Pattern Recognition 177 (2026) 113254

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

STOOD-X: Explainable out-of-distribution detection via nonparametric statistical testing on large-scale datasets

Iván Sevillano-García *, Julián Luengo , Francisco Herrera

Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain

ARTICLE INFO

Keywords:
Explainable artificial intelligence
Deep learning
Out-of-distribution

ABSTRACT

Out-of-distribution (OOD) detection is a critical task in machine learning, particularly in safety-sensitive applications where model failures can have serious consequences. However, current OOD detection methods often suffer from restrictive distributional assumptions, limited scalability, and a lack of interpretability. To address these challenges, we propose **STOOD-X**, a two-stage methodology that combines a Statistical nonparametric Test for OOD Detection with eXplainability enhancements. In the first stage, STOOD-X uses feature-space distances and a Wilcoxon-Mann-Whitney test to identify OOD samples without assuming a specific feature distribution. In the second stage, it generates user-friendly, concept-based visual explanations that reveal the features driving each decision, aligning with the BLUE XAI paradigm. Through extensive experiments on benchmark datasets and multiple architectures, STOOD-X achieves competitive performance compared to state-of-the-art post hoc OOD detectors, particularly in high-dimensional and complex settings. In addition, its explainability framework enables human oversight, bias detection, and model debugging, fostering trust and collaboration between humans and AI systems. Therefore, STOOD-X offers a robust, explainable, and scalable solution for real-world OOD detection tasks.

1. Introduction

Out-of-Distribution (OOD) detection has emerged as a challenge within machine learning [1], which consists of differentiating between In-Distribution (ID) and OOD samples. In particular, when dealing with Artificial Intelligence (AI) models, where any instance that can be introduced into the model obtains a prediction, it is essential to recognize when an introduced instance matches the data distribution for which the model has been trained. In safety-critical scenarios, the absence of OOD algorithms can lead AI models to make incorrect decisions instead of deferring to human judgment. As a result, the ability to reliably detect OOD samples has become a fundamental requirement for building robust, reliable, and trustworthy AI systems.

Various algorithms have been developed to address this challenge. OOD detection algorithms can be broadly categorized based on where to start applying the algorithm, where we can differentiate between training-based and post hoc algorithms. Training-based algorithms apply their approximation from the training stage. These algorithms modify this stage by adding regularizations to increase separability between ID and OOD samples [2] or even add trainable layers from which to obtain the OOD score [3]. Post hoc methods work on already trained models, chosen when retraining costs are prohibitive.

Within post hoc algorithms, there are different approaches depending on the basis of the algorithm. Classification-based algorithms are based on the model output to detect OOD samples [4]. Gradient-based algorithms focus on analyzing the gradients of ID samples to distinguish them from OOD samples [5]. Distance-based algorithms take advantage of the feature space to detect OOD samples by measuring the distance between ID and OOD samples. Some approximations use parametric assumptions such as Gaussianity in the feature space [6] while others do not, using nonparametric analysis [7]. Furthermore, recent research has explored the combination of the strengths of multiple algorithms [8], integrating scores from parametric and nonparametric methods.

Despite these advances, existing OOD detection methods face several limitations. Many approaches rely on strong assumptions about the data distribution, such as Gaussianity, which may not hold in real-world scenarios, and others lack explainability, making it difficult to understand how and why a sample is classified as OOD. In addition, most of these

* Corresponding author.
E-mail addresses: isevillano@go.ugr.es (I. Sevillano-García), julianlm@decsai.ugr.es (J. Luengo), herrera@decsai.ugr.es (F. Herrera).

https://doi.org/10.1016/j.patcog.2026.113254
Received 4 April 2025; Received in revised form 5 February 2026; Accepted 6 February 2026
Available online 7 February 2026

algorithms propose scores without theoretically robust meaning, such as statistical tests. These challenges highlight the need for a more robust and explainable solution for OOD detection.

In parallel to these algorithmic advances, there has been a growing interest in explainable AI (XAI) techniques that provide insight into how models make decisions [9]. A well-established definition of XAI is provided in [10] as "given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand". We can also differentiate between different queries to ask an AI system, such as who, when, what, and how to explain a decision [11]. A recent deep analysis and reflection on XAI is done in [12]. Two ways of considering XAI raised in [13], BLUE (responsiBle, Legal, trUst, Ethics) and RED (Research, Explore, Debug) XAI, bringing the first into a sphere of analysis of expert understanding and trustworthy AI. The BLUE XAI approach will be used in this paper, considering the distinction between the stakeholders analyzed in [12] (see Fig. 4 of this article, which contains a diagram showing different audience profiles). Specifically, on the OOD detection task, recent work uses explanations to validate AI decisions [14]. These approaches take advantage of visualizations to help users understand why a sample is classified as OOD, foster trust, and enable human-AI collaboration.

In this work, we address the OOD detection problem from an XAI perspective by introducing STOOD-X (Statistical Test for OOD detection enhanced with eXplainability), a two-stage methodology. The first stage of this methodology is a novel post hoc OOD detection algorithm that leverages feature space distances and statistical tests to detect OOD samples. The second stage consists of explanation generation, which provides a clear and user-friendly visualization and reasons for the decisions made during the OOD detection process. STOOD-X uses nonparametric statistical tests with meaningful probability-based scores and provides explanations for OOD detection decisions.

Our experiments show the competitive performance of the STOOD-X in multiple benchmark datasets and architectures and its explainability potential, establishing it as a promising solution for real-world OOD detection.

The remainder of this paper is organized as follows. Section 2 reviews related work in OOD detection and XAI. Section 3 presents STOOD-X, which differentiates between the detection and explanation stages. Section 4 describes the experimental setup, while Section 5 presents results and comparisons with the state-of-the-art methods. Section 6 shows the explainability capabilities. Finally, Section 7 concludes the paper and discusses future directions.

2. Related works

In this section, we present a comprehensive review of the literature on OOD detection, focusing on both theoretical advancements and practical applications. In Section 2.1, we begin with a discussion of theoretical proposals and improvements for OOD detection, differentiating the different types of algorithms by their principles and methodologies. Then, in Section 2.2, we explore how XAI approaches are used to bring the OOD decision to human understanding.

2.1. OOD detection algorithms

In this section, we analyze the different approaches that have been used to tackle the problem of OOD detection. Specifically, here is a brief summary of the different OOD proposals on which the STOOD-X methodology has been inspired. For a more detailed study of the taxonomy of generalized OOD methods, we refer to [15].

We can differentiate between OOD detection algorithms between training-based and post hoc methods. Training-based algorithms are developed to impose a set of constraints in training time so that the resulting model performs better on the OOD detection task. An example of this training-based algorithm can be found in [3], where a modified model is trained to estimate its confidence as a scalar between 0 and 1. Other approximations of this training-based perspective impose regularization factors, such as multi-prototype contrastive learning [16], on the training stage to facilitate separability between OOD and ID samples. Post hoc algorithms study the OOD detection problem for an already trained AI model, without modifying the training stage. This property is important in a real-world environment where the cost of retraining a model with new OOD restrictions is prohibitive. An example of this approximation is ODIN [4], which uses temperature scaling and input perturbation to improve the separability between ID and OOD samples. ODIN's approach to amplifying differences in softmax scores has inspired subsequent research, including methods that use energy scores [17] for the detection of OOD. Algorithms such as ASH [18] use the energy score obtained from logits by simplifying the internal representation of features to maximize the difference between the ID and OOD samples. In Fig. 1, we show graphically how the ASH algorithm simplifies the internal representation of the features of the model to infer whether the behavior of the model in a new sample is an OOD or ID sample based on the energy score of the modified representation.

Other algorithms base their OOD scores on gradients instead of logits or feature representation. This is the case with GradNorm [19], which uses the magnitude of gradients as an OOD detector. These methods focus on improving the discriminative power of OOD scores by taking advantage of different aspects of the model output.

More recently, distance-based methods have emerged as a promising direction in OOD detection. These methods rely on the assumption that OOD samples are relatively far from ID classes in the feature space. This intuition of these methods appears naturally, as the model learns a feature space where ID sample classes are distinguishable from each other. However, if an OOD sample is brought into the feature space, it may have features of different classes anecdotally. These features may not include all the characteristics required to belong to a particular class, nor be as present as in a sample of the class itself.

These methods usually work in the same way: the distance from the ID distribution is calculated. If this distance exceeds a certain threshold, it is considered an OOD sample. The search for this threshold is done experimentally and is optimized to separate ID and OOD data. However, this threshold is empirically searched and evaluated by the experimental result. It lacks theoretical robustness to support the results.

There are distance-based algorithms with very different principles. Parametric algorithms such as the minimum Mahalanobis distance score (MDS) [6], or its variation Relative-MDS(RMDS) [20], study the distances from the OOD sample to the centroid of each class assuming that the features are normally distributed. Alternatively, there are nonparametric algorithms that make no strong assumptions about the underlying feature distribution, making it a versatile and effective method. The work proposed in [5] demonstrated the effectiveness of K-nearest-neighbor (KNN) distances as a nonparametric OOD detection method. Finally, a last avenue explored is the combination of parametric and nonparametric algorithms. This is the case with CombOOD [8], which blends the RMDS and KNN scores to obtain a combined OOD score.

