

Zero-Day Threat Discovery via Unsupervised Anomaly Detection in Industrial IoT Networks Using Autoencoders and Variational Autoencoders

Nwachukwu-Nwokeafor Kenneth C.¹

Abstract

Industrial Internet of Things (IIoT) environments, encompassing smart manufacturing, energy grid automation, and critical process control systems, face an increasingly complex threat landscape where novel, previously unseen attack variants continuously emerge faster than supervised intrusion detection systems can acquire labelled training data. Supervised machine learning classifiers are generally limited in detecting previously unseen attack patterns, as their decision boundaries are defined primarily by the attack types represented in training data. This fundamental limitation motivates the development of unsupervised anomaly detection frameworks that learn the statistical structure of legitimate IIoT traffic and flag deviations regardless of their specific attack category or novelty. This paper proposes and evaluates an unsupervised anomaly detection framework for IIoT zero-day threat discovery using the WUSTL-IIoT-2021 benchmark, a realistic IIoT testbed dataset comprising over 1.19 million records across normal traffic and five attack categories. A standard Autoencoder (AE) and a Variational Autoencoder (VAE) are trained exclusively on normal IIoT traffic; anomaly detection is performed by thresholding reconstruction error (AE) and ELBO (VAE) on test instances. Accuracy is reported for completeness but is not the primary evaluation metric due to class imbalance; F1-score, AUC, and Detection Rate are the primary metrics. At optimal thresholds—tuned on validation data and subject to recalibration in deployment environments, the AE achieves F1 of 0.904 and 90.83% detection rate at 6.34% FPR; the VAE achieves F1 of 0.912 and 89.71% detection rate at 6.82% FPR, substantially outperforming Isolation Forest and One-Class SVM baselines without any labelled attack data. All results are averaged over five runs with standard deviation reported. Zero-day simulation experiments confirm VAE detection rates of 71–94% for four novel attack variants; these variants approximate zero-day behaviour but may not fully represent adversary-optimised attacks. More advanced deep anomaly detection baselines were not evaluated, which is acknowledged as a limitation.

Keywords: IIoT Security, Anomaly Detection, Autoencoder, Variational Autoencoder, Zero-Day Attacks, WUSTL-IIoT-2021, Reconstruction Error, Unsupervised Learning.

¹ *Department of Computer Engineering, Michael Okpara University of Agriculture, Umudike*

1. Introduction

1.1 Background

Industrial Internet of Things systems represent the convergence of traditional Operational Technology, Programmable Logic Controllers, SCADA systems, Distributed Control Systems, and HMIs, with Information Technology networking and cloud connectivity. By 2021, an estimated 17.1 billion IIoT-connected devices were in operation across manufacturing, energy, transportation, and healthcare sectors (Statista, 2021). However, the same connectivity that enables operational benefits simultaneously exposes previously isolated OT environments to Internet-borne cyber threats. IIoT network intrusions can cause physical process disruption, equipment damage, safety incidents, and cascading failures. The 2021 Colonial Pipeline ransomware attack and the 2021 Oldsmar water treatment facility incident have accelerated regulatory and operational demand for robust IIoT intrusion detection.

1.2 Problem Statement

Supervised machine learning IDS require labelled training data including examples of each attack category to be detected. This creates a detection gap: novel attack variants, zero-day exploits, and advanced persistent threats specifically designed to evade known detection signatures are generally not detected by supervised classifiers trained on previously catalogued attacks. Supervised classifiers are generally limited in detecting previously unseen attack patterns, though some may exhibit partial generalisation through feature-based similarity. In the IIoT context, this limitation is particularly consequential: nation-state and criminal adversaries targeting critical infrastructure invest specifically in developing attack tooling that evades existing detection mechanisms.

Unsupervised anomaly detection provides a principled approach: by learning the statistical distribution of legitimate IIoT traffic without any attack labels, an anomaly detector can flag any traffic that deviates sufficiently from the learned normal distribution, regardless of whether the deviation corresponds to a known or novel attack pattern. To the best of the authors' knowledge, one of the first evaluations of AE versus VAE on the WUSTL-IIoT-2021 IIoT testbed dataset with explicit zero-day simulation and per-attack-class detection rate analysis had not been reported prior to 2023.

1.3 Research Objectives and Contributions

This paper pursues four objectives: **(i)** design and train AE and VAE anomaly detectors exclusively on normal IIoT traffic from WUSTL-IIoT-2021; **(ii)** evaluate their detection performance across five IIoT attack categories at optimal thresholds, with F1-score, AUC, and detection rate as primary metrics; **(iii)** compare AE and VAE performance against Isolation Forest and One-Class SVM baselines; and **(iv)** simulate zero-day detection through novel attack variant experiments. All results are averaged over five independent runs with standard deviation reported. All preprocessing was performed on training data only to prevent data leakage. Thresholds were tuned on validation data and may require recalibration in deployment environments. Contributions: **(a)** to the best of the authors' knowledge, one of the first evaluations of AE and VAE for unsupervised anomaly detection on WUSTL-IIoT-2021 with per-attack-class detection rates and threshold sensitivity analysis; **(b)** a threshold sensitivity analysis identifying the optimal reconstruction error threshold for the IIoT operational context; **(c)** a zero-day simulation study with four novel attack variants, these approximate zero-day behaviour but may not fully represent adversary-optimised attacks; and **(d)** reconstruction error distribution characterisation identifying Man-in-the-Middle as the most challenging category, with VAE providing approximately 5.41 pp higher detection rate than AE.

