UNCERTAINTY AWARE DEEPVIEW: GLOBAL VISUALIZATION OF UNSUPERVISED QUANTIFIABLE UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

While research literature has expanded rapidly regarding explanations in the uncertainty domain, existing methods focus almost exclusively on individual predictions, offering little insight into the distribution of the boundaries that separate confidently modeled regions from unreliable ones. We propose Uncertainty Aware Deepview, the first globally faithful uncertainty visualization technique that is applicable to any type of model, and comes with a robustness estimate. Uncertainty Aware DeepView removes the requirement of an uncertainty output of the model and combines two concepts for the determination of unsupervised quantifiable uncertaintes: An adversarial score estimating the local instability of a sample, and an out-of-distribution score quantifying the distributional anomality of a sample. We demonstrate Uncertainty Aware Deepview's ability to accurately identify out-of-distribution data as well as adversarial examples on different models including Autoencoders, CNNs, Resnets and Vision Transformers on multiple benchmark datasets.

1 INTRODUCTION

Deep learning models are increasingly deployed in settings where knowing *when not to trust a prediction* is as important as the prediction itself Jiang et al. (2018). Consequently, the literature on uncertainty quantification and its explanation has expanded rapidly: Bayesian approximations Gal & Ghahramani (2016), deep ensembles Liu et al. (2019), conformal predictions Angelopoulos & Bates (2021) and calibration methods Kuleshov et al. (2018) equip models with different types of uncertainty estimates, while a growing body of work in explainable AI seeks to communicate these estimates to users Newen & Müller (2022); Bhatt et al. (2021). Yet, existing research focuses mostly on explaining uncertainty of individual predictions in isolation, offering little insight into how uncertainty is structured across an entire dataset Antorán et al. (2020); Watson et al. (2023); Wang et al. (2023). Only a few methods focus on global explanations Schulz et al. (2019); Newen et al. (2025b); Ley et al. (2022), which consider robustness estimates and outliers. A notable exception is the global XAI approach Unsupervised Deepview Newen & Müller (2022), which incorporates robustness estimates but requires the model to provide a probabilistic output in the first place, typically a softmax posterior or an explicit likelihood. Autoencoders and other likelihood-free models do not satisfy this needed prior, despite their widespread use in exactly the unsupervised settings where understanding model uncertainty is most critical, such as anomaly detection Chandola et al. (2009).

In this work, we propose *Uncertainty-Aware DeepView*. This uncertainty visualization technique is applicable to all type of models. The key idea is to remove the requirement of an uncertainty output from the model altogether. To this end, we derive uncertainties from two label-free, model-agnostic scores: an *instability score*, which quantifies the local instability of the model's behavior under bounded input perturbations, for example by consideration of adversarial examples and an *out-of-distribution (OOD) score*, which quantifies the distributional atypicality of a sample with respect to the unlabeled dataset. To this end, we derive the Unsupervised Quantifiable Uncertainty Score (Definition 1) of a model on a datapoint. Compared to the classical concepts of uncertainties Chan et al. (2024), this approach follows a more intuitive understanding of uncertainties, following geometric ideas such as geometric ideas such as local intrinsic dimensionality, which has been

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shown to correlate with both outlierness and susceptibility to adversarial attack Ma et al. (2018), motivating a definition of uncertainty that is behavioral and geometric rather than probabilistic. While expanding the original Deepview methodology Schulz et al. (2019), it still works for scenarios such as classifiers, where our adversarial component further admits an exact instantiation: using different robustness proxies, including formal verification Böing & Müller (2022), we compute a robustness score and thereby capture uncertainty with robustness estimates. This gives a stronger basis for uncertainty estimation compared to purely probabilistic estimates Newen & Müller (2022); Schulz et al. (2019). It is expandable to any new proxies, that measure outlierness and robustness.

The choice is motivated by the success of other methods such as local intrinsic dimensionality that have shown that local behavior can correlate to both outliers and susceptibility to adversarial attack Ma et al. (2018). The key idea of our approach is to remove any requirements on the model in question. To this end, we derive the *Unsupervised Quantifiable Uncertainty* (Definition 1) of a model on a sample from two label-free, model-agnostic scores: an *instability score*, which quantifies the local instability of the model's behavior under bounded input perturbations and an *out-of-distribution (OOD) score*, which quantifies the distributional atypicality of the sample with respect to the unlabeled dataset.

The combined score drives a global visualization. Building on the discriminative dimensionality reduction of DeepView Schulz et al. (2019), we embed the dataset into two dimensions under a Fisher-type metric and evaluate the identical uncertainty score on synthesized samples spanning the visualization plane. The result is a continuous uncertainty depiction in which every depicted value captures a combination of uncertainty as well as a robustness estimate. We demonstrate on multiple benchmark datasets that the resulting combination of the score acts as a faithful proxy for uncertainty in a variety of different network architectures, including Convolutional Neural Networks, a Resnet architecture and a Vision Transformer.

2 Related Work

Explainable artificial intelligence (XAI) comprises countless approaches for local interpretations, such as LIME Ribeiro et al. (2016), Shapley values Lundberg & Lee (2017), Anchors Ribeiro et al. (2018), and extensions based on these approaches. Especially regarding unsupervised learning, the field of explanations is evolving with methods such as EXPLAIN-IT in network traffic analysis Morichetta et al. (2019), work in fault detection Brito et al. (2022), and many other areas. Some explanations even aim for properties such as algorithm agnosticism Ellis et al. (2021). While local explanations have been extensively studied, global explanations, especially in the uncertainty domain are comparably scarce. Examples include Roberts et al. (2025), where they propose a pipeline that uses concept activation vectors (extracted via NMF over a foundation model's activations) to produce both local and global explanations of a classifier's predictive uncertainty.Most often, the explainable AI literature is still based on local properties such as local attribution mechanisms Ibrahim et al. (2019), or by approximating the local area of the complex deep learning decision boundary Guo et al. (2018). One other exception is DeepView Schulz et al. (2019), the first global explanation attempt to depict decision boundaries of a deep neural network. It was later extended to Unsupervised Deepview Newen & Müller (2022), removing the need for labels. However, the approach cannot be applied to autoencoder models because of a missing prior. Unsupervised DeepView requires an uncertainty estimation that is not commonly given for autoencoders, and therefore not straightforward to implement. Another recent work is Roberts et al. (2024), which adapts Deepview to encoder LLMs, but still focuses on classification uncertainty. In this work, we remove the requirement for likelihoods and the focus on classification tasks such that only data input and any model input is required as necessary input for a global explanation output of uncertainties. We provide a baseline for comparison that captures outlierness as well as uncertainties by combining proxies that are known to capture outlierness in data Houle et al. (2018) with uncertainty quantification frameworks.