Although distance-based algorithms can use various distances, recent work [21] shows that the cosine distance between features is useful for differentiation between OOD and ID examples. This distance works especially well when the vectors to be measured are sparse, with many dimensions with a value of 0.

2.2. XAI in OOD detection

OOD detection algorithms aim to detect OOD samples and differentiate them from ID samples. However, once detected, these algorithms must provide reasons why these samples have been detected as OOD. This is where XAI is introduced to propose human-understandable reasoning so that humans can evaluate whether each sample is OOD.

Within the field of study of XAI, two perspectives can be distinguished [13]. The RED XAI (Research, Explore, Debug XAI) refers to a field of explainability focused on the development and resolution of

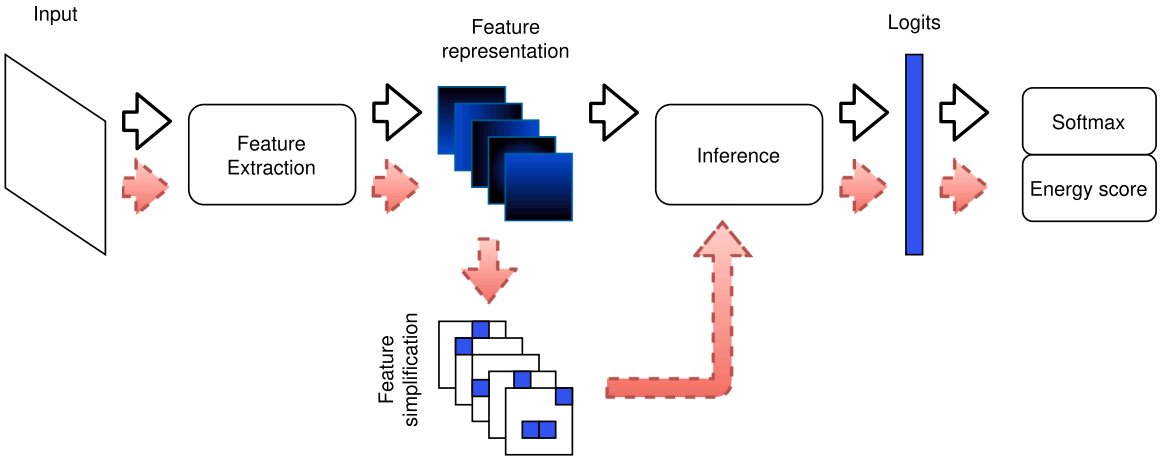


Fig. 1. Diagrammatic representation of the ASH methodology for estimating OOD confidence via Energy scores of features simplification. Following the chart, first block represents the feature extraction from the input sample. Second part of the chart represents the simplification of the feature representation, which is the main contribution of the ASH algorithm. The third part of the chart represents the energy score calculation based on the simplified features. White arrows represent the natural flow of the prediction model, while red arrows represent the flow used by ASH to calculate the OOD confidence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

AI. In contrast, the BLUE XAI (responsiBle, Legal, trUst, Ethics) aims to improve the proposal of explanations to a final user, ensuring its good behavior. Within both approaches, ensuring that an OOD detection algorithm is understood by a final user is part of the BLUE XAI perspective.

Answering the question of how to present an explanation, we can differentiate between several explanation proposals [11]: Explanations based on the importance of the feature or explanations based on examples. On the one hand, explanations based on feature importance assign a percentage of influence to each part of the sample input, showing how much it has influenced the final decision. It is presented to the final user by a heatmap highlighting these importances. On the other hand, explanations based on examples show samples similar in some sense to the intended example, either to enforce the decision (prototypes) or to show the changes needed to change these decisions (counterfactual).

Explanations based on feature importance have two different perspectives, black-box and white-box. Black-box algorithms do not use the internal structure of a model [22] so that the explanation is not biased by the specifications of the model itself. However, these explanations require a large number of model evaluations to provide a reliable explanation. The white-box perspective takes advantage of the knowledge of the internal structure to develop more specific explanations, such as Layer-wise Relevance Propagation (LRP) [23], which preserves the importance of the decision along the internal layers of the model. By being able to use the knowledge of the internal structure, these algorithms can propose explanations in reasonable time and cost.

Based on LRP, Concept-based Relevance Propagation (CRP) [24] has been developed. CRP introduces the term 'concept' to separate different features within the same explanation to obtain GLocal explanations (Global-Local). GLocal explanations offer a local perspective (where the concept is located on the example to explain) and a global perspective (which examples have the same concepts present as the example to explain).

Based on CRP and making use of prototypes, Prototypical Concept-based Explanations (PCX) [14] is proposed as an explanation proposal for different tasks. For the OOD detection approximation, this proposal can be classified as a parametric distance-based OOD detector, which computes the distances to class prototypes. However, the main contribution of this work in the XAI perspective is the explanation proposal, which presents prototypical dataset examples identified by the AI model, highlighting similar features to the analyzed sample and their locations. This method bases its explanation proposal on the selection of prototypes. For a method that does not use prototypes to detect OOD examples, the explanation must be adapted to align with the underlying approach to remain meaningful.

3. STOOD-X methodology: OOD detection algorithm using feature space analysis and statistical tests enhanced by explainability

In this section, we describe the fundamentals of the STOOD-X methodology, a novel explainable two-stage methodology designed to detect OOD samples using feature space distances to Nearest Neighbors (NNs) and statistical tests with improved explainability. Due to its feature-based construction of observed ID samples, it provides an approximation that can contribute to the explainability of the OOD detector. This methodology ensures a unified framework that combines accurate OOD detection with robust explainability, empowering stakeholders to make informed decisions while maintaining trust in the AI system.

We organize the description of the STOOD-X methodology as follows. Section 3.1 outlines the overall workflow. Section 3.2 formulates the mathematical foundations of the detection algorithm. Section 3.3 details the explanation generation stage that aligns with the BLUE XAI paradigm.

3.1. Flowchart of the STOOD-X methodology

This section describes the natural workflow of STOOD-X methodology, from sample presentation to final decision and explanation. Fig. 2 illustrates how statistical analysis and explainability mechanisms interact to provide reliable detection and human-understandable justifications.

The STOOD-X methodology executes through two sequential yet interdependent stages:

1. In Stage 1, the original model processes its features and maps them to the learned feature space. Within this space, the distances to previously observed samples are analyzed and a statistical hypothesis test is performed to determine whether the new sample is sufficiently close to the validated ID samples. Based on a predefined significance threshold, the detection algorithm classifies the new sample as ID or OOD. Two possible outcomes follow:
   - If the new sample is classified as ID, it is deemed ready to be processed by the model, as there are validated nearby samples.
   - If the sample is classified as OOD, it should be excluded from processing, as no similar instances exist in the feature space.

