KU LEUVEN

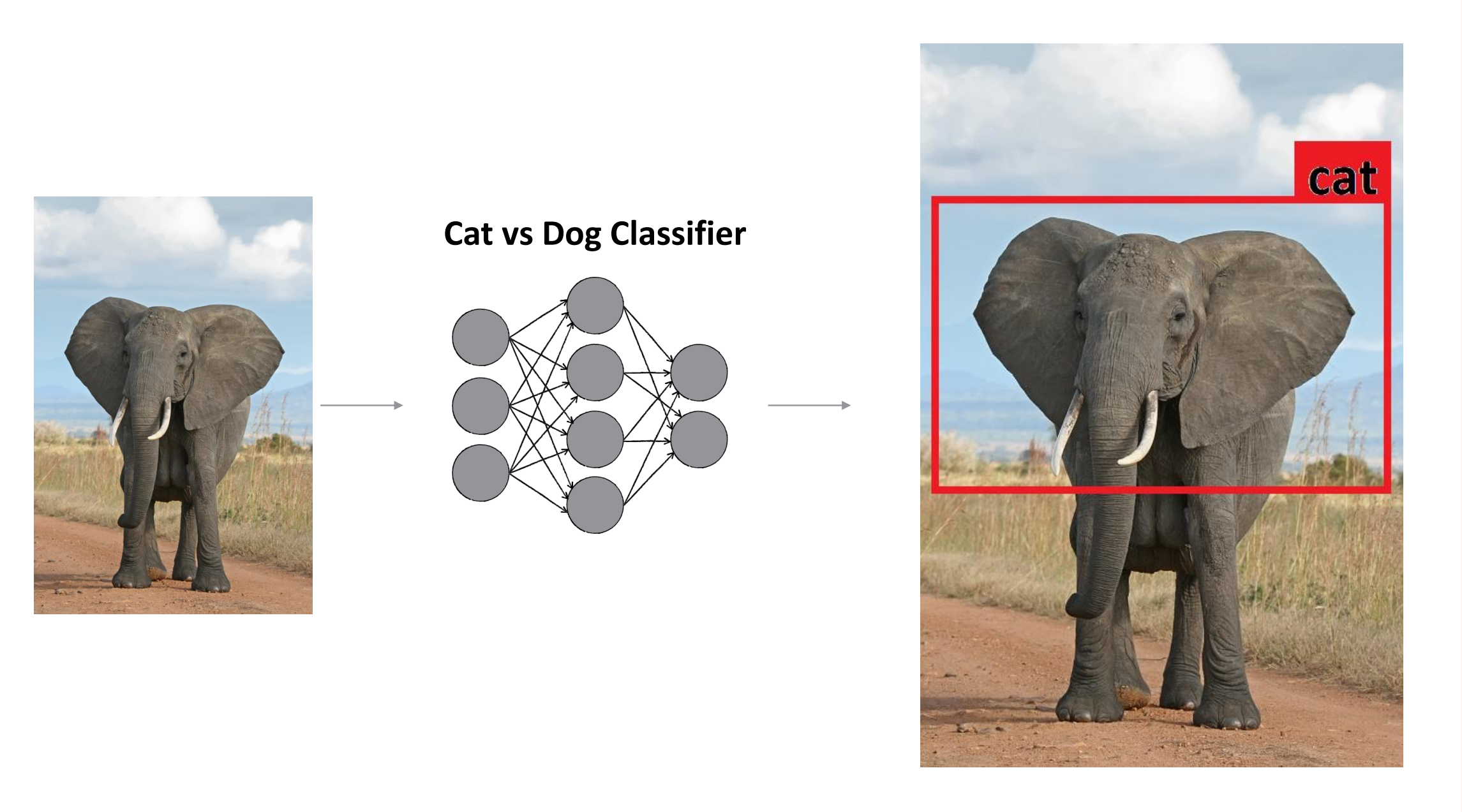
Francesco Tonin, Arun Pandey, Panagiotis Patrinos and Johan A. K. Suykens
Department of Electrical Engineering ESAT-STADIUS, KU Leuven

Unsupervised Energy-based Out-of-distribution Detection using Stiefel-Restricted Kernel Machine

Motivation

- Deploying Deep Learning classifiers in an open-world setting is not trivial: how to detect test samples that do not belong to the training distribution?
- Requirement in safety-critical applications: ML systems should flag potentially **anomalous test samples** so that erroneous predictions.

Example



In the open-world setting, the test distribution can be very different from the training distribution. A user might input an image of an elephant to a cat-vs-dog classifier and an erroneous prediction is made. Instead, we would like the ML system to produce a warning that the given input is out of distribution.