3 Methodology

Our goal is the definition and visualization of a new notion of uncertainty that (i) applies to any trained model $f\colon \mathcal{X} \to \mathcal{Y}$, regardless of whether f provides a probabilistic output, (ii) requires no labels, and (iii) can be evaluated on arbitrary points $x \in \mathcal{X}$ of the input space. This notion takes

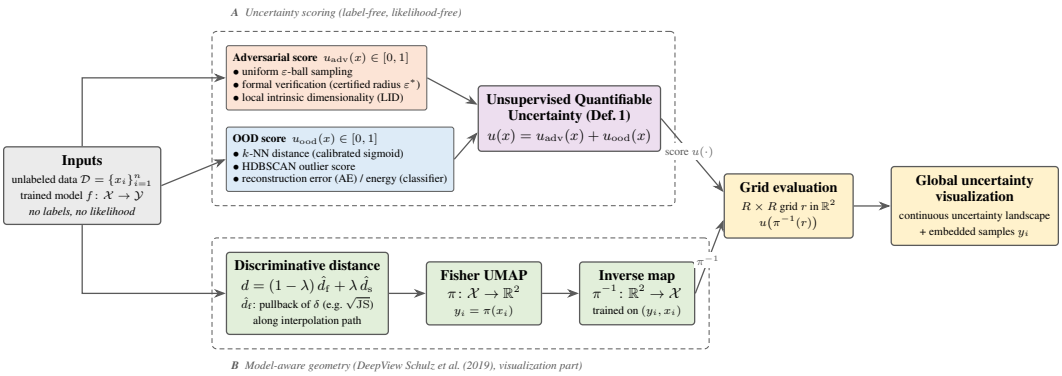


Figure 1: Overview of Uncertainty-Aware DeepView. Given only an unlabeled dataset and a trained model, branch A computes the unsupervised quantifiable uncertainty $u(x)$ (Def. 1) from an adversarial and an OOD component, each instantiated by exchangeable plug-in proxies. Branch B generalizes the DeepView pipeline: a discriminative distance organizes the embedding by model behavior, and an inverse mapping π^{-1} lifts a regular grid in the visualization plane back to input space. Evaluating the identical score $u(\cdot)$ on the lifted grid yields a continuous, globally faithful uncertainty landscape.

into account two aspects: 1. How unstable is the model's behavior in a bounded neighborhood of x? 2. How atypical is x relative to the data distribution the model was trained on? This motivates the definition of Unsupervised Quantifiable Uncertainty. An overview of the methodology is provided in Figure 1.

Definition 1 (Unsupervised Quantifiable Uncertainty)**.** *Let $f\colon \mathcal{X} \to \mathcal{Y}$ be a model and $\mathcal{D} \subset \mathcal{X}$ an unlabeled dataset. For an instability score $u_{\text{adv}}(x) \in [0,1]$, quantifying the local instability of f under bounded input perturbation, an outlier score $u_{\text{ood}}(x) \in [0,1]$, quantifying the distributional atypicality of x with respect to $\mathcal{D}$, and a regularization factor$\lambda \in [0,1]$, we define the* unsupervised quantifiable uncertainty *(UQU) of a sample $x \in \mathcal{X}$ as*

$$u(x) = \lambda \cdot \underbrace{u_{\text{adv}}(x)}_{\textit{instability}} + (1-\lambda) \cdot \underbrace{u_{\text{ood}}(x)}_{\textit{OOD}}. \tag{1}$$

The UQU score $u \in [0,1]$ combines the model's local instability/ sensitivity around x (e.g. adversarial vulnerability) with the out-of-distribution position of x. The higher the UQU score, the more uncertain is the model to the sample. Examples for the proxies are described in the following, among them estimators for robustness and adversarial attacks Ma et al. (2018).

3.1 Label-free Likelihood-Free Uncertainty Scoring:

For every sample $x \in X$ given as input by the user, compute the composite UQU score $u(x)$. In this section we present a list of all proxies for both instability and outlier score currently implemented in our framework. In addition our framework is completely modular and can easily be extended by any other proxy developed in the future.

Instability Score: For the practical implementation, we use the following proxies for the instability score which approximate likelihood of a sample to be an adversarial example. The general idea of all these methods is that adversarial examples often lay in some sort of small 'pocket' Schulz et al. (2019); Feinman et al. (2017) and behave drastically different from the points around them.

- **Uniform sampling:** This proxy describes a sampling-based approach that samples a large number of points (1000 in our case) uniformly random from a ϵ-ball around the given sample. It then compares the networks behavior of the reference sample and the randomly generated points: For a classifier it compares the predicted classes and for an autoencoder it computes the reconstruction errors of the model. Now, the assumption is that for adversarial

examples we will see drastically more cases in which the models output disagree than agree, i.e., different predicted classes for classifiers and vastly different reconstruction errors for autoencoders. Comparing the number of disagreements against a threshold and applying a sigmoid function will then yield a instability score $u_{\text{adv}}(x) \in [0, 1]$ where samples with a high number of disagreements will get a higher score.

- **Local Intrinsic Dimensionality:** LIDs are a common proxy for likelihood of adversarial attack Ma et al. (2018), and therefore implemented in our visualization method as proxy. Again we use a sigmoid function to turn the real-valued LID into a score score $u_{\text{adv}}(x) \in [0, 1]$. Furthermore, we compute the LIDs of the training data and use the 0.95 quantile as a threshold for mapping values above or below a score u_{adv} of 0.5.