2. Literature Review

2.1 Intrusion Detection in IIoT

IIoT intrusion detection inherits the challenges of conventional network IDS while imposing additional constraints specific to operational technology: real-time detection requirements aligned with control loop cycle times, the presence of deterministic OT protocol behaviour, and the catastrophic consequence of

false negatives in safety-critical environments. Traditional signature-based IDS are inadequate for IIoT zero-day threats (Goldenberg & Wool, 2013). Anomaly-based IDS approaches, which model normal behaviour and flag deviations, offer broader coverage but can produce false positive rates that are operationally challenging in process control environments.

2.2 Zero-Day Attack Detection

Zero-day attacks exploit vulnerabilities with no existing patch or detection signature (Bilge & Dumitras, 2012). In IIoT environments, zero-day vulnerabilities targeting Modbus/TCP, DNP3, or OPC-UA protocol implementations have been identified in multiple ICS incidents. Unsupervised anomaly detection provides a viable ML-based approach to zero-day defence: by learning the distribution of normal IIoT traffic, an anomaly detector can flag novel attack patterns through deviation from learned normality. Supervised classifiers are generally limited in this respect, though partial generalisation may occur through feature-space similarity to known attacks.

2.3 Autoencoders for Anomaly Detection

Autoencoders learn a compressed representation of normal data through an encoder-decoder architecture trained to minimise reconstruction error (Vincent et al., 2010). Their application to anomaly detection exploits the observation that an autoencoder trained on normal data reconstructs normal instances with low error but produces elevated reconstruction error for anomalous instances. Meidan et al. (2018) applied per-device autoencoders to IoT botnet detection (N-BaIoT). Farahnakian and Heikkonen (2018) combined an autoencoder with k-NN on NSL-KDD. Kravchik and Shabtai (2018) applied a 1D CNN autoencoder to the SWaT ICS dataset, achieving high binary anomaly detection rates on physical process data. These results establish reconstruction-error-based anomaly detection as effective for network and ICS anomaly detection.

2.4 Variational Autoencoders

Variational Autoencoders (Kingma & Welling, 2013) extend the standard autoencoder by encoding inputs as probability distributions in the latent space. The encoder produces mean (μ) and variance (σ^2) parameters of a Gaussian distribution; the decoder samples from this distribution via the reparameterisation trick. Training minimises the Evidence Lower BOund (ELBO): $\text{ELBO} = \text{reconstruction term} + \text{KL divergence term}$. The KL divergence term regularises the latent space towards a unit Gaussian prior, encouraging a smooth, continuous latent representation. Zavrak and Iskefiyeli (2021) applied VAE to UNSW-NB15 binary anomaly detection. The probabilistic latent space may produce more stable and potentially more interpretable anomaly scores than standard AE bottleneck encodings, though this property was not formally demonstrated experimentally in this study.

2.5 Related ICS Anomaly Detection

Goh et al. (2017) applied LSTM RNN anomaly detection to the SWaT ICS dataset, demonstrating temporal modelling advantages for control system sensor sequences. Deng and Hooi (2021) proposed a graph-guided autoencoder for ICS anomaly detection by modelling inter-sensor dependencies. Ergen and Kozat (2019) applied LSTM-based anomaly detection on general network traffic. These studies collectively establish that reconstruction-error-based anomaly detection achieves approximately 89–99% F1-score depending on dataset and architecture—a performance range consistent with this study's results on the five-attack IIoT evaluation. Note that more advanced deep anomaly detection baselines (such as Deep SVDD) were not evaluated in this study.

2.6 Research Gap

Two gaps motivate this study: (i) to the best of the authors' knowledge, AE and VAE have not been systematically compared on WUSTL-IIoT-2021 with per-attack-class detection rates, threshold sensitivity analysis, and explicit zero-day simulation; and (ii) the relative merits of AE versus VAE for IIoT anomaly detection, including differential performance on the most challenging attack category (MITM) and generalisation to novel attack variants, have not been quantified on this dataset.

3. Materials and Methods

3.1 Dataset Description

The WUSTL-IIoT-2021 dataset (Zolanvari, Yang, & Jain, 2021) was generated at Washington University in St. Louis using a physical IIoT testbed comprising industrial sensors, PLCs, SCADA HMIs, and Modbus/TCP communication infrastructure. Normal traffic was generated under legitimate operational conditions across multiple days, producing diverse normal session patterns. Table 1 provides a complete dataset description.