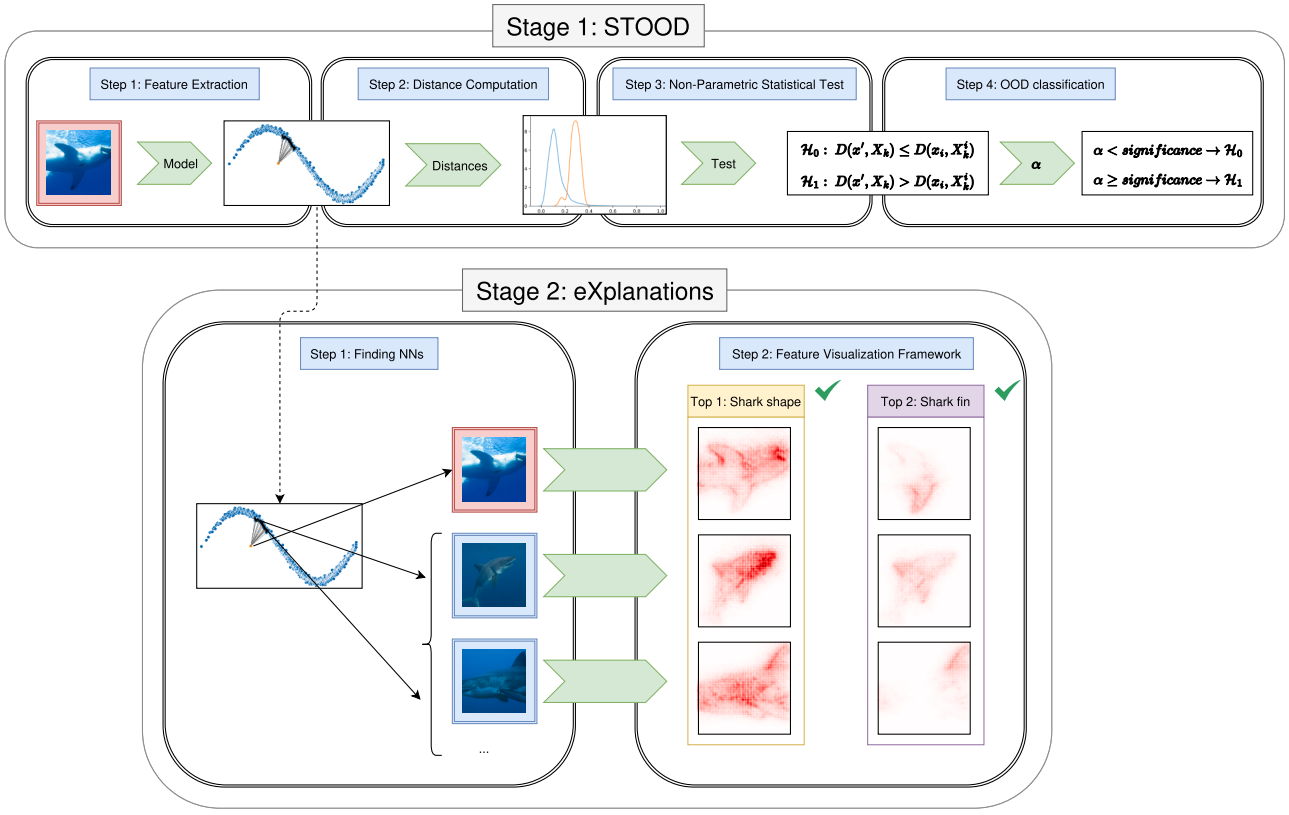


Fig. 2. Flowchart of the STOOD-X methodology.

2. Stage 2 of the STOOD-X methodology becomes relevant, providing additional insights into the model's decision-making process by presenting nearby samples along with the most important features shared between the new sample and its closest counterparts. The explanation generated is valuable for the understanding of the model of the final user, regardless of whether the sample is classified as ID or OOD.
   - In the ID case, the user should be able to verify that the features used by the model align with those of validated samples.
   - In the OOD case, the final user should also validate that the identified features do not provide meaningful or comparable information relative to previously validated samples.

   Although we have already mentioned that Stage 2 is useful in both cases, the explanations generated are particularly valuable in the case of OOD classification scenario, the model's uncertainty can be effectively translated into a meaningful query for the final user. By highlighting the closest validated samples, the model implicitly asks whether the new sample should be processed despite not having encountered a similar instance before. This enables the user to assess a validation of the model's decision and take appropriate action. As a result, the approach helps prevent OOD samples from being mistakenly processed as ID while also allowing the reconsideration of ID samples that were misclassified as OOD, ensuring a more reliable decision-making process.

In the next sections, we will go into the specifics of both stages. These detailed explanations will provide a clearer understanding of the processes involved in each stage, building on the general description given earlier.

3.2. STOOD-X methodology first stage: OOD detection algorithm

In this section, we introduce the OOD detection algorithm that constitutes the first stage of the STOOD-X methodology. We begin by formally defining the key concepts and notation that establish the basis of our approach, ensuring a clear and rigorous foundation. This formalization provides the necessary framework to systematically derive the STOOD-X methodology, allowing a principled approach to OOD detection.

Let $\mathcal{X}, \mathcal{V}, \mathcal{Y}$ be the input, feature, and output space, respectively. Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a machine learning model. This model f can be modeled as follows:

$$\begin{aligned} V &: \mathcal{X} \rightarrow \mathcal{V} = \mathbb{R}^d \\ C &: \mathcal{V} \rightarrow \mathcal{Y} \\ f &= C \circ V : \mathcal{X} \rightarrow \mathcal{Y}, \end{aligned} \tag{1}$$

where V is the feature extractor and C is the inference function from the features.

This representation separates machine learning models into feature extraction and inference components. In Fig. 3, we show this representation for a classification model where features are concentrated in the same feature space region.

The model f has been trained and tested in all $(\mathbb{X}_{train}, \mathbb{Y}_{train})$ and $(\mathbb{X}_{test}, \mathbb{Y}_{test})$, respectively. Intuitively, the train and test sets are samples of the same distribution in $\mathcal{X}$. Since f has been trained with these data, the features extracted with V are also of the same distribution. In a classification problem, we may consider each class as a different random variable. In our approximation, the features of a sample x, $V(x)$, will be near other samples of the same class and will be separated from the features of other classes. Moreover, in the case of an OOD sample, its features will be separated from all the sample features of the original classes.

In classification tasks we can differentiate between several distributions, one for each class. In Fig. 4, we show the distributions of features and distances in the feature space of two different classes in the feature space. We can model this scenario with several approaches:

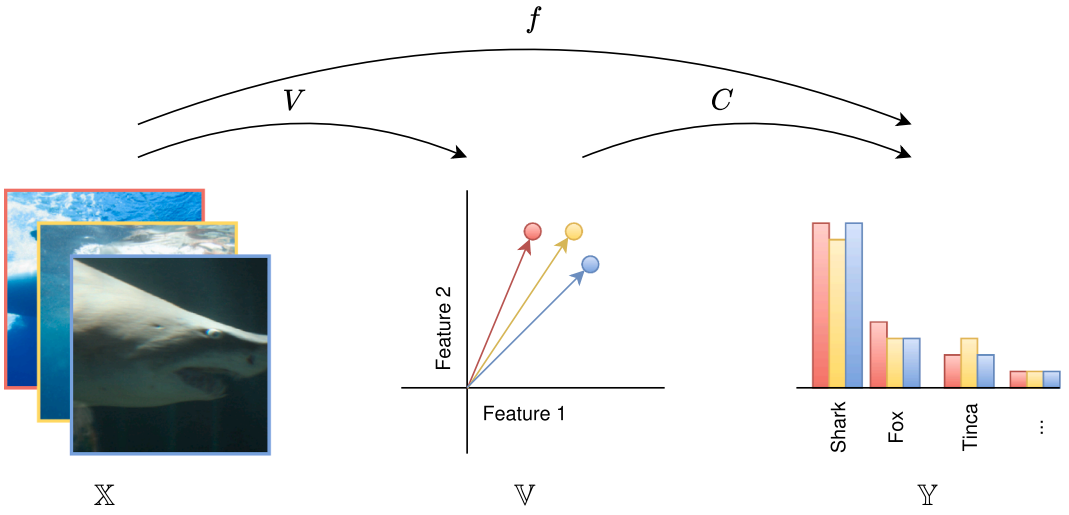
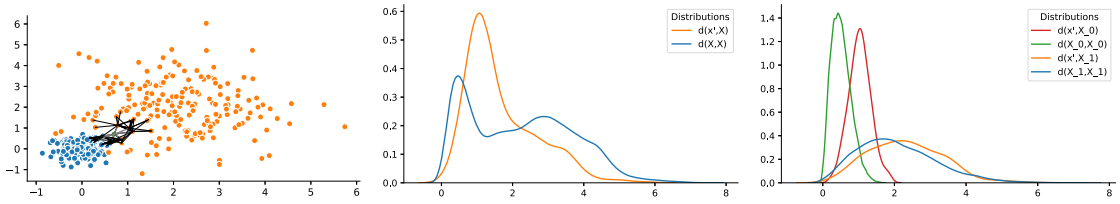


Fig. 3. Representation of the machine learning models into two separated functions V and C.



(a) Feature distribution (b) Distances distribution (c) Distances distributions per class

Fig. 4. Intuition of the behavior in the feature space of ID (orange and blue) and OOD (green) samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

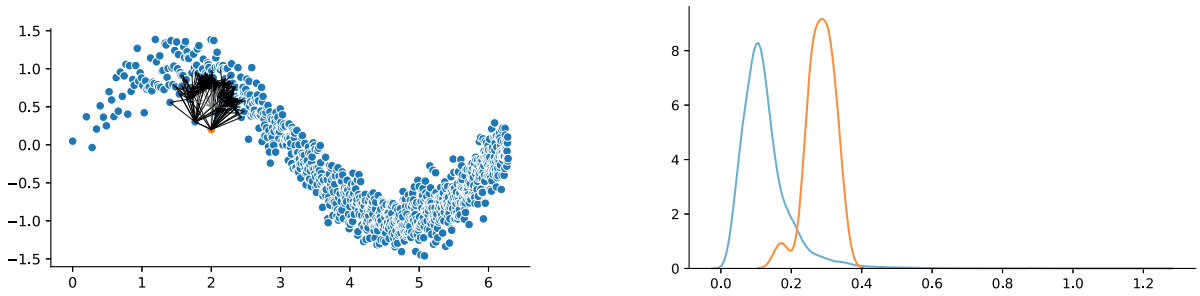
- **Distribution of distances to samples in X_{set}.** This case is the most general. We consider that the whole dataset is part of the same distribution. However, we do not exploit the a priori knowledge concerning the class that the model attributes. In Fig. 4b, we show the distribution of the distances of the OOD sample compared to the distribution of the ID samples. We differentiate between the two distributions, but we would not be able to determine whether one distribution lies above or below the other.
- **Distribution of distances to each sample in X_{set} separated by class.** This case considers distance distributions differently per class. The ID set membership depends on whether the sample x' belongs to one of these distributions. In Fig. 4c, we show the difference between the distances of the OOD sample x' to each class compared to the distances of each class. In this example, we may observe that the distances from x' to class 0 are slightly larger than they might be, but the distances to class 1 are quite similar.
- **Distribution of distances corresponding to the samples in X_{set} whose class is the one selected by the model.** Looking at the distributions in the figure, we can differentiate between the distances to class 0 and class 1. If a model were to classify this OOD sample as class 0, its ID membership could be low because its distances to samples of class 0 are larger than usual. However, the distances to class 1 are equivalent, so one could classify this example as ID with the above approach, even with its misclassification. Therefore, from a present viewpoint, the approach is to use only the class that the model has predicted.