- **Maximal verified adversarial robustness:** The notion of adversarial robustness was first introduced by Szegedy et al. Szegedy et al. (2014) and the lack thereof describes a common phenomenon of neural networks that small perturbations to the networks input can cause large changes in the networks output. More formally, a classifier $f\colon \mathcal{X} \to \mathcal{Y}$ is called ϵ-robust for a sample $x \in X$ if the following holds:

$$\forall x'.\|x, x'\| < \epsilon \to argmax(f(x)) = argmax(f(x')).$$

  Centered around the annual Verification of Neural Network Competition (VNNCOMP) Kaulen et al. (2025), a variety of neural network verification tools have been developed over the last years. These tools aim at formally (i.e., mathematically) proving that a given neural network fulfills a certain correctness property like the network being ϵ-robust for some ϵ. The *maximal verified adversarial robustness* of a network and a sample describes the largest value of ϵ for which the network can be proven to be ϵ-robust. The use of the maximal verified adversarial robustness as a proxy is motivated by the assumption that this value is often much smaller for adversarial examples than for regular samples due to the network being less stable in their neighborhood. In our implementation we use the verification tool α, β-CROWN Wang et al. (2021) to find the maximal verified adversarial robustness of a sample.

Outlier Score: For the outlier detection proxies of our framework, any state-of-the-art anomaly detector outputting a score $u_{OOD} \in [0, 1]$ would work and could be additionally integrated Ruff et al. (2021). In summary, our outlier estimation proxies consist of the following methods:

- **k-Nearest Neighbor** estimate outlierness with respect to training data by characterizing each point x by its mean distance to its k nearest neighbors. To obtain a normalized out-of-distribution score $u_{\text{ood}}(x) \in [0, 1]$, we map this distance through a calibrated sigmoid mapping the outliers outside of the 95% quantile Barnett et al. (1994) to a score below 0.5 as uncertain. This approach is a common baseline for outlier detection methods Cover & Hart (1967).

- **HDB-SCAN** McInnes et al. (2017): A well-known state-of-the-art hierarchical density-based clustering algorithm with an implemented outlier score which compares the density at the point of maximum density of the closest cluster in the hierarchy. It is fully non-parametric with respect to the data distribution, does not require a likelihood output from a model nor a parametric density assumption on the data. Furthermore, it takes into consideration the local neighborhood more than DBSCAN Hahsler et al. (2019), which operates on a more global scale and is more closely aligned to the property of local intrinsic dimensionality Ma et al. (2018). The outlier score is a continuous, bounded score, which allows the additive combination of the adversarial and ood score.

- **Reconstruction error:** For models such as autoencoders, the model itself provides a natural outlier proxy An & Cho (2015). Since the autoencoder is trained to minimize the reconstruction loss on the training distribution, it learns a compressed representation of the data manifold; inputs that lie off this manifold may be reconstructed less faithfully, and thus incur a systematically higher reconstruction error. Thus, it is a natural outlier detection proxy for this specific model type, and therefore included in our implementation Sakurada & Yairi (2014).

- **Confidence Score:** For classification models whose outputs are raw logits $f_k(x)$, we obtain a normalized confidence signal by applying the softmax function over the K classes,

$$p(y = k \mid x) \;=\; \frac{\exp\big(f_k(x)\big)}{\sum_{j=1}^{K} \exp\big(f_j(x)\big)}, \tag{2}$$

  which maps the unbounded logits to a probability vector $p(\cdot \mid x) \in [0,1]^K$ with $\sum_k p(y = k \mid x) = 1$. The resulting posterior serves as the model's confidence estimate and provides the softmax-based signal underlying the original DeepView (Schulz et al., 2019); we retain it as one easy-to-compute proxy within our framework.

- **Energy Score:** For classification models, we additionally implement the energy-based out-of distribution score Liu et al. (2020). Given the logits $f_k(x)$ of a classifier, the free energy is defined as

$$E(x; f) \;=\; -T \cdot \log \sum_{k=1}^{K} e^{f_k(x)/T}, \tag{3}$$

  with temperature T (we use $T = 1$). The negative energy $-E(x; f)$ is theoretically aligned with the log of the data density induced by the classifier: in-distribution inputs tend to attain low energy, while out-of-distribution inputs attain high energy Liu et al. (2020).

Choosing k-nearest neighbor Ramaswamy et al. (2000) is an intuitive choice for a strong baseline, which scores a data point by its distance to the training data and remains remarkably strong even against deep detectors Ruff et al. (2021). Its main limitation, a single global distance threshold, motivated new approaches such as DBSCAN Hahsler et al. (2019) or its further advancement HDBSCAN McInnes et al. (2017), which implements a hierarchical density-based cluster hierarchy, essentially removing the drawback that the k-nearest Neighbor approach only performs well in settings where the density deviations are not too large. We implement HDBSCAN as a second proxy for quantifying atypicality. Beyond purely data-driven detectors, the model under inspection can itself serve as an outlier proxy: for autoencoders, the reconstruction error is the classical anomaly signal Sakurada & Yairi (2014); An & Cho (2015), as inputs off the learned data manifold cannot be reconstructed faithfully. There exist other deep hybrid methods such as DAGMM Zong et al. (2018), which combine reconstruction features with a fitted parametric density model. We deliberately avoid such parametric components to keep the framework assumption-free. For classifiers, post-hoc out-of-distribution scores can be derived directly from the model output, ranging from the maximum softmax probability (Hendrycks & Gimpel, 2016), which recovers the original DeepView (Schulz et al., 2019) behavior if desired. We extend its functionality, to the theoretically better-grounded energy score (Liu et al., 2020), which provides the analogous model-based proxy, and by extending beyond classifiers, including likelihood-free models. In combination, the instability score and the outlier score yield the unsupervised quantifiable uncertainty score.

