Latent Embedding–Based Isolation Forests for Out-of-Distribution Detection

Michael Onyekwelu, Yooncheol Choi, and Dongweon Yoon
Department of Electronic Engineering, Hanyang University, Seoul, Korea
dwyoon@hanyang.ac.kr

***Abstract*—Non-cooperative communication contexts, such as spectrum surveillance and cognitive radio, increasingly rely on automatic modulation classification (AMC) for intelligent signal processing. However, AMC models without out-of-distribution (OOD) detection risk misclassifying unknown modulations with high confidence. Existing OOD detection methods, including autoencoders, operate under the assumption that the reconstruction loss of OOD inputs is smaller than that of ID inputs. This assumption does not always hold, and when it fails, detection may degrade. To address this, this paper proposes AE-iForest, which integrates an isolation forest (iForest) with the autoencoder's latent embeddings, serving as the feature space for OOD detection. In the proposed method, the iForest isolates OOD signals by recursively partitioning the latent feature space, producing fewer partitions (shorter path lengths) for OOD regions and more partitions for dense ID regions. For evaluation, we adopt a quantile-based thresholding rule on held-out ID samples to retain a fixed proportion of ID, while also considering threshold-free measures of separability. Experiments conducted under two scenarios demonstrate that the proposed method effectively addresses the limitations of reconstruction-loss-based approaches.**

***Index Terms*—Autoencoder, automatic modulation classification, isolation forest, non-cooperative context, out-of-distribution detection**

I. Introduction

Out-of-distribution (OOD) detection is crucial for ensuring the reliability of intelligent signal processing in non-cooperative communication contexts, such as spectrum surveillance and cognitive radio. In these contexts, automatic modulation classification (AMC) serves as a key enabling function, identifying modulation types without prior knowledge of the received signal [1]. While deep learning approaches have significantly improved AMC accuracy, they are often vulnerable to OOD signals, waveforms, or modulation formats unseen during training, which can lead to incorrect predictions with high confidence [2], [3].

Detecting OOD signals in AMC is challenging. Autoencoder (AE)-based methods have been widely proposed for OOD detection, as they rely on the assumption that OOD inputs yield higher reconstruction loss than in-distribution (ID) inputs. However, this assumption often fails, particularly when OOD data closely resemble ID signals, leading to deceptively low reconstruction loss and missed detections [4], [5]. To address this limitation, in this paper, we propose AE-iForest, which integrates an autoencoder's latent embedding space with an isolation forest (iForest) for OOD detection. In AE-iForest, the latent features extracted by the autoencoder serve as input to the iForest, which isolates OOD signals by recursively partitioning the feature space—assigning shorter path lengths to OOD samples and longer paths to ID samples [6]. For evaluation, we adopt a quantile-based thresholding rule on held-out ID samples to retain a fixed proportion of ID while also considering threshold-free measures of separability. Two distinct scenarios are considered, enabling a comprehensive assessment of AE-iForest's performance.

II. Isolation Forest on Latent Embeddings for OOD Detection

We consider the received baseband signal over an observation window of length N, corresponding to the number of complex symbols. The received signal is modeled as

$$x[k] = s[k] + n[k], \quad k \in 0, \ldots, N-1,\ N \in \mathbb{Z}^+, \tag{1}$$

where $s[k]$ denotes the transmitted symbol and $n[k] \sim \mathcal{N}(0,1)$ is additive white Gaussian noise. Collecting the N complex samples yields the vector $\mathbf{x} = [x[0], x[1], \ldots, x[N-1]]^T \in \mathbb{C}^T$. To enable real-valued processing, we define the vectorized form $\tilde{\mathbf{x}} = \mathrm{vec}(\mathbf{x}) = \begin{bmatrix}\Re\{\mathbf{x}\}^T & \Im\{\mathbf{x}\}^T\end{bmatrix}^T \in \mathbb{R}^{2T}$, where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote the real and imaginary parts, respectively. With $\tilde{\mathbf{x}}$ defined in real-valued form, we proceed to train an AE that maps inputs into a lower-dimensional latent space suitable for OOD detection.