The STOOD-X methodology has two main advantages: it allows us to check whether a sample is OOD or not, and it also allows us to differentiate between samples that are on the decision boundary, that is, almost equally distant from two or more different ID classes samples. These samples, although they belong to the ID distribution, are samples in which the model is not truly confident and may give erroneous inferences. Therefore, we base our proposal on the final prediction class of the model.

Based on the previous argument, and specializing the proposal in distinguishing whether a sample is ID or OOD of an unique distribution, we formalize the following approach: Let $x', x_1, \dots, x_N$ samples in $\mathcal{X}$ where x', x_i are samples of the same distribution and $X_{set} = \{x_1, \dots, x_N\}$ be samples of the distribution $\mathbb{X}$. Then, for a distance $d : V \times V \to [0, \inf)$, we compute the distances of $V(x')$ and $V(x_i)$ for $i \in X_{set}$, that is, $d(V(x'), V(x_i))$. Without loss of generality, we can assume that x_i are sorted with respect to the distance $d(V(x'), V(x_i))$ in ascending order. For x' and $1 \leq k \leq N$, we can define $X_k = x_1, \dots, x_k$ the k NNs to x' with distance $D : \mathbb{X} \times \mathbb{X} \to [0, \inf)$, $D(z_1, z_2) = d(V(z_1), V(z_2))$. Also, for each $x_i \in X_k$ we can define $X_k^i = x_1^i, \dots, x_k^i$ the k NNs of x_i in $X_{set} \setminus \{x_i\}$. Both X_k and X_k^i are the k NNs of x' and x_i, respectively.

For illustration purposes, in Fig. 5 we show an example in which the ID data set is distributed as $(x, sin(x) + \mathcal{N}(0, 0.2)$ (blue) where we have introduced the OOD sample $(2, 0)$ (orange). In Fig. 5a, we notice that most NNs in the OOD sample have their own NNs at a smaller distance on average than the NNs of the OOD sample x'. In Fig. 5b, we confirm this point by showing the distribution of the distances of the different NNs (blue) and the OOD sample (orange).

The main hypothesis of the STOOD-X methodology consists of the following assumption: If x' belongs to the original distribution, there will be no significant differences between the distance distributions $D(x', X_k)$ and $D(x_i, X_k^i)$. As we assume x' and x_i are samples of the same distribution, the set of distances $D(x', X_k)$ and $D(x_i, X_k^i)$ are also samples of the same distribution: *"Distance on the feature space of a sample of X to their k NNs"*. Otherwise, if x' does not belong to the original distribution, the distances $D(x', X_k)$ should be greater than the distances $D(x_i, X_k^i)$.



(a) Samples of $(x, sin(x) + \mathcal{N}(0, 0.2)$ (blue) and an OOD sample(orange). In gray, the connection of the OOD sample and its NNs. In black, the connections of the OOD's NNs and their NNs.

(b) Distribution of distances of a sample of $(x, sin(x) + \mathcal{N}(0, 0.2)$ and its NNs (blue) and distances of the OOD sample and its NNs (orange).

Fig. 5. Illustrative example of a distribution and the distances between samples from the same distribution and an OOD sample.

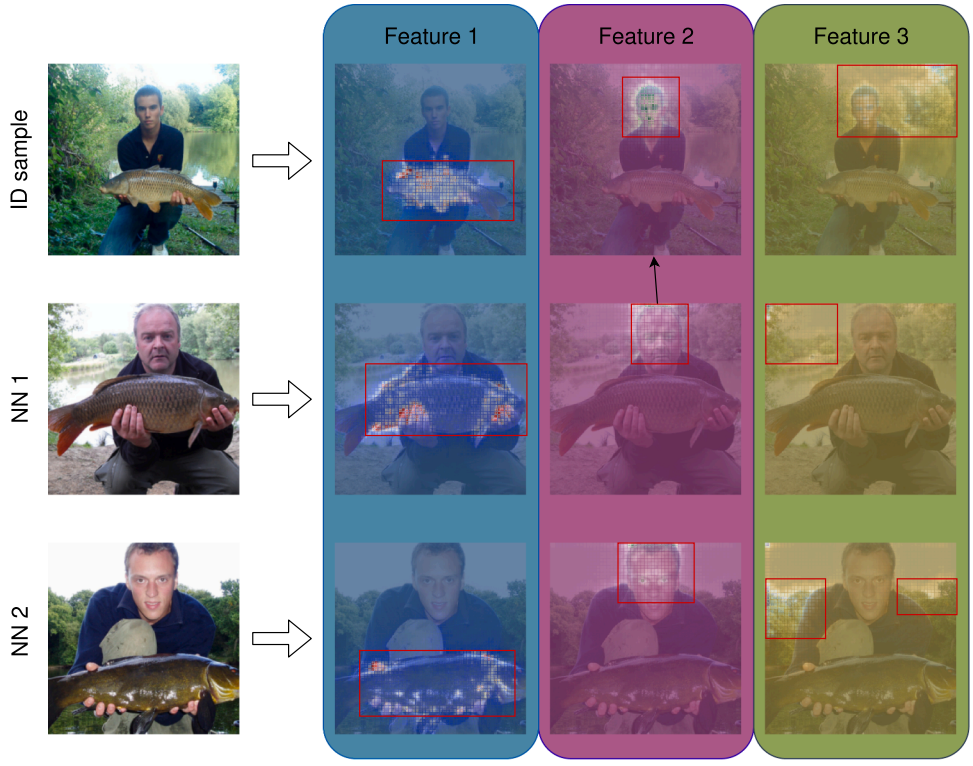


Fig. 6. Example of OOD study results shown to an end user of an ID example with an 52% ID score. Focusing on the 3 most important features, feature 1 highlights the Tinca fish, feature 2 and 3 reveal potential biases.

This translates into the specific case of Fig. 4b. We should differentiate whether the OOD has a larger distance to samples from the original distribution than its NN.

The distribution of features does not necessarily follow any concrete distribution. Therefore, the set of distances $D(x', X_k)$ and $D(x_i, X_k^i)$ also does not necessarily follow any particular distribution. To be able to statistically differentiate between both distributions, we use the nonparametric Wilcoxon-Mann-Whitney test for the differences of two variables in its positive version. The usage of a nonparametric test is justified because parametric tests require normality assumptions that distance data (strictly positive values) typically violate. Shapiro-Wilk testing confirmed non-normality in our feature datasets, and prior research demonstrates that nonparametric tests outperform parametric alternatives for non-normal distributions while showing only marginally inferior performance on normal data [25]. This test takes as null hypothesis that the random variable $X - Y \leq 0$ and as an alternative hypothesis, $X - Y > 0$. The p-value obtained by this test will be the score that the STOOD-X methodology assigns to the new sample. This is a percentage that indicates how likely the new sample is to belong to the original distribution. By selecting a specific significance level, it can be determined whether the sample is ID or OOD.

As shown in Fig. 2, the first stage of the STOOD-X methodology can be summarized as follows in 4 steps:

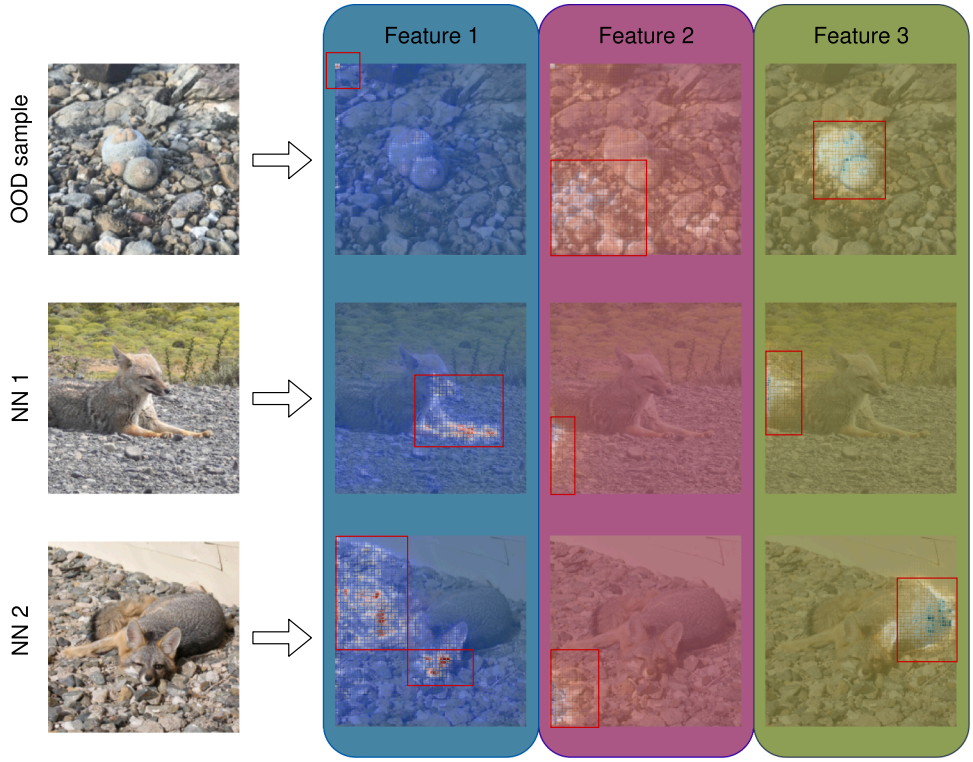


Fig. 7. Example of OOD study results shown to an end user of an OOD example with a 2% ID score. Focusing on the three most important features: Feature 1 and Feature 3 highlight fox-related characteristics in neighboring regions, while Feature 2 reveals background stone textures. The OOD sample lacks a coherent Feature 1 representation and exhibits a misinterpretation of Feature 3.

1. **Feature extraction:** Extract the features of the new sample x' with the feature extractor of our model.
2. **Distance computation:** Compute the distances of the new sample x' to the samples in the train set X_{set} with the distance function D on the feature space.
3. **Nonparametric Statistical Test:** Apply the Wilcoxon-Mann-Whitney test to the distances of the new sample x' to the NNs X_k and the distances of the NNs x_i to their NNs X_k^i.
4. **OOD classification:** Depending on the significance level, classify the new sample x' as ID or OOD. The significance level indicates the confidence with which we want to classify a new sample as ID.

Computational Complexity Analysis: The computational complexity of the STOOD-X methodology can be analyzed by examining each of its main components. For a training set of size N, a feature dimensionality d and K the number of nearest neighbors considered, the algorithm requires:

- Computing cosine distances between the new sample and each training sample requires $O(N \cdot d)$ operations, where d represents the feature dimensionality (typically no more than 1024 for the models considered).
- Sorting the distances to find the K nearest neighbors requires $O(N \log N)$ operations.
- The Wilcoxon-Mann-Whitney test requires $O(M \log M)$ operations, where M represents the total number of distances considered. In our formulation, $M = K + K^2$, as we consider the distances of the new sample to its K nearest neighbors plus the distances of each of these neighbors to their own K nearest neighbors.

The overall computational complexity of the STOOD-X methodology is $O(N \cdot d + N \log N + M \log M)$. Given that $M = K + K^2 \approx K^2$ for large K, and considering that d is typically bounded by the network architecture, the dominant complexity becomes $O(N \log N + K^2 \log K)$ for the distance sorting and statistical testing components, respectively. As $K < N$, the overall complexity can be approximated as $O(N^2 \log N)$.

In summary, this first stage of the STOOD-X methodology provides a principled method for OOD detection by using feature-space distances and statistical tests. Its ability to handle nonparametric distance distributions and its adaptability to various machine learning models make it a versatile tool to improve the reliability of AI systems.

3.3. STOOD-X methodology second stage: eXplanation generation

To justify STOOD-X methodology decisions to users, we study how this methodology explains decisions from a BLUE XAI perspective. The p-value score represents distance from the training distribution, but explanation quality can be enhanced through different explainability techniques.

As an example-based explanation, we can show NNs within the class so that a user can evaluate whether the neighbors have similar features. The STOOD-X methodology score is calculated based on how far the new sample is from the set of trains in the feature space. Therefore, showing the NNs is a good explanation of the features that make the algorithm increase or decrease the score.

As an explanation based on the importance of the features, we can show the locations of crucial details for the decision of the model on the new sample. Based on algorithms such as PCX, we can also analyze the presence of concepts studied a priori within the new sample, distilling the final decision of the algorithm into separate features analysis of the new sample.

By the nature of the STOOD-X methodology, the combination of both perspectives can provide more explanation for OOD detection. We can show the local explanations separately per feature of the new sample while we show the explanations of the NNs. In this way, the final user can qualitatively evaluate whether the explanations offered for each feature are similar to those offered for the NNs.

Table 1
Specifications of the benchmark scenarios provided by OpenOOD.

ID dataset	Near OOD	Far OOD
CIFAR10 [27]	CIFAR-100, ImageNet200	MNIST, SVHN, Textures, Places365
CIFAR100 [28]	CIFAR-10, ImageNet200	MNIST, SVHN, Textures, Places365
ImageNet200 [29]	SSB-hard, NINCO	iNaturalist, Textures, OpenImage-O
ImageNet1K [30]	SSB-hard, NINCO	iNaturalist, Textures, OpenImage-O

To justify the OOD score obtained, the STOOD-X methodology proposes to combine the explanations of importance of the features of the new sample and the NN of the new example explanations, thus categorizing this explanation as a BLUE XAI proposal. As shown in Fig. 2, the second stage of the STOOD-X methodology is based on the following two steps:

1. **Finding NNs in the feature space:** We select the train samples with the closest features in the feature space of the STOOD-X methodology formulation. Since the STOOD-X methodology score is based on the distances to train samples on the feature space, the closest samples must have similar visualizations if classified as ID. The final user will be able to establish relationships within those samples and checking whether the closest samples have similar visualizations, thus justifying the score proposed by the STOOD-X methodology first stage.
2. **Feature Visualization Framework:** For the most present features in these train samples, the STOOD-X methodology visualizations show their feature importance localization from the train samples and the new sample. This step is essential to confirm whether specific features of the new sample correspond to biases or to features present in the original dataset. It is in this visualization where we can establish whether we validate the features learned by our model. Alternatively, we can determine whether the model introduces a bias that should be avoided.

This explanatory stage is valuable in uncertain cases, presenting explanations as queries to users for collaborative decision-making.

4. Experimental setup

This section provides a detailed description of the experimental setup used to evaluate STOOD-X performance in OOD detection. We describe the benchmark selection Section 4.1, neural network architectures Section 4.2, performance metrics Section 4.3, hyperparameters considered Section 4.4, and state-of-the-art comparison algorithms Section 4.5.

4.1. Benchmark selection

To evaluate the performance of our method, we use the OpenOOD framework [26], which provides a variety of ID datasets along with the corresponding OOD datasets. This benchmark collection was curated within the OpenOOD framework to provide standardized evaluation scenarios for OOD detection methods, with particular attention to ensure that classes common to multiple datasets are not used as OOD samples. For each ID dataset, OpenOOD distinguishes between Near and Far OOD datasets based on the degree of similarity between the ID and OOD datasets. We show the specifications of each experimentation scenario in Table 1.

4.2. Neural network architectures selection

The experimental setup involves testing various models, depending on the ID dataset. For the selected benchmarks and post hoc methods to be compared, specific models and weights have been commonly used in the literature for the correct comparison between OOD detection methods.

For CIFAR-10, CIFAR-100, and ImageNet200, we use ResNet18 architecture [31] with OpenOOD pre-trained checkpoints for fair comparison with existing work.

For ImageNet, we use ResNet50 [31] and ViT-B/16 [32] with torchvision pre-trained checkpoints to evaluate across different architectural paradigms.

4.3. Metrics for OOD detection performance evaluation

To perform the performance analysis, we evaluated our method using the metrics most commonly used for OOD detection. There are several proposals for OOD detection, including unsupervised proposals [33]. However, we use the Area Under the Receiver Operating Characteristic curve (AUROC) and the False Positive Rate under a True Positive Rate of 95% (FPR@95), which are the main comparative metrics in the literature. The implementation of both metrics can be found in the OpenOOD framework [26].

4.4. STOOD-X hyperparameters

The design decisions for STOOD-X are guided by the goal of achieving a balance between performance and computational efficiency. The algorithm's scalability depends critically on these hyperparameter choices, particularly the number of neighbors K and the training set size N, which directly affect the computational complexity. In the following, we outline the key hyperparameters and the rationale behind their selection.

- **Distance Metric (d)**: We use the cosine distance as a distance metric for the feature space. The cosine distance is particularly suitable for high-dimensional spaces, such as the feature space.
- **Neighbor Set**: For the set of possible neighbors, we use the training subset. This choice gives us a large number of possible neighbors without introducing data snooping into the experiment.
- **Number of Neighbors (K)**: We first analyze the influence of K, the number of neighbors considered, by testing values such as 9, 18, 36, 72, 144, 288, 500, and N (the size of the entire training set). Our analysis reveals that increasing K improves detection performance, but with diminishing returns beyond a certain point. Based on this trade-off between performance and computation time, we set $K = 500$ for subsequent experiments.
- **Number of important Features (N_f)**: We also analyze the influence of N_f for distance computation. The importance of features is determined by the absolute relevance of the features in the feature layer and is calculated using the zennit-crp library. We tested various percentages of features, including 12.5%, 25%, 37.5%, 50%, 62.5%, 75%, 87.5%, and 100%.