3.2 Discriminative Embedding and Visualization

To realize an uncertainty visualization, we must argue that the discriminative embedding remains feasible in our label- and likelihood-free setting. The visualization is done by learning an inverse mapping from a dimensionality reduction from the inputs, as visible in Figure 1. A visualization that is faithful to the model must arrange samples not by their proximity in input space, but by how similarly they lie with regard to a certain criterion (decision boundary, or in this case, uncertainty). DeepView Schulz et al. (2019) formalizes this idea by equipping the input space with a Riemannian structure: locally, distances are measured by the Fisher information matrix of the class posterior $p(y \mid x)$ Kaski et al. (2001); Peltonen et al. (2004):

$$J(x) \;=\; \mathbb{E}_{p(y|x)}\big[\nabla_x \log p(y \mid x)\, \nabla_x \log p(y \mid x)^\top\big]\,.$$

Directions along which the posterior changes rapidly are assigned large lengths.

This construction is tied to the existence of a predictive distribution $p(y \mid x)$ and is therefore undefined for autoencoders. The underlying principle, however, does not depend on it: the metric is designed to stretch regions in which the *model's output* changes and to contract regions in which it is constant, so that the subsequent embedding is organized by model behavior rather than by input geometry. This principle requires only (a) a discretized path between two samples and (b) a local

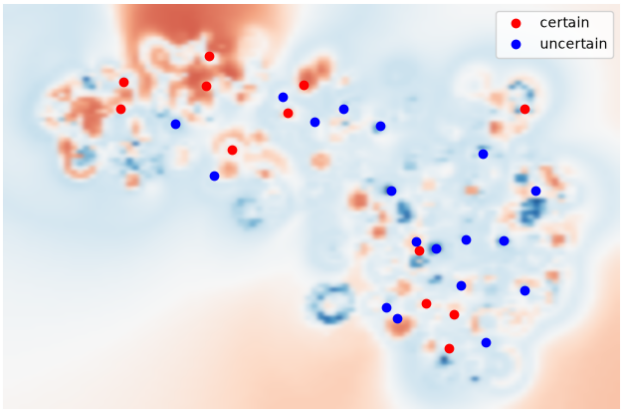


Figure 2: This is an example visualization of the method Uncertainty Aware DeepView on the dataset CIFAR-10 using LIDs as adversarial estimator and confidence score for the uncertainty visualization Ma et al. (2018). Lighter colors or whiteness show that we have no clear information for an estimation in those areas.

dissimilarity δ on the model's output space. We follow the general construction of Deepview Schulz et al. (2019) for the embedding and visualization of the samples. More specifically, our contribution framework proceeds as follows:

1. **Discriminative Distance:** We measure the dissimilarity of two samples x_i, x_j not in input space but by how differently the model f processes them along the straight line connecting them, following the methodology of Deepview Schulz et al. (2019). Details of their setup can be checked within their paper.
2. **Embedding:** Apply the dimensionality reduction technique Fisher UMAP first proposed in Schulz et al. (2019), yielding the two-dimensional projection $\pi\colon \mathcal{X} \to \mathbb{R}^2$ with $y_i = \pi(x_i)$. In theory, this could be replaced with another dimensionality reduction technique.
3. **Inverse Mapping:** Create an inverse mapping $\pi^{-1}\colon \mathbb{R}^2 \to \mathcal{X}$ on the pairs (y_i, x_i), similarly to the original supervised Deepview methodology Schulz et al. (2019). The visualization part is the same, but we introduce the novel unsupervised quantifiable uncertainty score for the alternative construction of uncertainties.
4. **Uncertainty visualization:** Span a regular $R \times R$ grid over the embedding, map the corresponding inputs via π^{-1}. This results in a continuous background showing the estimated binary unsupervised quantifiable uncertainty score.

We replace the coloring component: Rather than drawing class-decision regions, we evaluate the composite unsupervised quantifiable uncertainty (Def. 1) on the identical grid, giving us a more intuitive overview of data points that are likely to lead to problems for the model predictions. An example visualization of our method is depicted in Figure 2. In this Figure, we provide a two-dimensional visualization of the model's UQU in two dimensions: each data sample is assigned to one of two classes, *certain* or *uncertain*, while the continuous background extends these estimates to unseen regions of the input space, indicating where the model is expected to be reliable or unreliable.

4 Experiments

We evaluate Uncertainty-Aware DeepView along the two axes that define its contribution: whether the unsupervised quantifiable uncertainty (UQU) score reliably identifies inputs on which a model is likely to fail, and whether it does so across model types without access to labels or a probabilistic output. To this end, we measure how well the composite score separates clean in-distribution samples from adversarial and out-of-distribution (OOD) inputs, reporting a composite detection accuracy together with its clean, adversarial, and OOD components. Because the framework is designed

around exchangeable plug-in proxies, we further study how the choice of instability and outlier proxy affects performance, spanning five benchmark datasets and four architectures. The following paragraphs detail the datasets, the evaluated models, and the resulting observations; qualitative visualizations produced by the method are shown in Figure 2.

Dataset Choices: We experimentally validated our results on the datasets Fashion-MNIST Xiao et al. (2017), MNIST Deng (2012), CIFAR-10 Krizhevsky et al. (2009), SVHN Netzer et al. (2011) and TinyImagenet Le et al. (2015). These choices ensure that our testsets include grayscale lower resolution (MNIST Deng (2012)), as well as natural colored images (CIFAR-10 Krizhevsky et al. (2009)) or potentially different intrinsic dimensionality (SVHN Netzer et al. (2011)). We also make sure to include higher resolution samples with more classes (TinyImagenet Le et al. (2015)), and that we therefore cover a wider variety of standard datasets with different focuses.

Network Architectures: The network architectures are chosen to test the uncertainty visualization capabilities on various model types, proving empirically that the ability to capture unsupervised quantifiable uncertainties is in this case a general capability of the framework itself. We chose popular model representatives, such as Autoencoders Bank et al. (2020), Convolutional Neural Networks LeCun & Bengio (1998) both with and without a softmax layer in order to account for a probability or no probability output. We also include a trained Resnet He et al. (2016), as well as a Vision Transformer (ViT) Dosovitskiy (2020). The Resnet tests whether the discriminative distance and inverse mapping remain faithful for a deep residual architecture of the kind actually deployed in safety-critical vision pipelines, while the ViT removes the convolutional prior entirely: its attention-based processing makes our evaluation pipeline truly model-agnostic.