OOD Detection Task

Consider a training dataset $\mathcal{D}_{\text{in}}^{\text{train}}$ drawn i.i.d. from a data distribution P (the *in-distribution*). Let Q be an unknown distribution (the *out-distribution*), from which anomalous examples $\mathcal{D}_{\text{out}}^{\text{test}}$ are drawn.
The **out-of-distribution (OOD) detection task** involves computing an anomaly score $s(x) \in \mathbb{R}$, where $x \in \mathbb{R}^D$ is a test sample.

LEUVEN.AI

STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics

Proposed approach

- New energy-based OOD detector leveraging the St-RKM: the model parameters are learned in an unsupervised manner via manifold optimization where the interconnection matrix U lies on the Stiefel manifold.
- We propose multiple energy function definitions.

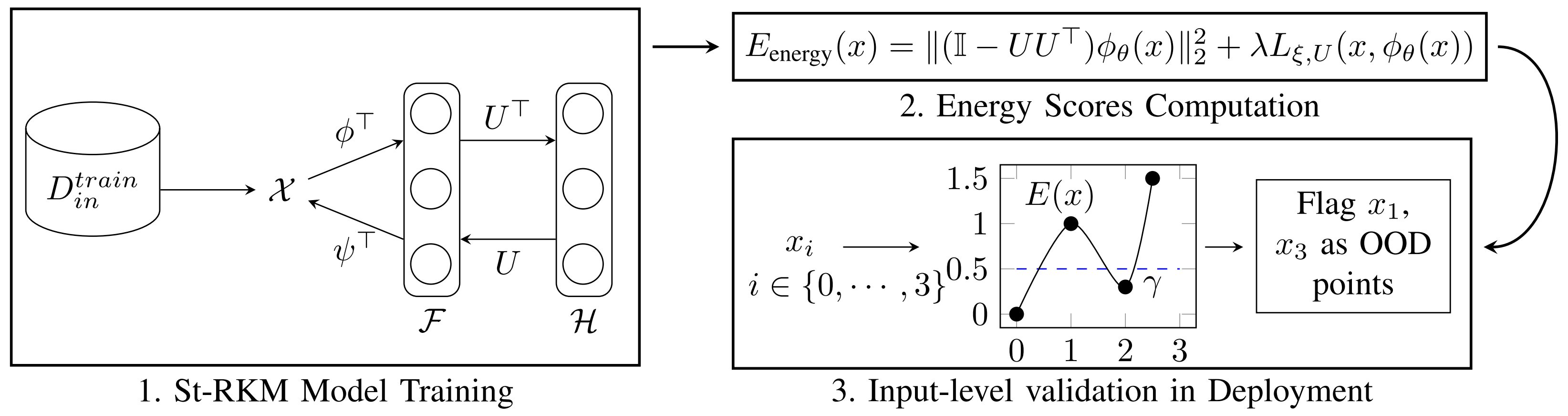

D_{in}^{train}
$\mathcal{X}$
$\phi^\top$
$\psi^\top$
$U^\top$
U
$\mathcal{F}$
$\mathcal{H}$
1. St-RKM Model Training
$E_{\text{energy}}(x) = \|(\mathbb{I} - UU^\top)\phi_\theta(x)\|_2^2 + \lambda L_{\xi,U}(x, \phi_\theta(x))$
2. Energy Scores Computation
x_i
$i \in \{0, \cdots, 3\}$
$E(x)$
γ
1.5
1
0.5
0
0
1
2
3
Flag x_1, x_3 as OOD points
3. Input-level validation in Deployment

Figure: Proposed pipeline of training and detection of OOD points: first, the model is trained on the in-distribution dataset with the full energy function, where $\mathcal{F}$ is the feature space and $\mathcal{H}$ is the latent subspace. Then, the threshold γ is selected such that the scores of 95% of training points are below the threshold value. Lastly, in the evaluation phase, a test sample x_i is passed through the model and its energy score is calculated. If the score is below/above the threshold, the test point is flagged as in/out-of-distribution sample.