4.5. State of the art OOD algorithm

We compare STOOD-X methodology with state-of-the-art post hoc OOD detection algorithms, selecting top-performing methods from the OpenOOD leaderboard for fair evaluation.

As discussed in Section 2, post hoc algorithms encompass various approaches. The selected algorithms cover a diverse set of these perspectives, ensuring a comprehensive comparison. Table 2 provides an overview of the underlying perspectives for each algorithm included in our evaluation.

5. Evaluating the STOOD-X capabilities in OOD detection

This section analyzes the behavior of the first stage of STOOD-X, which involves the OOD detection algorithm. The analysis is carried out in three steps: two to empirically optimize STOOD-X and one to evaluate its performance. In Section 5.1, we analyze the influence of the number of neighbors used. In Section 5.2, we evaluate the impact of the number

Table 2
Categorization of state of the art post hoc OOD Detection algorithms.

Method	Distance-Based	Energy-Based	Gradient-Based	Perturbation	Prediction-Based
ASH [18]		Yes		Yes (features)	
CombOOD [8]	Yes				
Gradrm [19]			Yes (input gradients)		
KNN [7]	Yes				
MDS [6]	Yes				
NNGuide [5]	Yes		Yes (auxiliary)		
ODIN [4]			Yes (for perturbation)	Yes (gradient)	Yes
ReAct [34]				Yes	
RMDS [20]	Yes				
TempScaling [35]					Yes
VIM [36]		Yes			

Table 3
Comparison of the number of neighbors (K) in the ImageNet dataset with the ViT-B16 architecture in terms of different metrics of STOOD-X.

K	AUROC Near (↑)	FPR@95 Near(↓)	AUROC Far(↑)	FPR@95 Far(↓)	Time(s in cpu)
9	72.328	100	74.837	100	4
18	74.571	85.428	77.831	76.269	4
36	76.545	74.859	80.791	68.719	5
72	78.545	71.747	85.498	52.618	6
144	80.217	67.328	88.443	41.972	10
288	81.252	63.954	90.31	35.855	19
500	81.681	**62.001**	91.196	33.213	30
5000	**81.899**	62.11	**92.181**	**30.167**	269

of features used to calculate the distances. Finally, in Section 5.3, we compare the performance of the optimized STOOD-X with other state-of-the-art algorithms for OOD detection.

5.1. Influence of the number of neighbors K

In this section, we investigate the impact of varying the number of neighbors, K, on the performance of STOOD-X for detection. By testing different values of K in the ImageNet dataset with ViT-B16 architecture, we aim to determine the optimal number of NNs that balance model performance with computational efficiency. The results are presented in Table 3, where we compare different K values in various performance metrics.

Upon reviewing the performance metrics, we observe a clear trend in the relationship between the number of neighbors and the model's performance in the AUROC and FPR@95 metrics. As K increases, both metrics show a noticeable improvement in the detection of near- and far-OOD samples. This suggests that considering a larger number of neighbors helps the model better capture the underlying structure of the data. This thereby enhances its ability to distinguish between ID and OOD samples. However, as highlighted in the time column of Table 3, the computational time increases significantly with larger values of K. This is expected as evaluating a greater number of neighbors requires more calculations. This directly impacts the execution time of the method.

Provided that there is a trade-off between performance and computational efficiency, we conducted a detailed analysis to find an optimal balance. Although increasing K results in higher performance, the improvement decreases after a certain threshold. For example, the jump in quality from $K = 500$ to $K = 5000$ the maximum number of NNs is relatively small compared to the large increase in execution time. Based on these observations, and to decrease the complexity of the algorithm as mentioned previously in Section 3.2, we decided to select $K = 500$ for the rest of our experiments. This choice strikes a reasonable balance, offering satisfactory performance with a manageable computational cost, avoiding excessive time consumption for minimal gains in quality.

5.2. Influence of the percentage of features considered

In this section, we explore the impact of the number of features selected on the performance of STOOD-X, specifically analyzing how excluding less important features affects detection accuracy. We evaluate STOOD-X using different percentages of features and compare its performance across several datasets provided by OpenOOD. Furthermore, we test two different architectures for ImageNet: ResNet50 and ViT-B16, to examine how the number of features influences both neural network structures. These architectures are chosen because they are commonly used in the OpenOOD framework, which provides a standardized basis for comparison. We show the results in Table 4.

From the results on the CIFAR-10 and CIFAR-100 datasets using the ResNet18 architecture, we observe a consistent trend: reducing the number of features used to calculate the distance between neighbors leads to a deterioration in performance. This suggests that the feature space for each class is dispersed across various dimensions, and removing any of these dimensions negatively impacts the model's ability to detect OOD examples. In particular, performance does not degrade significantly when only 12.5% of the most important features are excluded, implying that the remaining 87.5% of the features considered less important do not substantially affect the distinction between ID and OOD samples.

However, the performance on ImageNet200 (using ResNet18), as well as on ImageNet with the ResNet50 and ViT-B16 architectures, presents an interesting countertrend: reducing the number of features slightly enhances performance. This result is initially counterintuitive, as the removal of features would typically lead to a loss of information, as in the previous experiment. However, it suggests that many of the features in these neural networks are redundant or even detrimental to the OOD detection task. Specifically, the less important features seem to introduce noise, negatively impacting the model's ability to differentiate between ID and OOD samples.

This finding also has notable implications from an XAI perspective. By reducing the number of features considered, the complexity of the model is reduced, making it easier to explain the decisions made by the

Table 4
Performance of STOOD-X on the datasets provided in OpenOOD with different percentage of features used.

	CIFAR10: 95,22%		CIFAR100: 77.17%		ImageNet200: 86.38%		ImageNet (R50): 80.38%		ImageNet (ViT): 81.14%	
Nº Features	Near	Far	Near	Far	Near	Far	Near	Far	Near	Far
100.0%	**89.527**	**91.884**	**80.363**	81.179	82.004	**90.538**	**77.66**	85.308	81.51	90.742
87.5%	89.520	91.853	80.347	81.165	82.048	90.527	77.652	85.304	81.521	90.806
75.0%	89.433	91.772	80.255	**81.215**	82.076	90.454	77.649	85.301	81.548	90.892
62.5%	89.267	91.553	79.99	81.079	82.151	90.397	77.647	85.294	81.576	90.947
50.0%	88.833	90.912	79.662	80.929	**82.172**	90.29	77.649	85.294	81.63	91.065
37.5%	88.305	89.981	79.211	80.301	82.16	90.112	**77.66**	**85.311**	81.681	91.196
25.0%	88.118	89.531	78.728	79.1	82.104	89.943	77.65	85.264	**81.695**	**91.241**
12.5%	87.926	89.197	77.597	77.946	81.822	89.687	77.538	84.989	81.574	91.118

Table 5
Near and Far AUROC metric of STOOD-X. *Implemented by our team, due to the absence of a suitable pre-existing OpenOOD implementation.

	CIFAR10		CIFAR100		ImageNet200		ImageNet ResNet50		ImageNet ViT B16	
Method	Near	Far	Near	Far	Near	Far	Near	Far	Near	Far
ASH [18]	74.111	78.360	78.394	79.701	82.119	94.226	36.312	30.469	53.206	51.555
CombOOD [8]*	90.101	92.759	80.708	**82.947**	83.348	90.519	80.512	88.072	79.601	92.652
GradNorm [19]	53.772	58.553	69.734	68.816	73.327	85.293	38.817	32.636	39.281	41.746
KNN [7]	**90.699**	93.105	80.248	82.317	81.750	93.474	70.100	88.640	74.112	90.812
MDS [6]	86.716	90.201	58.794	70.062	62.507	74.939	76.038	**93.473**	79.042	92.599
NNGuide [5]	52.261	46.820	77.089	76.357	76.150	90.683	38.765	53.355	40.906	54.387
ODIN [4]	80.253	87.210	79.798	79.440	80.320	91.897	67.944	67.970	64.306	76.058
React [34]	86.468	91.019	80.705	79.844	80.484	93.096	36.142	36.227	69.261	85.687
RMDS [20]	89.534	92.427	80.27	82.528	82.904	88.54	**80.612**	87.535	80.088	92.6
TempScaling [35]	82.215	87.906	**80.94**	81.421	**85.114**	**94.307**	67.741	75.543	58.896	75.037
VIM [36]	88.506	**93.136**	74.833	82.114	78.814	91.52	64.542	92.112	77.029	**92.837**
STOOD-X	89.527	92.013	80.363	81.215	82.172	90.538	77.660	85.311	**81.948**	92.198

OOD detector. This can be particularly valuable when understanding the rationale behind the model's decisions is crucial.

Based on these observations, we select the optimal percentage of features for each dataset, which corresponds to the configuration that yields the best performance in Near AUROC for each architecture and dataset.