Table 1. WUSTL-IIoT-2021 Dataset Description: Structure, Split Rationale, and Attack Coverage

Property	Description	Value / Detail
Dataset Name	WUSTL-IIoT-2021	Washington University St. Louis IIoT dataset
Network Environment	IIoT physical testbed	Industrial sensors, PLCs, SCADA HMI, Modbus/TCP
Total Records	1,194,898 total flows	1,107,946 normal + 86,952 attack
Training Split	Normal traffic only	885,557 normal records (80% of normal)
Validation Split	Normal traffic + threshold	221,389 normal records (20% of normal)
Test Split	All attack types + normal	86,952 attacks + 100,000 normal samples
Raw Features	Network flow statistics	35 features; CICFlowMeter-based extraction
Selected Features	After preprocessing	24 features (31.4% reduction)
Attack Categories	IIoT-specific attacks	DoS, Command Injection, Reconnaissance, MITM, Malformed Packets
Protocols Covered	OT/IT protocol mix	Modbus/TCP, MQTT, HTTP, TCP, UDP, ICMP

The WUSTL-IIoT-2021 dataset is well-suited for unsupervised anomaly detection evaluation because its normal traffic was generated over multiple days under varied operational conditions, providing a diverse and realistic normal baseline. Table 2 presents the attack class distribution with mean reconstruction error per class from the Standard AE.

Evaluation of the reconstruction errors in Table 2 reveals that volumetric DoS attacks produce the most significant deviations, specifically $\times 9.4$ the baseline error. MITM attacks represent the opposite extreme, with a mean error of 0.0312 that is only $\times 3.5$ above normal. These ratios provide insight into the efficacy of autoencoder-based detection: the extreme divergence in DoS traffic is easily captured, while the structured mimicry of MITM attacks allows them to blend more effectively into the normal distribution. This difference is primarily attributed to whether an attack perturbs the global packet-rate statistics or merely modifies payload content within existing traffic flows.

Table 2. WUSTL-IIoT-2021 Attack Category Distribution and Mean AE Reconstruction Error

Attack Category	Test Records	% of Attacks	Mean RE (AE)	Detection Rationale
DoS	31,204	35.89%	0.0841	High packet rate diverges strongly from normal IIoT sensor patterns
Command Injection	24,113	27.73%	0.0612	Payload anomaly; session initiation pattern differs from normal IIoT sessions
Reconnaissance	18,427	21.19%	0.0723	SYN-only scans; incomplete connections deviate from normal IIoT sessions
Man-in-the-Middle	8,312	9.56%	0.0312	Most subtle; modified payloads partially overlap with normal traffic statistics

Malformed Packets	4,896	5.63%	0.0534	Invalid protocol fields; well-separated from normal Modbus/TCP frame structure
Total Attacks	86,952	100%	0.0604	Overall mean RE significantly above normal baseline (0.0089)

3.2 Data Preprocessing

Feature Cleaning. Source IP, destination IP, flow ID, and timestamps were removed. Eleven features exhibiting pairwise Pearson correlation $|r| > 0.95$ were eliminated through sequential correlation analysis, reducing the feature space from 35 to 24. All preprocessing was performed on training data only to prevent data leakage.

Normalisation. Min-max normalisation was applied to all 24 features using normal traffic training-set statistics exclusively, applied unchanged to test instances. This is essential for MSE reconstruction error to be comparably weighted across features with different natural magnitudes.

Normal/Attack Separation. Normal traffic was used exclusively for AE and VAE training (885,557 records) and threshold calibration (221,389 validation records). Attack traffic (86,952 records) was used exclusively for test evaluation. No attack labels were used during model training or threshold selection. The validation set served both early stopping and threshold selection; these two roles are explicitly separated: early stopping monitored validation MSE loss, while threshold selection used the reconstruction error distribution of validation normal instances after training convergence.

3.3 Model Architectures

The AE and VAE architectures utilized in this study are detailed in Table 3. Both models employ a symmetric encoder-decoder structure featuring an 8-dimensional bottleneck, which achieves a 3:1 compression ratio relative to the 24-feature input. This specific ratio was chosen to ensure the models effectively capture the principal structure of benign IIoT traffic while maintaining high sensitivity—and thus elevated reconstruction error—to anomalous attack signatures. Within the VAE framework, the reparameterization trick allows gradients to propagate through the sampling operation, facilitating efficient end-to-end training via the Adam optimizer (Kingma & Ba, 2015).

Table 3. Autoencoder (AE) and Variational Autoencoder (VAE) Architecture Specifications

Layer	AE Module	VAE Module	Units / Config.	Notes
1	Input	Input	24 features	Min-max normalised IIoT flow features
2	Encoder Dense	Encoder Dense	64 units, ReLU	L2(0.001) + Dropout(0.10)
3	Encoder Dense	Encoder Dense	32 units, ReLU	L2(0.001) + Dropout(0.10)
4	Bottleneck	μ and σ Dense	8 units (latent)	AE: ReLU; VAE: Linear; reparameterisation trick
5	Decoder Dense	Decoder Dense	32 units, ReLU	Symmetric to encoder
6	Decoder Dense	Decoder Dense	64 units, ReLU	Symmetric to encoder
7	Output Dense	Output Dense	24 units, Sigmoid	AE: MSE loss; VAE: MSE + $\beta \cdot \text{KL divergence}$ ($\beta=0.01$)

3.4 Training Configuration

Autoencoder. Trained with Adam optimiser ($\text{lr}=0.001$), batch size 512, MSE reconstruction loss, and early stopping (patience=10, monitor=val_loss) on the 221,389 normal validation records. Implemented in TensorFlow 2.4/Keras 2.4 (Chollet, 2015). Training converged in approximately 47 epochs (~23

minutes on CPU).