An AE is then trained in an unsupervised manner using only ID samples, drawn from the distribution $\mathcal{D}_{\text{ID}}$. The encoder $f_\theta : \mathbb{R}^{2T} \to \mathbb{R}^d$ maps the input $\tilde{\mathbf{x}}$ to a d-dimensional latent embedding $\mathbf{z} = f_\theta(\tilde{\mathbf{x}}) \in \mathbb{R}^d$, and the decoder $g_\theta : \mathbb{R}^d \to \mathbb{R}^{2T}$ reconstructs the input as $\hat{\mathbf{x}} = g_\theta(\mathbf{z})$. The AE is trained by minimizing the squared ℓ_2-norm ($\|\cdot\|_2^2$) reconstruction error

$$\min_\theta\ \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{ID}}}\left[\|\tilde{\mathbf{x}} - g_\theta(f_\theta(\tilde{\mathbf{x}}))\|_2^2\right], \tag{2}$$

which encourages the encoder to produce embeddings that preserve the underlying structure of ID signals. After training, the decoder is discarded, and only the encoder is retained to generate embeddings $\mathbf{z}$, which serve as the feature space for OOD detection.

To score embeddings for OOD detection, we employ an iForest [6], which isolates samples by recursively partitioning the feature space. OOD samples are typically more easily isolated, yielding shorter average path lengths. Using the standard iForest normalization, the OOD score for an embedding $\mathbf{z}$ is defined as

$$s(\mathbf{z}) = 2^{-\frac{\bar{h}(\mathbf{z})}{c(\psi)}} \in (0, 1], \tag{3}$$

where $\bar{h}(\mathbf{z})$ is the average path length of $\mathbf{z}$ across the isolation trees and $c(\psi)$ is the expected path length of a random binary tree of subsample size ψ. Larger values of $s(\mathbf{z})$ indicate a higher likelihood of OOD. After computing the OOD score $s(\mathbf{z})$ for each latent embedding $\mathbf{z} \in \mathbb{R}^d$, we formalize OOD detection as a statistical hypothesis test: $H_0 : \tilde{\mathbf{x}} \sim \mathcal{D}_{\text{ID}}$ (ID signal), $H_1 : \tilde{\mathbf{x}} \sim \mathcal{D}_{\text{OOD}}$ (OOD signal). During evaluation, test samples are drawn from the mixture $\tilde{\mathbf{x}} \sim \gamma\mathcal{D}_{\text{OOD}} + (1-\gamma)\mathcal{D}_{\text{ID}}$, where $\gamma \in (0, 1)$ denotes the OOD proportion. A decision rule is then applied by thresholding the score:

$$\delta(\tilde{\mathbf{x}}) = \mathbb{I}\{s(\mathbf{z}) \geq \tau_\alpha\}, \tag{4}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function and τ_α is chosen from ID validation data to achieve a significance level α (i.e., false positive rate (FPR) $1 - \alpha$). Here, $\delta(\tilde{\mathbf{x}}) = 1$ denotes an OOD decision.