5.3. Performance analysis

We compare STOOD-X against state-of-the-art OOD detection algorithms using AUROC for near- and far-OOD scenarios. Results were obtained through the OpenOOD library when available, with careful implementation of other methods based on their publications. The optimal hyperparameter configurations from the previous analysis were used: $K = 500$ and dataset-specific optimal feature percentages.

For STOOD-X, the optimal configuration of each hyperparameter (number of neighbors K and percentage of features) was selected based on the previous sections: From Section 5.1, we choose $K = 500$ as a balance in performance and computational time, and from Section 5.2, we choose the best number of features considered for each architecture and dataset based on the highest AUROC achieved in the near-OOD scenario.

The proposed method demonstrates robust and competitive performance in all datasets evaluated, consistently ranking among the top performing algorithms. The results summarized in Table 5 highlight that STOOD-X is on par with or superior to several state-of-the-art approaches. Here is a breakdown of the results by dataset and architecture:

- **CIFAR10**: STOOD-X achieves AUROC scores of 89.527% in the Near OOD scenario and 92.013% in the Far OOD scenario. These scores are competitive with the best methods such as KNN (90.699%, 93.105%) and CombOOD (90.101%, 92.759%). The results indicate that STOOD-X can effectively handle near- and far-OOD scenarios, making it a robust option for a variety of detection scenarios. Other algorithms, such as ViM with the highest Far AUROC (93,136%) has a worse performance on Near AUROC (88.506%), realizing that there is a trade-off between Far and Near scenarios.
- **CIFAR100**: In the Near OOD scenario, STOOD-X achieves an AUROC of 80.363%, slightly below top methods like TempScaling (80.94%) and CombOOD (80.708%). In the Far OOD scenario, STOOD-X reaches 81.215%, which is competitive, but still slightly behind methods like CombOOD (82.947%) and KNN (82.317%). These results indicate that the method performs well in this dataset, although there is room for improvement compared to the leading methods. On this dataset, we notice again a tradeoff between Near and Far scenarios.
- **ImageNet200**: In the Far OOD scenario, STOOD-X achieves an AUROC of 90.538%, outperforming methods such as RMDS (88.54%) but far from the top performing TempScaling (94.307%). In the Near OOD scenario, it reaches 82.172%, which also performs worse than TempScaling (85.114%). These results underline the method's capability to handle complex datasets, demonstrating its strength in high-dimensional spaces with more diverse and challenging OOD examples, though with room for further improvements.
- **ImageNet (ResNet50 architecture)**: In Far scenarios, STOOD-X performs with an 85.294%, behind VIM (92.112%) and MDS (93.473%). However, when we compare with the Near scenarios, the performance achieved by STOOD-X (77.660%) outperforms both methods (ViM: 64.542%, MDS:76.038%). RMDS(80.612% Near, 87.535% Far) and CombOOD(80.512% Near, 88,077% Far) are examples of balance between Near and Far AUROC, with worse performance in the Far scenario but outperforming imbalanced algorithms in the Near scenario. STOOD-X is competitive compared to the balanced algorithms.
- **ImageNet (ViT-B16 architecture)**: STOOD-X achieves 81.948% (Near) and 92.198% (Far), outperforming state-of-the-art balanced algorithms, such as RMDS with 80.088% and 92.6% in the Near and Far scenarios, respectively. Imbalanced algorithms like ViM (77.029% Near, 92.837% Far) or KNN (74.112% Near, 90.812% Far) could achieve a good result in the Far scenarios with worse performance in the Near. The results suggest that STOOD-X is suitable for modern architectures such as transformers, which are gaining prominence in computer vision tasks.

In conclusion, STOOD-X shows strong performance in multiple datasets, maintaining a balance between near- and far-reach scenarios. Although it is competitive with the best performing algorithms, there are specific areas where it can be further optimized, particularly for convolutional networks. STOOD-X excels in more complex detection scenarios, especially with transformer-based models, indicating its versatility and adaptability. Additionally, the number of classes in each dataset may have influenced the quality of the OOD detector, as larger class sets can present more challenging scenarios for detection. Since STOOD-X performs better with the transformer model, it might be worth considering that it is in this scenario where the features extracted from transformers don't follow an easy-to-model distribution where the non-parametric tests maintain their potential. However, in order to confirm this statement, more extensive experimentation than the OpenOOD benchmark offers is required. With further refinement, particularly in feature extraction and architectural adjustments, the proposed method has the potential to be a leading solution for OOD detection in a variety of computer vision tasks.

6. Evaluating the STOOD-X explainability

In this section, we evaluate the explainability capabilities of STOOD-X from a BLUE XAI perspective. Our evaluation examines explanation quality across multiple dimensions: feature relevance alignment, neighbor similarity consistency, and bias detection capability. We use the zennit-crp library to generate feature importance visualizations.

We first provide a comprehensive assessment of explanation quality. We establish three key evaluation criteria:

1. **Feature-Neighbor Consistency**: Comparison of highlighted features between test samples and their nearest neighbors to assess explanation coherence.
2. **Class-Relevant Feature Detection**: Evaluation of whether explanations focus on semantically meaningful features rather than spurious correlations.
3. **Bias Identification Capability**: Assessment of the methodology's ability to reveal problematic feature dependencies through neighbor analysis.

With these criteria in mind, we proceed to evaluate the explainability of STOOD-X through two representative case studies: a mid-confidence ID case and a low-confidence OOD case. As discussed in Section 3.3, these cases occur when the algorithm does not find nearby examples validated by humans, making explanation-driven validation particularly valuable.

In Fig. 6, we present a detailed analysis of a mid-confidence ID case (Tinca fish class, 52% ID score). Quantitatively, and according to the relevances computed by CRP, the three features displayed are responsible of 76,21% of the model's decision: 73,09%, 2.19% and 0.94% respectively. The evaluation reveals:

- **Feature-Neighbor Consistency**: All three analyzed features show consistent activation patterns between the test sample and its two nearest neighbors, indicating coherent feature-based similarity detection.
- **Class-Relevant Feature Detection**: Feature 1 correctly focuses attention on the fish itself, validating semantic relevance. However, Feature 2 highlights the person's head, revealing a bias toward human presence in fish classification. Feature 3 emphasizes background elements, indicating potential spurious correlations.
- **Bias Identification**: The analysis successfully identifies two problematic feature dependencies (human presence and background correlation) that should be excluded from fish classification, showing the methodology's capability to reveal model biases.

In Fig. 7, we present a comprehensive analysis of a clear OOD case (NINCO dataset, classified as fox, 2% ID score). Quantitatively, and according to the relevances computed by CRP, the three features displayed are responsible of 11,90% of the model's decision: 8,55%, 2,84% and 0.52% respectively. The evaluation shows:

- **Feature-Neighbor Consistency**: The analysis reveals clear inconsistencies between the OOD sample and nearest neighbors, with no coherent fox-related features shared across samples, confirming the low ID membership score.
- **Class-Relevant Feature Detection**: Feature 1 attempts to detect fox-like patterns in neighbors, but incorrectly highlights the corner in the OOD sample, revealing its OOD nature. Feature 3 focuses on stone textures rather than fox characteristics, indicating misclassification of irrelevant visual elements.
- **Bias Identification**: The methodology successfully identifies bias toward background stone textures (Feature 2) that appears consistently in neighbors, revealing that the model has learned spurious correlations with environmental context rather than animal-specific features.

This thorough analysis enables users to identify specific model limitations and provides actionable insights for bias mitigation, demonstrating the methodology's capability to reveal problematic feature dependencies that compromise classification reliability.

Both analyses highlight the importance of providing structured explanations for STOOD-X. Through feature analysis and neighbor consistency evaluation, users can understand how and why the model classifies examples in particular ways. This approach identifies potential areas for improvement and biases through consistent evaluation criteria. The framework enables reliable bias detection (e.g., stone textures in the fox sample 7) and structured validation of model decisions (e.g., the feature analysis in the tinca fish sample 6). Therefore, incorporating organized visual explanations in STOOD-X allows users to make more informed decisions and refine model behavior through structured evaluation protocols.

We must highlight the difference in percentage of the model explained by the same number of important features when studying the ID and OOD samples. While the 78% of the decision of the ID sample is explained by the 3 most important features, the OOD sample decision of the 3 most important features can explain only 12%. It should be studied whether this behavior is anecdotal or alternatively, whether it is generalizable and potentially can be used to improve both the model and the OOD detection algorithm.

In summary, STOOD-X stands out for its comprehensive approach to bridging algorithmic decision making and human comprehension, fostering effective human-AI collaboration through structured evaluation frameworks. The explainability evaluation empowers users to understand and refine OOD classifications through consistent analysis criteria, enhancing trust, and facilitating AI integration in critical domains. The methodology's bias detection capabilities and structured feature validation framework enable reliable model auditing and continuous improvements. Furthermore, the organized approach to explanation generation promotes seamless human-machine collaboration through clear visualizations and consistent evaluation protocols, ensuring that AI-driven decisions align with human expertise through rigorous validation processes.