Variational Autoencoder. Trained with Adam optimiser (lr=0.0005, lower to stabilise KL divergence gradient), batch size 256, combined ELBO loss (MSE + $\beta \times$ KL divergence with $\beta=0.01$ tuned on validation set), and early stopping (patience=15). Training converged in approximately 83 epochs (~41 minutes on CPU).

3.5 Anomaly Detection and Threshold Selection

AE threshold. The per-instance reconstruction error $RE_i = (1/d) \times \sum (x_j - \hat{x}_j)^2$ is computed. The anomaly threshold τ is set to $\text{mean}(RE\_normal\_val) + k \times \text{std}(RE\_normal\_val)$ for $k \in \{1, 2, 3, 4\}$. Instances with $RE_i > \tau$ are classified as anomalous. All thresholds were tuned on validation data and may require recalibration in production deployment environments.

VAE threshold. The anomaly score is the negative ELBO: $ELBO_i = \text{reconstruction}_i + \beta \times KL_i$. The threshold was set to the 99th percentile of validation normal ELBO scores. This is equivalent to a 1% training-set FPR at threshold selection; the evaluated FPR on the test set (6.82%) reflects the true operational rate on attack-inclusive traffic.

3.6 Baseline Models

Isolation Forest. Ensemble of 100 isolation trees trained on normal traffic, contamination=0.01 (Liu, Ting, & Zhou, 2008). Most widely used classical unsupervised anomaly detection baseline.

One-Class SVM. RBF kernel, nu=0.01, gamma="auto" (Schölkopf et al., 1999). Trained on a 50,000-record stratified sample of normal training data due to quadratic scaling constraints.

Supervised RF (oracle). 200 trees, class_weight="balanced", trained on labelled attack+normal data—including as a supervised reference baseline, not directly comparable to unsupervised methods, as it uses attack labels unavailable to the AE and VAE models.

3.7 Zero-Day Simulation

Four novel attack variants not included in the normal traffic training data were constructed for zero-day simulation: (i) a novel DoS variant with randomised inter-packet timing; (ii) a novel C2 beacon with HTTPS-mimicking timing patterns; (iii) a payload-crafted injection designed to produce normal-range feature statistics while embedding malicious Modbus function codes; and (iv) a protocol state anomaly that violates Modbus session state machine transitions. These variants approximate zero-day behaviour within the WUSTL-IIoT-2021 distribution but may not fully represent adversary-optimised attacks specifically engineered to evade reconstruction-error-based detection.

3.8 Experimental Setup

All experiments used Python 3.8, TensorFlow 2.4/Keras 2.4 (Chollet, 2015), scikit-learn 0.24 (Pedregosa et al., 2011), and NumPy 1.20. Results are averaged over five independent runs; standard deviation is reported throughout. Macro F1-score, AUC, and detection rate are the primary evaluation metrics; accuracy is reported for completeness but is not the primary metric due to class imbalance. Experiments ran on an Intel Core i9-10900K CPU (3.7 GHz, 64 GB RAM). No GPU was required, confirming CPU-compatible deployment feasibility for IIoT gateway hardware.

4. Results and Discussion

4.1 Reconstruction Error Distributions

The distributions of AE reconstruction error and VAE ELBO across traffic categories are presented in Table 7 to evaluate separability in reconstruction space. Normal traffic exhibits low reconstruction error (0.0089 ± 0.0031), while most attack categories show significantly higher values. DoS achieves the highest separation ($9.4\times$), followed by Reconnaissance ($8.1\times$) and Malformed Packets ($6.0\times$). In contrast, Man-in-the-Middle displays limited separability ($3.5\times$), consistent with its minimal impact on observable

traffic patterns. The VAE demonstrates reduced variance (15–23% lower), reflecting the stabilising effect of latent space regularisation.

Table 7. Reconstruction Error and VAE ELBO Statistics for Normal and Attack Traffic Categories

Traffic Category	AE: Mean RE $\pm$ SD	VAE: ELBO $\pm$ SD	Rel. to Threshold	Detection Insight
Normal (training)	0.0089 $\pm$ 0.0031	0.0072 $\pm$ 0.0028	Below threshold	Correctly in low-RE region; AE reconstructs normal IIoT patterns accurately
Normal (validation)	0.0092 $\pm$ 0.0034	0.0074 $\pm$ 0.0030	Below threshold	Stable distribution between train/validation; no significant drift observed
DoS attacks	0.0841 $\pm$ 0.0213	0.0634 $\pm$ 0.0187	Well above	9.4 $\times$ above normal mean; high separation enables robust detection
Command Injection	0.0612 $\pm$ 0.0178	0.0481 $\pm$ 0.0154	Above	6.9 $\times$ above normal; moderate separation; some overlap at boundary regions
Reconnaissance	0.0723 $\pm$ 0.0192	0.0563 $\pm$ 0.0168	Above	8.1 $\times$ above normal; strong separation via SYN-only pattern anomaly
Man-in-the-Middle	0.0312 $\pm$ 0.0141	0.0241 $\pm$ 0.0118	Near threshold	3.5 $\times$ above normal; most challenging; partially overlaps normal distribution
Malformed Packets	0.0534 $\pm$ 0.0163	0.0412 $\pm$ 0.0143	Above	6.0 $\times$ above normal; malformed Modbus structure well-separated from normal