Adversarial Construction: In order to fairly evaluate our framework, we require a ground-truth set of data including normal data, anomalous data, and adversarial data to test against. For all combinations of models and datasets, we construct a total of 100 adversarial examples using two well-established methods: the Fast Gradient Sign Method (FGSM) Goodfellow et al. (2015) and the Projected Gradient Descent (PGD) attack Madry et al. (2017). Both methods start with a random sample from our evaluation dataset as a reference and perturb it until the perturbed sample is an adversarial for the reference one. For our classifiers this means that the distance between the reference and the perturbed sample is below a set ϵ (we choose $\epsilon = \frac{4}{255}$) but the classes predicted by the network differ. In contrast, adversarial robustness of autoencoders is defined slightly different as they try to reconstruct a given input and do not expose any class posterior. Instead we follow a definition from the literature Ramanaik et al. (2024) and call an autoencoder $f\colon \mathcal{X} \to \mathcal{X}$ robust for a sample $x \in X$ if the following holds:

$$\forall x'.\|x, x'\| < \epsilon \to re(x', f(x')) \leq \delta \cdot re(x, f(x)),$$

where $re(y, f(y))$ describes the reconstruction error between y and $f(y)$ and $\delta \in \mathbb{R}^+$ is a hyperparameter regulating the allowed divergence.

Construction of Outliers: For out-of-distribution data, we repurpose the used datasets: For MNIST and Fashion-MNIST we used 100 random samples of the other dataset as outliers as they share the same 28×28, grayscale input. In a similar way we use samples from CIFAR-10 and SVHN, respectively. For TinyImageNet, we use samples from the MNIST dataset that we scale from 28×28, grayscale images to 64×64, RGB images matching the corresponding input dimension.

Experimental Results For MNIST Deng (2012), Fashion MNIST Xiao et al. (2017), and other datasets, it is clearly visible that especially with sampling as a proxy, we can see 89-95% overall accuracy. This shows the overall validity of the framework, note that the proxy methods are always interchangeable for the current state-of-the-art. Following observations stand out: First, the simplest proxy pairing—a sampling-based instability score together with a k-nearest-neighbor outlier score—is the most robust default across benchmarks, which matches the overall observation that KNN often performs very reliably and close to state-of-the-art despite its lack of complexity Sun et al. (2022) . Second, and consistent with the plug-in design, no single proxy dominates every regime: density-based detection (McInnes et al., 2017) is the most conservative on in-distribution data yet comparatively insensitive to adversarial inputs, confidence-based scoring becomes preferable on the attention-based ViT, and intrinsic-dimensionality scoring (Ma et al., 2018) paired with

Benchmark	KNN		CS		Energy		HDBSCAN	
	LID	Samp.	LID	Samp.	LID	Samp.	LID	Samp.
CIFAR-10/CNN	57	**77.5**	49	74	63	75.75	36.5	71.5
CIFAR-10/CNN-sm	60.25	77.25	56.75	77.75	71.25	**79.75**	36.5	73.5
CIFAR-10/ResNet	86.75	**88**	51.25	84.25	83.75	86.5	37	81.25
CIFAR-10/ViT	39.25	49	71.75	**76.75**	38.75	48.75	36.25	48.25
Fashion-MNIST/CNN	85.25	91	84.75	72.25	83	**91.5**	72	84.75
Fashion-MNIST/CNN-sm	84.5	86.75	84.25	67	**88.75**	74.25	72	86.75
Fashion-MNIST/ResNet	88	**92**	84.25	81.75	88.5	91	72	89.5
Fashion-MNIST/ViT	78	85.25	**87.75**	69.75	80.5	78.5	73	75.75
MNIST/CNN	91.5	**94.75**	93	79.5	92.5	**94.75**	82.25	87.75
MNIST/CNN-sm	91.75	**94.25**	92.5	72.75	50.75	25.5	83.5	85
MNIST/ResNet	92.56	**95.64**	91.79	67.18	92.56	92.05	80	84.62
MNIST/ViT	89.75	**95**	92.75	68	90.5	85	81.5	84.25
SVHN/CNN	83	84.5	75.75	72.5	83.5	**86.5**	67.5	78.25
SVHN/CNN-sm	83.5	**86.75**	77	74.25	86.25	86.5	67.5	79
SVHN/ResNet	91.25	**93**	80.75	70.5	90.75	90.5	67	77.75
SVHN/ViT	79	**82.5**	81.75	76.75	71.75	70.75	67	65
Tiny-ImageNet/CNN	38.75	53.75	65.25	**68**	61.75	65.75	48	60.25
Tiny-ImageNet/CNN-sm	36.75	48	65.5	**68.25**	38.25	46.75	48.25	56.25
Tiny-ImageNet/ResNet	72.25	74.25	70.25	74	75.25	**77.25**	47.75	60.75
Tiny-ImageNet/ViT	34.75	41.25	63.25	**65.25**	35.25	42.25	48	50.5

Table 1: Composite unsupervised quantifiable uncertainty (UQU) detection accuracy (%) of Uncertainty-Aware DeepView across benchmarks. Columns group the score type with the LID and sampling variants (`method1`); best per row in **bold**.

a confidence score is strongest for out-of-distribution detection. Note that this is a strength as well as a limitation: The framework's value therefore lies in its´ability to match the proxy to the failure mode and architecture at hand, rather than in any fixed configuration, but it also gives the freedom to interchange and update the method in the future to any state-of-the-art proxies available to the user. A more detailed breakdown of these results, including the split into normal, adversarial, and out-of-distribution accuracy, as well as the formal-verification instantiation, which did not terminate for every configuration—is provided in Appendix A.

