Power for an adaptively selected number of neighbors k. The results show the effect of varying the training dataset size n and the data dimension d. Our proposed method (`proposed`) outperformed other methods across all settings.

C.4 Details of Tabular Datasets

We used the following 10 real datasets from the Kaggle Repository. All datasets are licensed under the CC BY 4.0 license.

- *Heart*: Dataset for predicting heart attacks
- *Money*: Dataset on financial transactions in a virtual environment
- *Fire*: Dataset on fires in the MUGLA region in June
- *Cancer*: Dataset related to breast cancer diagnosis
- *Credit*: Dataset on credit card transactions
- *Student*: Dataset related to student performance
- *Bankruptcy*: Dataset on company bankruptcies
- *Drink*: Dataset on the quality of drinking water
- *Nuclear*: Dataset on pressurized nuclear reactors
- *Network*: Dataset on anomaly detection in virtual network environments

C.5 Experimental Results on Image Data Examples

We evaluated `proposed` and `naive` on the 10 datasets from MVTec AD dataset. The datasets used in this study are *Carpet*, *Grid*, *Leather*, *Tile*, *Wood*, *Bottle*, *Capsule*, *Metal Nut*, *Transistor*, and *Zipper*. Examples except for those shown in the Figure 5 from each dataset are shown in Figure 15. In each example, we present patches corresponding to true negative and true positive cases, along with both the naive p-value and the selective p-value.

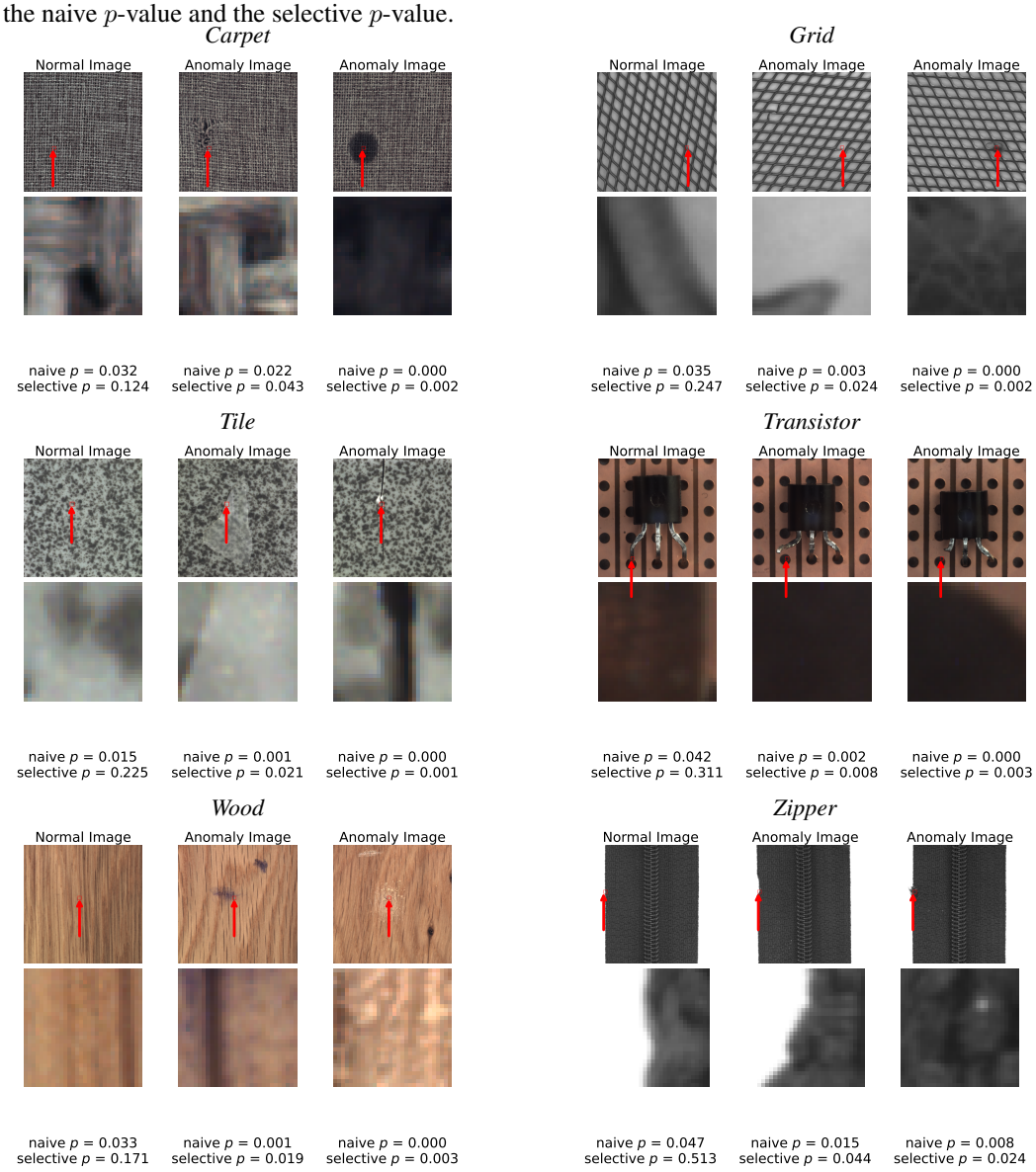


Figure 15: Experimental results of 6 datasets from MVTec AD dataset. For each dataset, one normal example (left) and two anomaly examples (center, right) are showed. For each example, the top row displays the original image used for testing along with the patch location (marked in red), while the bottom row presents the extracted patch image. For all normal examples, the naive p-value is below the significance level $\alpha = 0.05$ (false positive), whereas the proposed selective p-value correctly results in a true negative. For all anomaly examples, the selective p-value successfully detects anomalies.