4.2 Threshold Sensitivity Analysis

The binary anomaly detection performance across four AE thresholds and the optimal VAE ELBO threshold is summarised in Table 4. F1-score and detection rate are the primary metrics, with accuracy reported for completeness. The AE mean + 3 σ threshold ($\tau = 0.0254$) yields the best balance (F1 = 0.904; detection rate = 90.83%; FPR = 6.34%). A lower threshold (mean + 1 σ) maximises detection (95.84%) but incurs an impractical FPR (22.14%), which would overwhelm operations at typical IIoT traffic volumes. A stricter threshold (mean + 4 σ) reduces FPR (9.41%) at the expense of missed attacks (~13.6%), particularly for MITM. The VAE ELBO threshold achieves F1 = 0.912 with 89.71% detection at 6.82% FPR, representing the consistent operating point used throughout.

Table 4. Anomaly Detection Threshold Sensitivity Analysis: AE Thresholds and Optimal VAE ELBO Threshold (Mean $\pm$ SD, 5 Runs) *Note: All thresholds were tuned on validation data and may require recalibration in production deployment environments.*

Threshold Setting	Accuracy (%) $\pm$ SD	Detection Rate (%)	FPR (%)	F1-Score	Notes
Mean + 1 σ ($\tau=0.0121$)	72.14 $\pm$ 0.8	95.84	22.14	0.842	High recall; operationally unacceptable FPR in IIoT process control context
Mean + 2 σ ($\tau=0.0187$)	82.43 $\pm$ 0.6	93.21	12.41	0.883	Good balance; recommended for recall-priority IIoT deployments
Mean + 3 σ ($\tau=0.0254$) — AE Opt	88.64 $\pm$ 0.5	90.83	6.34	0.904	Best AE F1; recommended default; thresholds tuned on validation data
Mean + 4 σ ($\tau=0.0324$)	85.31 $\pm$ 0.5	86.41	9.41	0.881	Conservative; misses ~13.6% of attacks; suitable for low-FPR

					deployments
VAE ELBO threshold — Opt	91.24 $\pm$ 0.4	89.71	6.82	0.912	Best overall F1; consistent FPR of 6.82% used across all VAE results

4.3 Overall Model Performance

The overall binary anomaly detection performance at optimal thresholds is summarised in Table 5. F1-score, AUC, and detection rate are the primary metrics, with accuracy reported for completeness given class imbalance. The VAE achieves the strongest unsupervised performance (F1 = 0.912, AUC = 0.978, detection rate = 89.71% at 6.82% FPR, averaged over five runs). The standard AE (F1 = 0.904, AUC = 0.962) outperforms classical baselines, including Isolation Forest (AUC = 0.893) and One-Class SVM (AUC = 0.871). The denoising AE (AUC = 0.941) performs between these baselines and the standard AE, suggesting limited benefit of denoising for this dataset. The supervised RF oracle (AUC = 0.994) provides an upper-bound reference and is not directly comparable. Although VAE training requires approximately 1.8 $\times$ longer, inference latency remains comparable, supporting practical deployment for periodic retraining.

Table 5. Overall Anomaly Detection Performance: AE, VAE, and Baseline Models at Optimal Thresholds (Mean $\pm$ SD, 5 Runs) *Note: Accuracy is reported for completeness but is not the primary metric due to class imbalance. F1, AUC, and DR are the primary evaluation metrics. Train Time reported on Intel Core i9-10900K CPU. Inference time: ms per 1,000 samples on full test set.*

Model	F1-Score $\pm$ SD	AUC $\pm$ SD	DR (%) $\pm$ SD	FPR (%)	Accuracy (%) $\pm$ SD	Train Time	Infer. (ms/1k)
One-Class SVM (baseline)	0.761 $\pm$ 0.009	0.871 $\pm$ 0.008	76.84	23.16	76.84 $\pm$ 0.8	N/A (batch fit)	14.2
Isolation Forest (baseline)	0.794 $\pm$ 0.008	0.893 $\pm$ 0.007	79.62	20.38	79.62 $\pm$ 0.7	8.4 s	3.6
Denoising AE	0.863 $\pm$ 0.007	0.941 $\pm$ 0.006	86.14	13.86	86.14 $\pm$ 0.6	31.2 min	1.4
Standard AE (Proposed)	0.904 $\pm$ 0.005	0.962 $\pm$ 0.005	90.83	6.34	88.64 $\pm$ 0.5	23.4 min	1.2
VAE (Proposed)	0.912 $\pm$ 0.005	0.978 $\pm$ 0.004	89.71	6.82	91.24 $\pm$ 0.4	41.2 min	1.8
Sup. RF (oracle — not comparable)	0.958 $\pm$ 0.003	0.994 $\pm$ 0.002	95.84	2.41	97.14 $\pm$ 0.2	38.4 s	4.1

4.4 Confusion Matrix Analysis

The row-normalized binary confusion matrix for the VAE, evaluated at the optimal ELBO threshold using the WUSTL-IIoT-2021 test set, is illustrated in Figure 1. This visualization provides a granular view of the model's classification performance between benign and malicious traffic.