5 Conclusion and Future Work

In this work, we introduced Uncertainty-Aware Deepview, a globally faithful explanation method that requires neither labels nor probabilistic model output and allows a global evaluation of the potential unreliability of a model on a specific dataset in safety-critical, applied contexts Khedher et al. (2023). For this purpose, we introduce Unsupervised Quantifiable Uncertainty, a score which incorporates both robustness estimates as well as quantifies atypicality of data. Our framework integrates a variety of different proxies ranging from a models own confidence score or reconstruction error that are basically free to compute, to methods like local intrinsic dimensionality Ma et al. (2018) or methods from formal verification of neural networks that are well studied and highly prevalent in the literature. With its modular design, our framework is extendable and will continue to improve with the state-of-the-art outlier detection methods and detection methods for adversarial examples. It also underlines the interconnectedness of uncertainty quantification with robustness Newen et al. (2025a), and raises the question for future work of whether these connections are specifically visible when both phenomena exploit local geometric properties. Across five datasets and four different model types (convoltutional neural network (CNN), a more compressed version of a CNN LeCun et al. (1998), ResNet He et al. (2015) and Vision transformers Dosovitskiy (2020)), we show the capabilities of the framework and its abilities to detect both the outlier as well as the adversarial components with rising task difficulty, accuracy decreases as expected, less complex datasets such as MNIST perform much more reliably than TinyImageNet, where it is assuredly a hard problem to solve. Compared to previous Deepview versions Newen & Müller (2022); Schulz et al. (2019),

the performance remains stable while removing proxies otherwise needed for the construction of the visualization. Most importantly, by instantiating the same score on autoencoders through reconstruction error and formal verification, Uncertainty-Aware DeepView delivers a globally faithful uncertainty visualization for likelihood-free models, a setting no previous DeepView variant could address. We turn what was a structural limitation of prior work into the central capability of ours.

6 Reproducibility Statement:

AI use statement

This paper used Claude as a grammatical checker, and for the generation of the Overview Image (Figure 1). All content is human input, and can be defended by the authors. Claude was used in the code generation process, but all code was manually reviewed by at least one or two of the authors.

Ethics statement

In this paper, we discuss a general framework for explainable AI in a safety-critical context, however no human studies were performed. Our explanation algorithm aim to uncover potential adversarial attacks or other malicious behavior.

Reproducibility statement

The code will be published with the acceptance of the paper. Besides the main pipeline of our approach it contains all scripts used to produce the evaluation data and results presented in the paper (see Section 4). The construction of the adversarial examples for evaluation purposes, the train-test-eval split of the datasets and the shuffling of evaluation data where all conducted with a fixed random seed (seed=42) to allow easy reproducibility of all results. We use the α, β-CROWN verification tool in our experiments that is publicly available on GitHub. All datasets and methods used as proxies in the paper are cited with their corresponding publication and publicly available.

References

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

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

Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a clue: A method for explaining uncertainty estimates. *arXiv preprint arXiv:2006.06848*, 2020.

Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *CoRR*, abs/2003.05991, 2020. URL https://arxiv.org/abs/2003.05991.

Vic Barnett, Toby Lewis, et al. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

Benedikt Böing and Emmanuel Müller. On training and verifying robust autoencoders. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10. IEEE, 2022.

Lucas C Brito, Gian Antonio Susto, Jorge N Brito, and Marcus AV Duarte. An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. *Mechanical Systems and Signal Processing*, 163:108105, 2022.

Matthew A Chan, Maria J Molina, and Christopher A Metzler. Estimating epistemic and aleatoric uncertainty with a single model. *Advances in Neural Information Processing Systems*, 37:109845–109870, 2024.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Charles A Ellis, Mohammad SE Sendi, Eloy Geenjaar, Sergey M Plis, Robyn L Miller, and Vince D Calhoun. Algorithm-agnostic explainability for unsupervised clustering. *arXiv preprint arXiv:2105.08053*, 2021.

Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner. Detecting adversarial samples from artifacts, 2017. URL https://arxiv.org/abs/1703.00410.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. Lemna: Explaining deep learning based security applications. In *proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pp. 364–379, 2018.

Michael Hahsler, Matthew Piekenbrock, and Derek Doran. dbscan: Fast density-based clustering with r. *Journal of Statistical Software*, 91:1–30, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Michael E Houle, Erich Schubert, and Arthur Zimek. On the correlation between local intrinsic dimensionality and outlierness. In *Similarity Search and Applications: 11th International Conference, SISAP 2018, Lima, Peru, October 7–9, 2018, Proceedings 11*, pp. 177–191. Springer, 2018.

Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 279–287, 2019.

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Samuel Kaski, Janne Sinkkonen, and Jaakko Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12(4):936–947, 2001.

Konstantin Kaulen, Tobias Ladner, Stanley Bak, Christopher Brix, Hai Duong, Thomas Flinkow, Taylor T. Johnson, Lukas Koller, Edoardo Manino, ThanhVu H Nguyen, and Haoze Wu. The 6th international verification of neural networks competition (vnn-comp 2025): Summary and results, 2025. URL https://arxiv.org/abs/2512.19007.

Mohamed Ibn Khedher, Houda Jmila, and Mounim El Yacoubi. On the formal evaluation of the robustness of neural networks and its pivotal relevance for ai-based safety-critical domains. *International Journal of Network Dynamics and Intelligence*, 2(4):100018, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 1998.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Dan Ley, Umang Bhatt, and Adrian Weller. Diverse, global and amortised counterfactual explanations for uncertainty estimates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7390–7398, 2022.

Jeremiah Liu, John Paisley, Marianthi-Anna Kioumourtzoglou, and Brent Coull. Accurate uncertainty estimation and decomposition in ensemble learning. *Advances in neural information processing systems*, 32, 2019.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Leland McInnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

Andrea Morichetta, Pedro Casas, and Marco Mellia. Explain-it: Towards explainable ai for unsupervised network traffic analysis. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, pp. 22–28, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Carina Newen and Emmanuel Müller. Unsupervised deepview: Global explainability of uncertainties for high dimensional data. In *2022 IEEE International Conference on Knowledge Graph (ICKG)*, pp. 196–202. IEEE, 2022.

Carina Newen, Emmanuel Müller, and Albert Newen. Trust and uncertainties: Characterizing trustworthy ai systems within a multidimensional theory of trust. *Topoi*, pp. 1–22, 2025a.