III. Numerical Results and Analysis

We consider single-carrier signals from five modulations: binary phase shift keying (BPSK), quadrature phase shift keying (QPSK), 8-*ary* PSK (8-PSK), 16-*ary* quadrature amplitude modulation (16-QAM), and 64-QAM. Two OOD scenarios are evaluated: (i) Scenario 1 (PSK-vs-QAM): $\mathcal{D}_{\text{ID}}$ = BPSK, QPSK, 8-PSK, $\mathcal{D}_{\text{OOD}}$ = 16-QAM, 64-QAM; (ii) Scenario 2 (odd and even bits per symbol): $\mathcal{D}_{\text{ID}}$ = QPSK, 16-QAM, 64-QAM, $\mathcal{D}_{\text{OOD}}$ = BPSK, 8-PSK. Signals are generated under SNR $\in [-5, 10]$ dB. Each example is a complex IQ sequence $\mathbf{x} \in \mathbb{C}^{1000}$, represented as a real vector $\tilde{\mathbf{x}} \in \mathbb{R}^{2000}$ by concatenating the in-phase and quadrature components. Per scenario, the dataset is partitioned into 32,000 training, 6,000 validation, and 10,000 test sequences. The AE employs mirrored fully connected layers $(2000-1240-245-84-32)$, with bottleneck dimension 2. We train the AE-iForest with learning rate 3.33×10^{-5}, batch size 128, for 30 epochs using Adam, LeakyReLU activations, and mean squared error reconstruction loss; the iForest head uses 500 trees and sets the decision threshold τ_α to the 99% ID quantile.

For performance evaluation, we compare the proposed AE-iForest with a conventional AE (CAE), a variational AE (VAE), and a ResNet50 feature–distance baseline, evaluated at principal components $k \in \{15, 200\}$ [7]. Fig. 1 depicts the AE latent space and sets the stage: in Scenario 1, the ID and OOD clouds are separated, whereas in Scenario 2 they are visibly entangled—signaling a harder detection problem. Building on this geometry, Fig. 2 depicts the iForest decision field trained on the embeddings; the model forms broad inlier basins around dense ID regions, with cleaner margins in Scenario 1 but still meaningful separation in Scenario 2. This partitioning is reflected in the data: Fig. 3 depicts iForest decision–score histograms where ID mass shifts to higher scores and OOD to lower scores, with the contrast again sharper in Scenario 1.

Having established how the detector scores OODs, Fig. 4 connects these observations to performance. In Scenario 1, AE-iForest is nearly perfect ($\approx$ 0.999) and the baselines also perform strongly. In Scenario 2, AE-iForest remains high

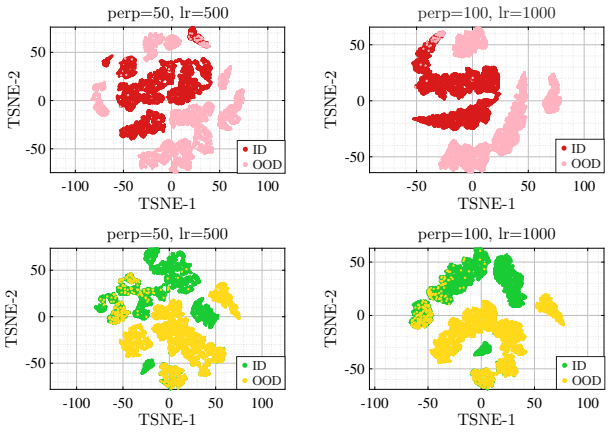

Fig. 1: t-SNE visualization of latent embeddings of the AE. Scenario 1(row 1) and Scenario 2 (row 2).

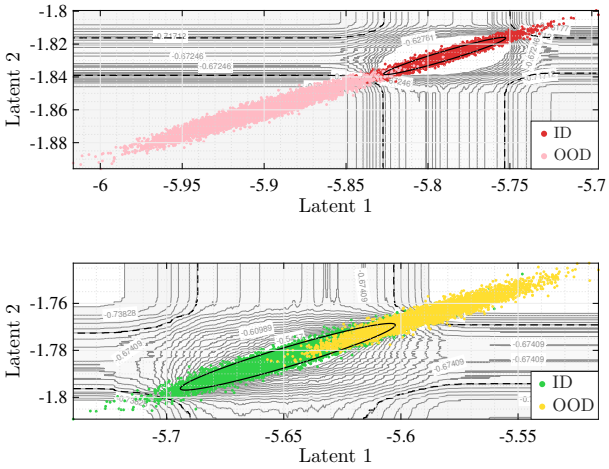

Fig. 2: Isolation-Forest decision fields in 2-D latent space. Top: Scenario 1; Bottom: Scenario 2.