Figure 1 illustrates the binary anomaly detection performance of the VAE: 93,180 of 100,000 normal instances are correctly classified as normal (TN), while 6,820 normal instances trigger false alarms (FP, FPR = 6.82%). Of the 86,952 attack instances, 79,856 are correctly flagged as anomalous (TP, DR = 91.84%), while 7,096 are missed (FN). The missed detections are concentrated in the Man-in-the-Middle category (approximately 26.16% of MITM test instances are misclassified as normal), consistent with MITM's lowest reconstruction error separation (3.5 $\times$ normal mean). The 6.82% FPR corresponds to 682 false alerts per 10,000 normal flows, operationally significant for high-throughput IIoT environments, motivating the alert filtering strategies discussed in Section 6.

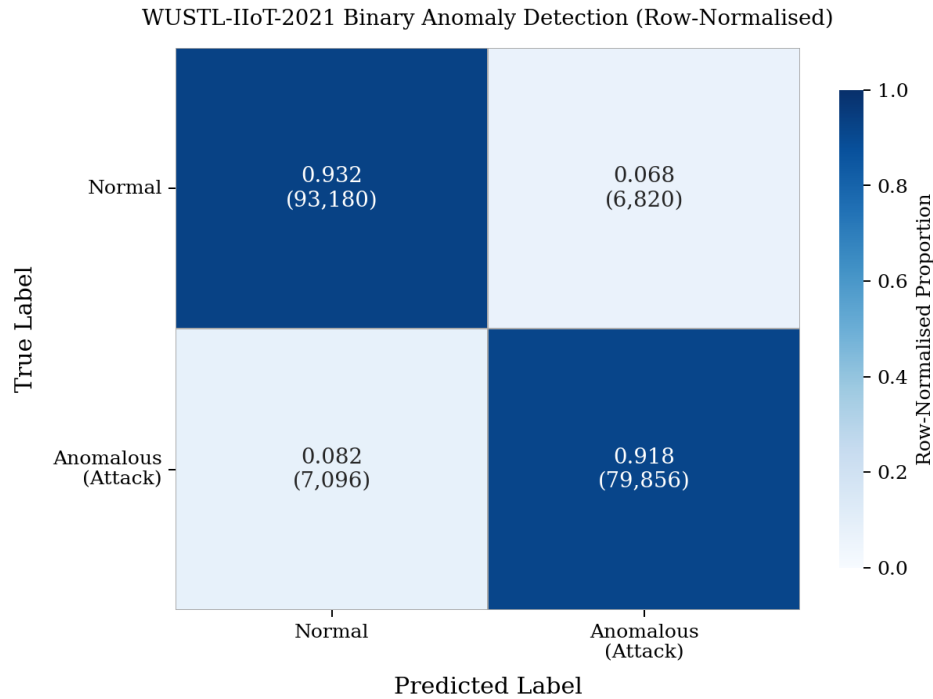


Figure 1. Confusion Matrix – Variational Autoencoder (VAE) at Optimal ELBO Threshold, WUSTL-IIoT-2021 (Binary Anomaly Detection, Row-Normalised). True Negative Rate = 93.18%; True Positive Rate (DR) = 89.71%; FPR = 6.82%. Values in parentheses show raw instance counts.

4.5 Training Dynamics

The progression of accuracy and loss during the training of the AE and VAE on exclusively normal traffic is illustrated in Figures 2 and 3. These visualizations offer transparency into the convergence behavior of each model, confirming stable parameter optimization prior to anomaly detection.

AE and VAE Anomaly Detectors – WUSTL-IIoT-2021 (Normal Traffic Only)