Carina Newen, Sofia Vergara Puccini, and Emmanuel Müller. Certainty attacks using explainability preprocessing. In *International Conference on Big Data Analytics and Knowledge Discovery*, pp. 209–224. Springer, 2025b.

Jaakko Peltonen, Arto Klami, and Samuel Kaski. Improved learning of riemannian metrics for exploratory analysis. *Neural Networks*, 17(8-9):1087–1100, 2004.

Chethan Krishnamurthy Ramanaik, Arjun Roy, and Eirini Ntoutsi. Adversarial robustness of vaes across intersectional subgroups, 2024. URL https://arxiv.org/abs/2407.03864.

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*, pp. 427–438, 2000.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pp. 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Isaac Roberts, Alexander Schulz, Sarah Schröder, Luca Hermes, and Barbara Hammer. Targeted visualization of the backbone of encoder llms. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 82–97. Springer, 2024.

Isaac Roberts, Alexander Schulz, Sarah Schroeder, Fabian Hinder, and Barbara Hammer. Conceptualizing uncertainty: A concept-based approach to explaining uncertainty, 2025. URL https://arxiv.org/abs/2503.03443.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pp. 4–11, 2014.

Alexander Schulz, Fabian Hinder, and Barbara Hammer. Deepview: Visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction. *arXiv preprint arXiv:1909.09154*, 2019.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199.

Hanjing Wang, Dhiraj Joshi, Shiqiang Wang, and Qiang Ji. Gradient-based uncertainty attribution for explainable bayesian deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12044–12053, 2023.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-CROWN: efficient bound propagation with per-neuron split constraints for complete and incomplete neural network robustness verification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 29909–29921, 2021.

David Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. Explaining predictive uncertainty with information theoretic shapley values. *Advances in Neural Information Processing Systems*, 36:7330–7350, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Benchmark	KNN	CS	Energy	HDBSCAN
CIFAR-10/CNN	56	63.33	**64.67**	37.33
CIFAR-10/CNN-sm	63.75	**73.33**	62.67	39.5
CIFAR-10/ResNet	–	**83.33**	80	72
CIFAR-10/ViT	41.25	**75.75**	48	40
CIFAR-10/AE	–	–	–	**44.67**
Fashion-MNIST/CNN	88	79.33	**90**	88
Fashion-MNIST/CNN-sm	**79**	69.5	67.33	75.5
Fashion-MNIST/ResNet	**94.67**	84.67	92	88
Fashion-MNIST/ViT	**91.25**	72.5	77	86.75
Fashion-MNIST/AE	–	–	–	**68**
MNIST/CNN	**97.33**	**97.33**	**97.33**	**97.33**
MNIST/CNN-sm	**94.75**	87.5	94.5	91.75
MNIST/ResNet	**96.67**	94	–	95.33
MNIST/ViT	90	78.75	81.25	**94.5**
MNIST/AE	–	–	–	**60.67**
SVHN/CNN	82	80.67	**84**	76.67
SVHN/CNN-sm	**83.25**	72.5	83	71.5
SVHN/ResNet	**89.33**	–	–	82
SVHN/ViT	68	**79.5**	68.67	66.25
Tiny-ImageNet/CNN-sm	**66**	–	–	–
Tiny-ImageNet/ViT	–	–	–	**57.33**
Tiny-ImageNet/AE	–	–	–	**44.67**

Table 2: Composite UQU detection accuracy (%) of Uncertainty-Aware DeepView using *formal verification* as the instability proxy, paired with each outlier score. "–" marks combinations for which no verification result is available. Best per row in **bold**.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.

A Appendix

Benchmark	KNN		CS		Energy		HDBSCAN		Recon.	
	LID	Samp.	LID	Samp.	LID	Samp.	LID	Samp.	LID	Samp.
CIFAR-10/CNN	80	79.33	73.33	82	82	80	**88**	84	–	–
CIFAR-10/CNN-sm	82.67	83.33	83.33	87.33	82.67	84.67	**88**	86.67	–	–
CIFAR-10/ResNet	80.67	83.33	85.33	**93.33**	79.33	83.33	88	92.67	–	–
CIFAR-10/ViT	85.33	84	46	59.33	86.67	87.33	88	**91.33**	–	–
CIFAR-10/AE	–	–	–	–	–	–	89	**96**	87	95
Fashion-MNIST/CNN	88	86	86	87.33	83.33	86.67	87.33	**88.67**	–	–
Fashion-MNIST/CNN-sm	86	85.33	84.67	**90**	80.67	85.33	87.33	89.33	–	–
Fashion-MNIST/ResNet	79.33	84	86.67	**90**	82	81.33	87.33	89.33	–	–
Fashion-MNIST/ViT	81.33	86.67	76.67	86.67	84.67	91.33	87.33	**94.67**	–	–
Fashion-MNIST/AE	–	–	–	–	–	–	86	**98**	87	**98**
MNIST/CNN	78.67	87.33	86	**99.33**	80	88	84.67	97.33	–	–
MNIST/CNN-sm	78.67	86	86.67	**100**	8.67	12	84.67	97.33	–	–
MNIST/ResNet	81.33	89.33	86	**99.33**	81.33	90	84.67	97.33	–	–
MNIST/ViT	81.33	90	84	**98**	82.67	92.67	84.67	96.67	–	–
MNIST/AE	–	–	–	–	–	–	87	**97**	85	93
SVHN/CNN	82.67	86.67	88.67	**93.33**	84	88.67	88	91.33	–	–
SVHN/CNN-sm	80	86	87.33	**92.67**	78	84	88	**92.67**	–	–
SVHN/ResNet	80.67	86	89.33	**96**	82.67	88.67	88	93.33	–	–
SVHN/ViT	85.33	87.33	75.33	82.67	84	89.33	88	**90.67**	–	–
SVHN/AE	–	–	–	–	–	–	87	**90**	89	89
Tiny-ImageNet/CNN	**88.67**	70.67	34.67	42.67	85.33	66.67	88	72	–	–
Tiny-ImageNet/CNN-sm	84.67	68	33.33	41.33	85.33	66.67	**88**	70.67	–	–
Tiny-ImageNet/ResNet	42.67	45.33	52	61.33	60.67	59.33	**88**	79.33	–	–
Tiny-ImageNet/ViT	81.33	78.67	4	9.33	80	80.67	**88**	84	–	–
Tiny-ImageNet/AE	–	–	–	–	–	–	86	**98**	89	97

Table 3: **Normal (clean-input) accuracy** (%): correct identification of in-distribution samples as reliable. Layout as in Table 1; best per row in **bold**. "–" marks proxies undefined for the model type.