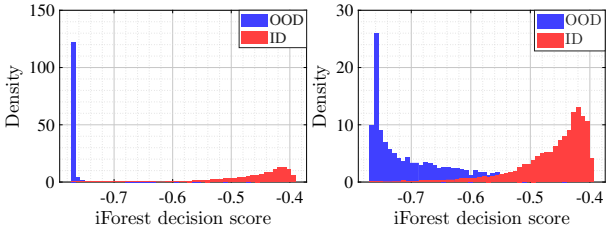

Fig. 3: Histogram distributions for ID and OOD samples for Scenario 1 and 2. Left: Scenario 1; Right: Scenario 2.

($\approx$ 0.97) while CAE/VAE collapse ($\approx$ 0.12) and the ResNet baselines degrade ($\approx$ 0.49 at k=15, $\approx$ 0.61 at k=200). The middle panel includes a CAE loss inset that depicts a loss–ordering inversion—OOD reconstruction loss becomes smaller than ID—explaining why loss–thresholded heads fail, whereas AE-iForest, which scores the latent geometry rather than the loss, remains robust. The right panel further shows that Scenario 1 is intrinsically easier than Scenario 2. To consolidate the evidence, Table I summarizes area under the receiver operating characteristic curve (AUROC), area under

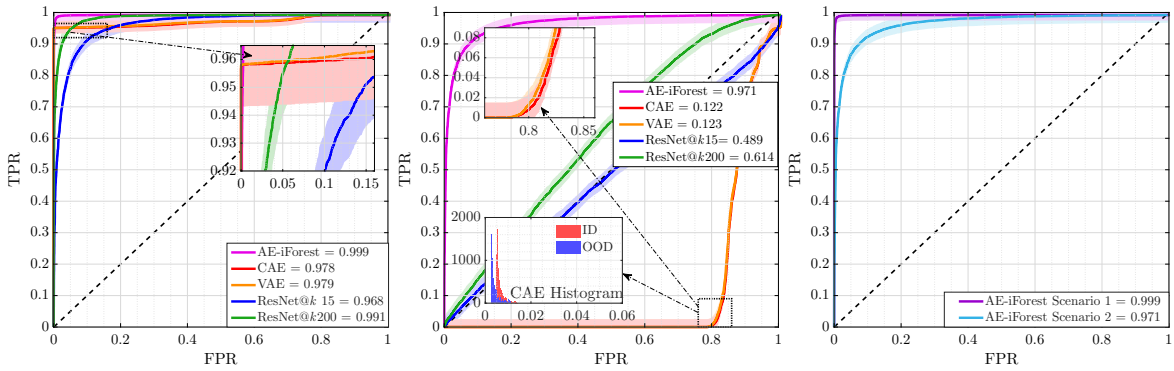

Fig. 4: AUROC: Scenario 1 (left), Scenario 2 (middle), and AE-iForest comparison (right). The middle panel includes a ROC zoom (top-left) and a CAE loss histogram (bottom; ID=red, OOD=blue) illustrating the Scenario 2 loss inversion.

TABLE I: Comparison of OOD Detection Metrics Across Scenarios and Models

Scenario	Method	AUROC	FPR@95TPR	AUPR (In)	AUPR (Out)	TPR	F1-Score
1	AE–iForest	**0.9994**	**0.0010**	**0.9992**	**0.9995**	**0.9930**	**0.9920**
	CAE	0.9783	0.0412	0.9862	0.9661	0.2626	0.4127
	VAE	0.9794	0.0406	0.9869	0.9680	0.2749	0.4279
	RESNET50 (@ k=15)	0.9676	0.1714	0.9659	0.9696	0.6342	0.7714
	RESNET50 (@ k=200)	0.9908	0.0452	0.9906	0.9913	0.8711	0.9262
2	AE–iForest	**0.9659**	**0.1548**	**0.9638**	**0.9647**	**0.5690**	**0.7210**
	CAE	0.1218	1.0000	0.3302	0.3380	0.0000	0.0000
	VAE	0.1230	1.0000	0.3304	0.3385	0.0000	0.0000
	RESNET50 (@ k=15)	0.4888	0.9375	0.5062	0.4697	0.0004	0.0008
	RESNET50 (@ k=200)	0.6145	0.9178	0.5839	0.6468	0.0819	0.1500