Proposed energy functions

$$E_{\text{FullEnergy}}(x) = \|h\|_2^2 - 2\phi_{\theta^\star}^\top(x) U^\star h + \|\phi_{\theta^\star}(x)\|_2^2 + \lambda L_{\xi^\star, U^\star}(x, \phi_{\theta^\star}(x)),$$

$$E_{\text{kPCAError}}(x) = \|h\|_2^2 - 2\phi_{\theta^\star}^\top(x) U^\star h + \|\phi_{\theta^\star}(x)\|_2^2.$$

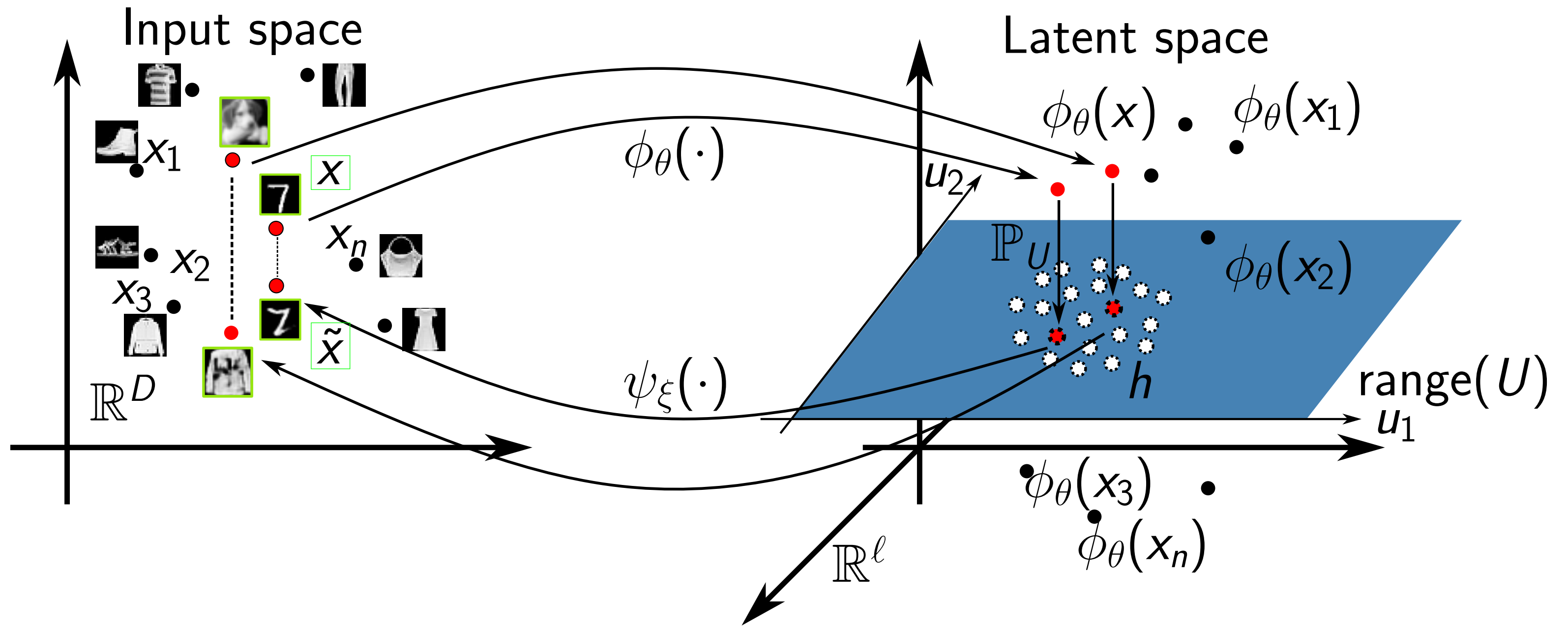


Figure: Illustration with $\mathcal{D}_{\text{in}}^{\text{train}}$ = Fashion-MNIST and $\mathcal{D}_{\text{out}}^{\text{test}} = \{$CIFAR-10, MNIST$\}$. For datasets whose distribution is closer to Fashion-MNIST, the AutoEncoder error is smaller (norm of the dashed line). Hence, the KPCA reconstruction error (norm of the solid line in latent space) becomes a more relevant metric for detecting OOD samples. For datasets whose distribution is dissimilar such as CIFAR-10, both the errors are significant and, hence, the $E_{\text{FullEnergy}}$ becomes more relevant to flag OOD samples.