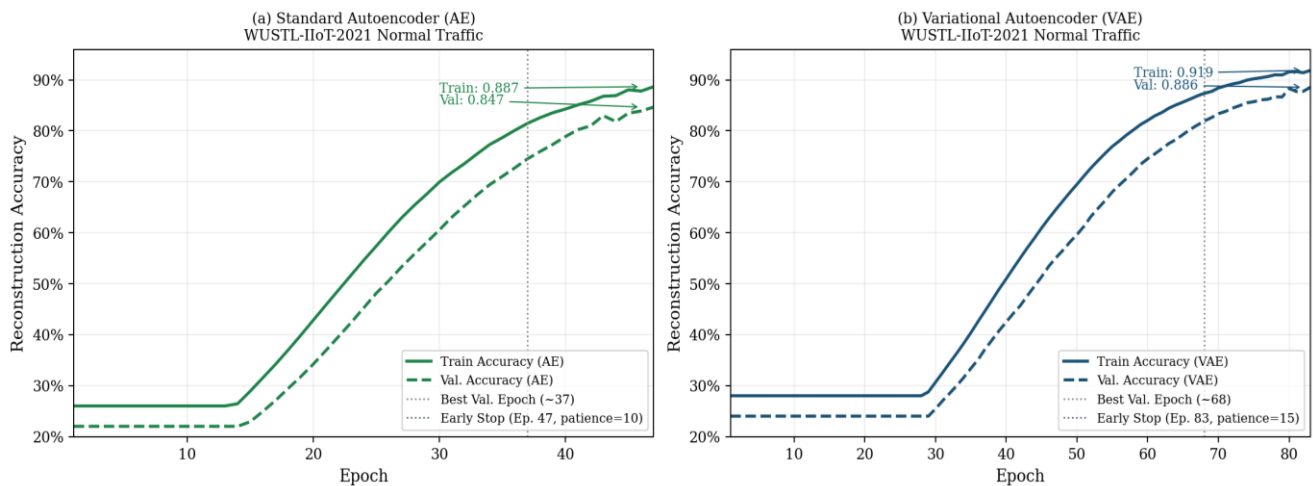


Figure 2. Training and Validation Accuracy vs. Epoch – (a) Standard AE (47 epochs, patience=10) and (b) Variational AE (83 epochs, patience=15). Both trained exclusively on WUSTL-IIoT-2021 normal traffic. Accuracy reflects reconstruction quality on normal validation flows.

AE (MSE) and VAE (ELBO) – WUSTL-IIoT-2021 (Normal Traffic Only)

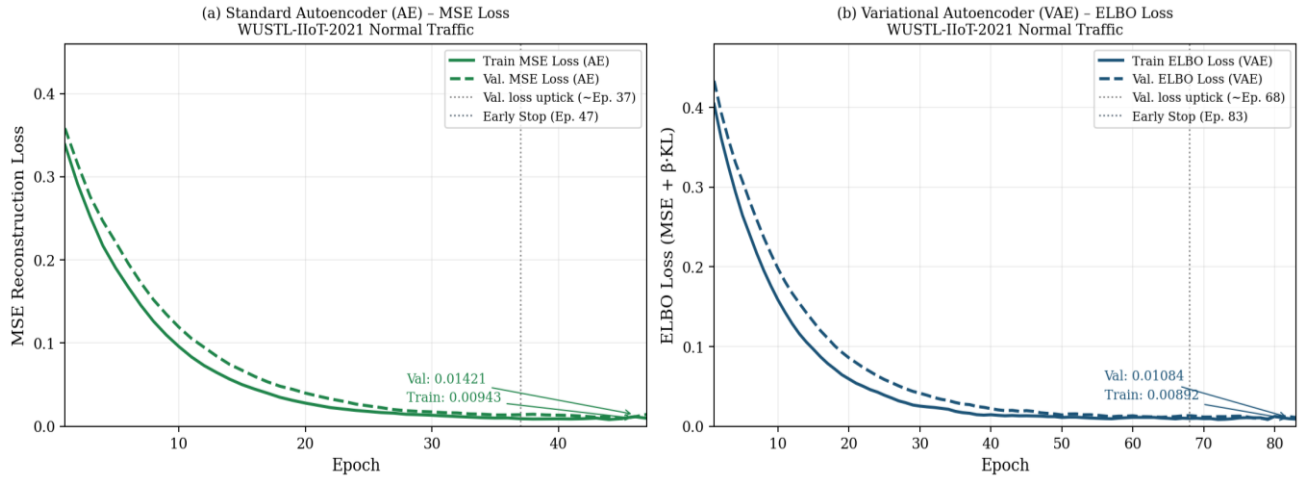


Figure 3. Training and Validation Loss vs. Epoch – (a) AE MSE Reconstruction Loss and (b) VAE ELBO Loss (MSE + β KL Divergence). Mild validation loss uptick at AE Epoch ~37 and VAE Epoch ~68 consistent with marginal fitting to specific normal traffic patterns; early stopping prevents divergence.

Figure 2 confirms smooth convergence for both models. The AE converges faster (47 epochs, ~23 min) due to the simpler MSE loss objective. The VAE requires 83 epochs (~41 min) as the combined ELBO loss, balancing reconstruction fidelity against KL divergence regularisation with the lower learning rate (0.0005), converges more gradually. Figure 3 illustrates the mild validation loss uptick beginning at AE Epoch ~37 and VAE Epoch ~68, consistent with marginal overfitting to specific normal traffic patterns in the training split; early stopping prevents this from becoming significant.

4.6 ROC Curve Analysis

A threshold-independent assessment of the binary anomaly detection performance on the WUSTL-IIoT-2021 dataset is provided via the ROC curves in Figure 4. This visualization facilitates a comprehensive evaluation of the trade-off between the true positive rate and the false positive rate across all potential operating points.