7. Conclusion

STOOD-X presents a significant advance in OOD detection by combining robust statistical analysis with human-centered explainability. Its two-stage framework, which uses nonparametric statistical tests for detection and explainable visualizations for decision support, offers a principled, scalable, and user-interpretable solution. Unlike many existing methods, STOOD-X does not rely on restrictive distributional assumptions and provides statistically meaningful confidence scores through the Wilcoxon-Mann-Whitney test.

Empirical results across diverse benchmarks, including CIFAR and ImageNet datasets, demonstrate the competitive performance of

STOOD-X compared to state-of-the-art post hoc OOD detectors applied to large-scale datasets. The method consistently balances performance across both near- and far-OOD scenarios while maintaining computational efficiency. Moreover, its alignment with the BLUE XAI perspective enhances trust and transparency, offering meaningful concept-driven visual explanations that help uncover both strengths and biases in model behavior.

Beyond technical contributions, STOOD-X highlights the importance of trustworthy and explainable AI in safety-critical domains such as healthcare, finance, and autonomous systems. Its ability to detect and visualize decision-relevant features empowers domain experts to audit and improve model behavior. In practice, STOOD-X tends to perform best under three conditions: (i) the learned feature representations are rich and discriminative (e.g., transformer-based models), so the non-parametric test can exploit complex feature structure; (ii) the training set is representative in the learned feature space and does not embed dominant spurious correlations; and (iii) moderate-to-high neighbor values and feature-selection choices (for example, $K \approx 500$ and a high fraction of relevant features) are used to balance sensitivity and computational cost. In these settings, STOOD-X reliably separates near- and far OOD cases while providing local, human-interpretable explanations. We recommend that practitioners validate the trade-off between K and runtime and confirm explanation visualizations on a small labeled subset of their domain before full deployment.

Despite these contributions, STOOD-X has several limitations that should be acknowledged. The method's computational scalability is constrained by its increase in execution time as the training set size grows, potentially making it prohibitive for very large datasets or real-time applications. Additionally, the explanations provided are based on subjective interpretation by end users, lacking quantitative metrics for systematic evaluation of explanation quality.

Future directions include extending STOOD-X to other modalities (e.g. time series), integrating with active learning pipelines, advancing user interfaces for richer interaction with explanations and providing a statistically grounded and quantitative assessment of the quality of the proposed explanations. With its foundation in statistical rigor and human-centered design, STOOD-X sets a promising path for reliable, transparent, and adaptive AI systems in real-world environments. The source code for STOOD-X is publicly available at: https://github.com/ari-dasci/S-STOOD-X

CRediT authorship contribution statement

Iván Sevillano-García: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization; **Julián Luengo:** Writing – review & editing, Validation, Supervision; **Francisco Herrera:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

F. Herrera is supported by the TSI-100927-2023-1 Project, funded by the Recovery, Transformation, and Resilience Plan from the European Union Next Generation through the Ministry for Digital Transformation and the Civil Service. All authors also receive support from the Spanish Ministry of Science and Technology under project PID2023-150070NB-I00 financed by MCIN/AEI/10.13039/501100011033. Funding for open access charge: Universidad de Granada / CBUA.

References

[1] J. Liu, Z. Shen, Y. He, X. Zhang, R. Xu, H. Yu, P. Cui, Towards out-of-distribution generalization: A survey, (2021). arXiv:2108.13624
[2] J. Tack, S. Mo, J. Jeong, J. Shin, CSI: novelty detection via contrastive learning on distributionally shifted instances, Adv. Neural Inf. Process. Syst. 33 (2020) 11839–11852.
[3] T. DeVries, G.W. Taylor, Learning confidence for out-of-distribution detection in neural networks, (2018). arXiv:1802.04865
[4] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, Proc. Int. Conf. Learn. Represent. (2018).
[5] J. Park, Y.G. Jung, A.B.J. Teoh, Nearest neighbor guidance for out-of-distribution detection, in: Proceedings of the International Conference on Computer Vision, 2023, pp. 1686–1695.
[6] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Adv. Neural Inf. Process. Syst. 31 (2018).
[7] Y. Sun, Y. Ming, X. Zhu, Y. Li, Out-of-distribution detection with deep nearest neighbors, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, 162 of *Proceedings of Machine Learning Research*, Proceedings of Machine Learning Research, 2022, pp. 20827–20840.
[8] M. Rajasekaran, M.S.I. Sajol, F. Berglind, S. Mukhopadhyay, K. Das, COMBOOD: a semiparametric approach for detecting out-of-distribution data for image classification, in: Proceedings of the 2024 Society for Industrial and Applied Mathematics(SIAM) International Conference on Data Mining (SDM), Society for Industrial and Applied Mathematics, 2024, pp. 643–651.
[9] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, et al., Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions, Inf. Fusion 106 (2024) 102301.
[10] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[11] Z. Wang, C. Huang, X. Yao, A roadmap of explainable artificial intelligence: explain to whom, when, what and how?, ACM Trans. Auton. Adapt. Syst. 19 (4) (2024) 1–40.
[12] F. Herrera, Reflections and attentiveness on explainable artificial intelligence (XAI). The journey ahead from criticisms to human-AI collaboration, Inf. Fusion (2025) 103133.
[13] P. Biecek, W. Samek, Position: explain to question not to justify, in: R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, F. Berkenkamp (Eds.), Proceedings of the 41st International Conference on Machine Learning, Volume 235 of Proceedings of Machine Learning Research, PMLR 2024, pp. 3996–4006. https://proceedings.mlr.press/v235/biecek24a.html.
[14] M. Dreyer, R. Achtibat, W. Samek, S. Lapuschkin, Understanding the (extra-)ordinary: validating deep model decisions with prototypical concept-based explanations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024, pp. 3491–3501.
[15] J. Yang, K. Zhou, Y. Li, Z. Liu, Generalized out-of-distribution detection: a survey, Int. J. Comput. Vis. 132 (12) (2024) 5635–5662.
[16] Y. Jia, J. Li, G. Zhao, S. Liu, W. Sun, L. Lin, G. Li, Enhancing out-of-distribution detection via diversified multi-prototype contrastive learning, Pattern Recognit. 161 (2025) 111214.
[17] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, Adv. Neural Inf. Process. Syst. 33 (2020) 21464–21475.
[18] A. Djurisic, N. Bozanic, A. Ashok, R. Liu, Extremely simple activation shaping for out-of-distribution detection, Int. Conf. Learn. Represent. (2022).
[19] R. Huang, A. Geng, Y. Li, On the importance of gradients for detecting distributional shifts in the wild, in: Advances in Neural Information Processing Systems, 2021.
[20] J. Ren, S. Fort, J. Liu, A.G. Roy, S. Padhy, B. Lakshminarayanan, A simple fix to mahalanobis distance for improving near-ood detection, (2021). arXiv:2106.09022
[21] X. Chen, X. Lan, F. Sun, N. Zheng, A boundary based out-of-distribution classifier for generalized zero-shot learning, in: European Conference on Computer Vision, Springer, 2020, pp. 572–588.
[22] M.T. Ribeiro, S. Singh, C. Guestrin, " Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[23] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: an overview, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (2019) 193–209.
[24] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From attribution maps to human-understandable explanations through concept relevance propagation, Nat. Mach. Intell. 5 (9) (2023) 1006-1019. https://doi.org/10.1038/s42256-023-00711-8
[25] A.J. Vickers, Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data, BMC Med. Res. Methodol. 5 (1) (2005) 35.
[26] J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, Y. Li, Z. Liu, Y. Chen, H. Li, OpenOOD v1.5: enhanced benchmark for out-of-Distribution

detection, J. Data-Centric Mach. Learn. Res. (2024). Dataset Certification, https://openreview.net/forum?id=cnnTnJQigs.
[27] A. Krizhevsky, V. Nair, G. Hinton, CIFAR-10 (Canadian institute for advanced research), URL http://www. cs. toronto. edu/kriz/cifar.html 5 (4) (2010) 1.
[28] A. Krizhevsky, Learning multiple layers of features from tiny images, URL http://www. cs. toronto. edu/kriz/cifar. html (2009) 32–33. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.
[29] J. Wu, Q. Zhang, G. Xu, Tiny ImageNet challenge, Technical report (2017).
[30] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.
[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[32] A. Dosovitskiy, An image is worth 16x16 words: transformers for image recognition at scale, Int. Conf. Learn. Represent. (2020). https://arxiv.org/abs/2010.11929.
[33] Y. Zhang, J. Hu, D. Wen, W. Deng, Unsupervised evaluation for out-of-distribution detection, Pattern Recognit. 160 (2025) 111212.
[34] Y. Sun, C. Guo, Y. Li, React: out-of-distribution detection with rectified activations, Adv. Neural Inf. Process. Syst. 34 (2021) 144–157.
[35] K. Xu, R. Chen, G. Franchi, A. Yao, Scaling for training time and post-hoc out-of-distribution detection enhancement, in: The Twelfth International Conference on Learning Representations, 2024.
[36] H. Wang, Z. Li, L. Feng, W. Zhang, Vim: out-of-distribution with virtual-logit matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4921–4930.