Benchmark	KNN		CS		Energy		HDBSCAN		Recon.	
	LID	Samp.	LID	Samp.	LID	Samp.	LID	Samp.	LID	Samp.
CIFAR-10/CNN	16	**57**	27	**57**	18	**57**	11	56	–	–
CIFAR-10/CNN-sm	24	**58**	22	56	31	55	11	56	–	–
CIFAR-10/ResNet	**78**	**78**	32	64	72	75	13	63	–	–
CIFAR-10/ViT	22	39	**81**	79	21	38	10	30	–	–
CIFAR-10/AE	–	–	–	–	–	–	**1**	0	**1**	0
Fashion-MNIST/CNN	59	86	63	83	57	**87**	16	83	–	–
Fashion-MNIST/CNN-sm	59	88	62	85	85	**89**	16	86	–	–
Fashion-MNIST/ResNet	83	**92**	60	87	81	**92**	16	85	–	–
Fashion-MNIST/ViT	40	62	**89**	85	46	63	20	47	–	–
Fashion-MNIST/AE	–	–	–	–	–	–	**4**	0	**4**	0
MNIST/CNN	98	99	93	82	**100**	**100**	52	73	–	–
MNIST/CNN-sm	**99**	**99**	90	80	41	65	57	65	–	–
MNIST/ResNet	**98.89**	**98.89**	87.78	75.56	**98.89**	97.78	38.89	60	–	–
MNIST/ViT	87	**96**	95	84	88	95	49	63	–	–
MNIST/AE	–	–	–	–	–	–	3	0	**14**	10
SVHN/CNN	65	**76**	35	62	60	75	15	59	–	–
SVHN/CNN-sm	66	82	46	66	82	**83**	15	64	–	–
SVHN/ResNet	**94**	**94**	66	72	92	92	13	62	–	–
SVHN/ViT	42	57	**69**	66	27	51	13	35	–	–
SVHN/AE	–	–	–	–	–	–	7	0	**15**	3
Tiny-ImageNet/CNN	20	38	**59**	58	12	33	13	33	–	–
Tiny-ImageNet/CNN-sm	17	38	**62**	61	14	32	14	34	–	–
Tiny-ImageNet/ResNet	75	**79**	63	66	61	71	12	43	–	–
Tiny-ImageNet/ViT	15	26	**97**	**97**	19	26	13	21	–	–
Tiny-ImageNet/AE	–	–	–	–	–	–	**0**	**0**	**0**	**0**

Table 4: **Adversarial accuracy** (%): correct flagging of adversarial inputs as unreliable. Layout as in Table 1; best per row in **bold**. "–" marks proxies undefined for the model type.

Benchmark	KNN		CS		Energy		HDBSCAN		Recon.	
	LID	Samp.	LID	Samp.	LID	Samp.	LID	Samp.	LID	Samp.
CIFAR-10/CNN	61.33	**89.33**	39.33	77.33	74	84	2	69.33	–	–
CIFAR-10/CNN-sm	62	84	53.33	82.67	86.67	**91.33**	2	72	–	–
CIFAR-10/ResNet	98.67	**99.33**	30	88.67	96	97.33	2	82	–	–
CIFAR-10/ViT	4.67	20.67	91.33	**92.67**	2.67	17.33	2	17.33	–	–
CIFAR-10/AE	–	–	–	–	–	–	1	1	1	**2**
Fashion-MNIST/CNN	**100**	99.33	98	50	**100**	99.33	94	82	–	–
Fashion-MNIST/CNN-sm	**100**	87.33	98.67	32	99.33	53.33	94	84.67	–	–
Fashion-MNIST/ResNet	**100**	**100**	98	70	**100**	**100**	94	92.67	–	–
Fashion-MNIST/ViT	**100**	99.33	98	42.67	99.33	76	94	76	–	–
Fashion-MNIST/AE	–	–	–	–	–	–	92	68	**100**	97
MNIST/CNN	**100**	99.33	**100**	58	**100**	98	**100**	88	–	–
MNIST/CNN-sm	**100**	99.33	**100**	40.67	99.33	12.67	**100**	86	–	–
MNIST/ResNet	**100**	**100**	**100**	30	**100**	90.67	**100**	86.67	–	–
MNIST/ViT	**100**	99.33	**100**	27.33	**100**	70.67	**100**	86	–	–
MNIST/AE	–	–	–	–	–	–	**100**	84	**100**	99
SVHN/CNN	95.33	88	90	58.67	**98.67**	92	82	78	–	–
SVHN/CNN-sm	**98.67**	90.67	87.33	61.33	97.33	91.33	82	75.33	–	–
SVHN/ResNet	**100**	99.33	82	44	98	91.33	82	72.67	–	–
SVHN/ViT	**97.33**	94.67	96.67	78	89.33	65.33	82	59.33	–	–
SVHN/AE	–	–	–	–	–	–	83	46	**87**	74
Tiny-ImageNet/CNN	1.33	47.33	**100**	**100**	71.33	86.67	31.33	66.67	–	–
Tiny-ImageNet/CNN-sm	2	34.67	**100**	**100**	7.33	36.67	31.33	56.67	–	–
Tiny-ImageNet/ResNet	**100**	**100**	93.33	92	99.33	99.33	31.33	54	–	–
Tiny-ImageNet/ViT	1.33	14	**100**	**100**	1.33	14.67	31.33	36.67	–	–
Tiny-ImageNet/AE	–	–	–	–	–	–	**34**	26	5	1

Table 5: **OOD accuracy** (%): correct flagging of out-of-distribution inputs as unreliable. Layout as in Table 1; best per row in **bold**. "–" marks proxies undefined for the model type.