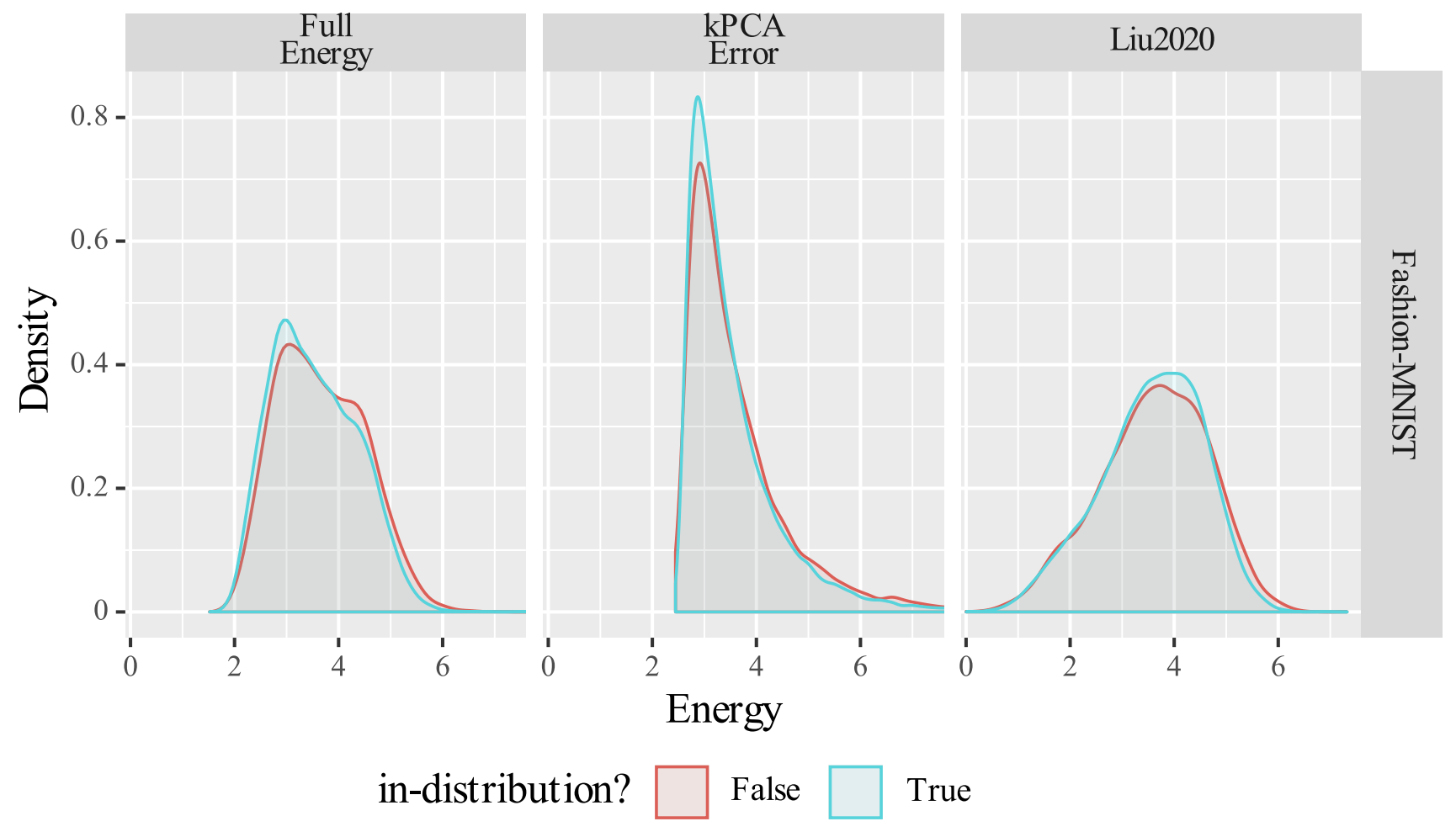


Figure: Distribution of the energy scores when the in/out-distributions are from the same dataset. $\mathcal{D}_{\text{in}}^{\text{train}}$ is the training set of Fashion-MNIST and $\mathcal{D}_{\text{out}}^{\text{test}}$ is the test set of the same dataset.