the precision–recall curve (AUPR), FPR at 95% True Positive Rate (FPR @95%TPR), TPR, and F1-score: AE-iForest consistently achieves the lowest FPR@95%TPR and the highest F1-score across both scenarios. Finally, Fig. 5 depicts per–class AUCs: results are uniformly strong in Scenario 1; in Scenario 2, AE-iForest stays accurate across BPSK and 8PSK while CAE/VAE remain uniformly poor—fully consistent with the flipping phenomenon observed in the AUROC inset.

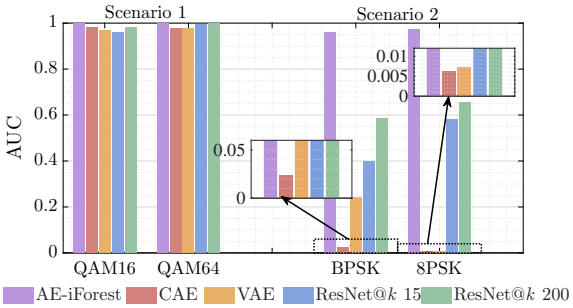

Fig. 5: Per-class AUC for Scenario 1 and Scenario 2.

IV. Conclusion

This paper proposed an OOD detection framework for AMC in non-cooperative communication contexts. The approach integrated an autoencoder's latent embedding space with an iForest, using the embeddings as the feature space for recursive partitioning. Thresholds for decision-making were obtained from held-out ID data using a quantile-based rule that retains a fixed proportion of ID, while performance was also evaluated with threshold-free metrics. Experimental results across two scenarios showed that the proposed AE-iForest framework enhances OOD detection and eliminates the prediction errors commonly observed with conventional autoencoder-based approaches. Per-class OOD AUC analysis further demonstrated consistent detection performance across modulation types.

References

[1] G. Song, M. Jang, and D. Yoon, "Automatic modulation classification for OFDM signals based on CNN with α-softmax loss function," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 60, no. 5, pp. 7491–7497, May 2024.

[2] M. Onyekwelu, M. Jang, and D. Yoon, "Deep learning-based anomaly detection using hybrid loss," in *Proc Int. Conf. Inf. Commun. Technol. Converg.*, Jeju Island, Korea, Oct. 2023, pp. 446–448.

[3] M. Onyekwelu, G. Song, Paulson Eberechukwu N., and D. Yoon, "Out-of-distribution detection for multiple signal sources in connected vehicles," in *Proc Int. Conf. Inf. Commun. Technol. Converg.*, Jeju Island, Korea, Oct. 2024, pp. 1233–1237.

[4] R. Bouman and T. Heskes, "Autoencoders for anomaly detection are unreliable," *arXiv preprint arXiv:2501.13864*, Jan. 2025.

[5] B. Min, J. Yoo, S. Kim, D. Shin, and D. Shin, "Network anomaly detection using memory-augmented deep autoencoder," *IEEE Access*, vol. 9, pp. 104 695–104 706, Jul. 2021.

[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. Int. Conf. Data Mining*, Pisa, Italy, Dec. 2008, pp. 413–422.

[7] M. A. Conn and D. Josyula, "Radio frequency classification and anomaly detection using convolutional neural networks," in *Proc IEEE Radar Conf.*, Boston, MA, USA, Apr. 2019, pp. 1–6.