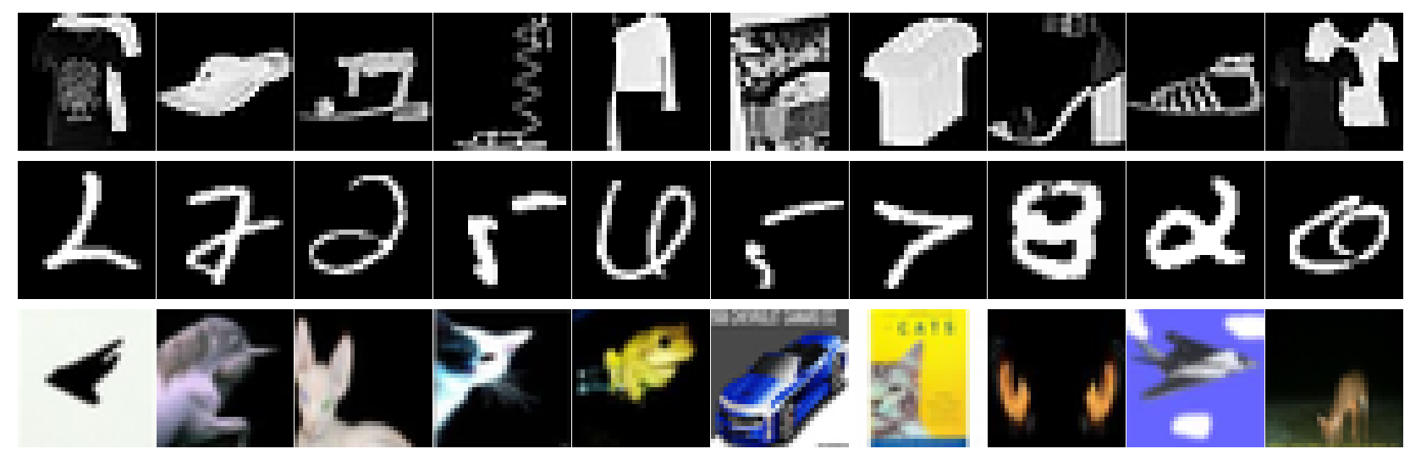


Figure: Samples from the test set of Fashion-MNIST (first row), MNIST (second row), and CIFAR-10 (third row) that are flagged as OOD by our method, illustrating that these samples often show unusual features.

$\mathcal{D}_{\text{in}}^{\text{train}}$: Fashion-MNIST [Mean (Std) over 10 iterations, values in %]

$\mathcal{D}_{\text{out}}^{\text{test}}$	Metric	St-RKM variants [U]		Liu2020 [S]	PCA [U]	VAE [U]
		$E_{\text{FullEnergy}}$	$E_{\text{kPCAError}}$			
MNIST	FPR95(↓)	75.73 (2.7)	**0.38** (0.2)	75.97 (11.1)	99.99	2.67 (1.0)
	AUROC(↑)	69.44 (1.5)	**99.70** (0.1)	78.05 (4.9)	73.17	99.36 (0.1)
	AUPR(↑)	66.26 (3.0)	**99.75** (0.1)	80.36 (4.0)	83.73	99.44 (0.1)
dSprites	FPR95(↓)	99.21 (0.8)	**2.71** (2.7)	96.23 (3.6)	99.79	69.43 (2.3)
	AUROC(↑)	11.61 (3.1)	**99.17** (0.4)	63.68 (7.5)	82.81	85.77 (0.8)
	AUPR(↑)	0.71 (0.02)	**92.82** (2.9)	20.82 (8.2)	70.89	36.87 (6.1)
SVHN	FPR95(↓)	**1.34** (0.2)	28.64 (10.1)	29.14 (8.9)	75.31	27.42 (6.1)
	AUROC(↑)	**99.59** (0.04)	95.61 (1.3)	94.04 (2.0)	51.36	94.56 (1.3)
	AUPR(↑)	**99.23** (0.1)	93.00 (1.8)	88.52 (3.3)	25.57	89.76 (2.4)
CIFAR-10	FPR95(↓)	**0.34** (0.01)	13.40 (5.7)	46.97 (10.4)	65.76	6.50 (2.8)
	AUROC(↑)	**99.76** (0.003)	97.70 (0.8)	90.60 (2.5)	67.86	98.63 (0.4)
	AUPR(↑)	**99.83** (0.003)	98.08 (0.6)	91.77 (2.0)	60.69	98.83 (0.3)

Table: Comparison of OOD detection performance. Lower scores (↓) are better for FPR95 and higher scores (↑) are better for AUROC and AUPR. [S] Supervised / [U] Unsupervised.

Metric	St-RKM variants		VRAE	GAN
	$E_{\text{FullEnergy}}$	$E_{\text{kPCAError}}$		
FPR95(↓)	**6.18** (0.2)	98.45 (1.7)	**6.27** (0.3)	87.45 (19.1)
AUROC(↑)	**94.02** (0.1)	50.39 (1.8)	93.89 (0.2)	37.08 (29.9)
AUPR(↑)	**95.62** (0.2)	85.67 (0.2)	**95.71** (0.2)	78.94 (9.6)

Table: Comparison of OOD detection performance in time series data of electrocardiogram (ECG) sequences. All values are in percentages.

F. Tonin, A. Pandey, P. Patrinos and J. A. K. Suykens. Unsupervised Energy-based Out-of-distribution Detection using Stiefel-Restricted Kernel Machine. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021.