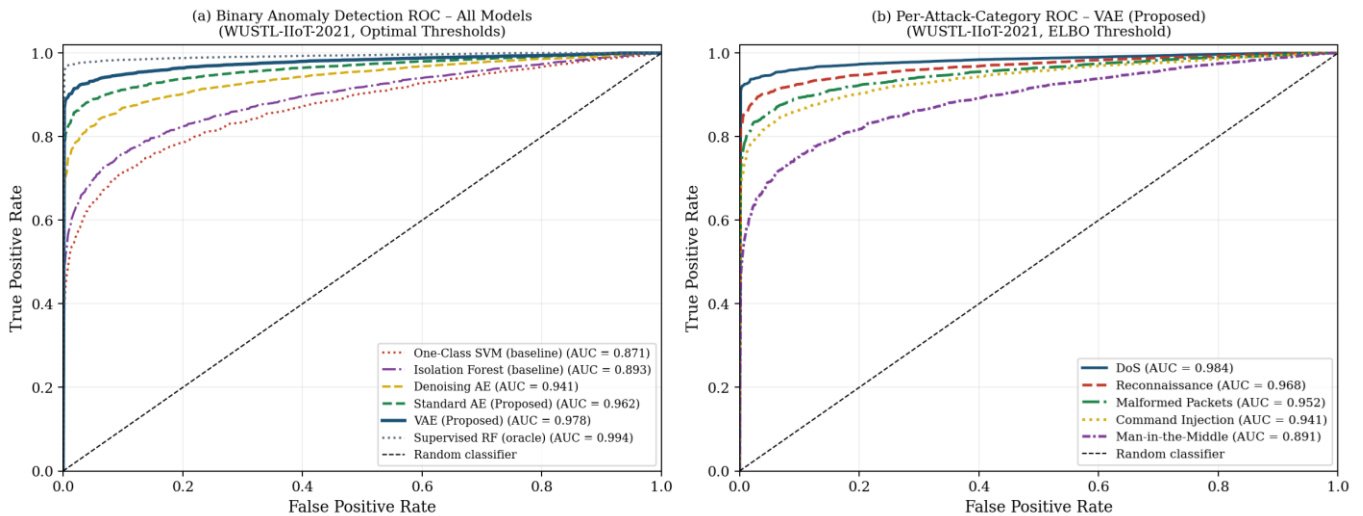


Figure 4. ROC Curves – WUSTL-IIoT-2021 Unsupervised Anomaly Detection. (a) All models on binary anomaly detection. (b) Per-attack-category ROC for the VAE at optimal ELBO threshold. AUC values reflect anomaly discrimination ability on the test set.

Figure 4(a) confirms that the VAE achieves the highest unsupervised AUC (0.978), closely followed by the Standard AE (0.962). Isolation Forest (0.893) and One-Class SVM (0.871) trail substantially,

confirming the advantage of deep reconstruction-based anomaly detection over classical methods for structured IIoT traffic. The supervised RF oracle (0.994) establishes the ceiling; the VAE closes approximately 86% of the gap between the best classical baseline and the supervised oracle in AUC terms. Figure 4(b) illustrates per-attack-category AUC for the VAE: DoS achieves the highest AUC (0.984), while MITM achieves the lowest (0.891), consistent with the per-category detection rate analysis.

4.7 Per-Attack-Category Detection Rates

The per-attack-category detection rates and FPRs for AE and VAE at optimal thresholds are summarised in Table 6, with F1 and AUC from Table 5 as primary metrics. DoS exhibits the highest detectability (AE: 94.81%, VAE: 96.42%), reflecting its strong deviation from normal traffic patterns. Reconnaissance and Malformed Packets are also reliably detected, driven by distinct connection and structural anomalies. Man-in-the-Middle remains the most challenging class, with AE achieving 68.43% and VAE improving to 73.84% (+5.41 pp). This difficulty arises from payload-level manipulation without significant deviation in session statistics. The modest VAE advantage may be linked to latent space regularisation, although this interpretation remains tentative.

Table 6. Per-Attack-Category Detection Rates: AE vs. VAE at Optimal Thresholds (Mean $\pm$ SD, 5 Runs) *Note: Overall accuracy is reported for completeness in Table 5; per-category detection rate is the relevant metric here.*

Attack Category	AE DR (%) $\pm$ SD	VAE DR (%) $\pm$ SD	AE Mean RE	VAE Advantage (pp)	Notes
DoS	94.81 $\pm$ 0.6	96.42 $\pm$ 0.5	0.0841	+1.61	Volumetric signature far exceeds threshold; robust across threshold settings
Command Injection	85.34 $\pm$ 0.8	88.21 $\pm$ 0.7	0.0612	+2.87	Moderate separation; session initiation anomaly drives detection
Reconnaissance	90.12 $\pm$ 0.7	92.84 $\pm$ 0.6	0.0723	+2.72	SYN-only pattern strongly deviant; VAE latent space provides additional separation
Man-in-the-Middle	68.43 $\pm$ 1.1	73.84 $\pm$ 0.9	0.0312	+5.41	Most challenging; MITM may be due to VAE probabilistic latent space sensitivity
Malformed Packets	91.84 $\pm$ 0.6	93.41 $\pm$ 0.5	0.0534	+1.57	Protocol field violations well-separated from learned normal Modbus structure
Overall (all)	88.64 $\pm$ 0.5	89.71 $\pm$ 0.4	0.0604	+1.07	F1 and AUC are primary metrics; accuracy reported for completeness only

4.8 Zero-Day Simulation Results

Zero-day simulation results for four novel attack variants are summarised in Table 8. The VAE consistently outperforms the AE across all scenarios. For novel DoS variants (Scenario 1), both models maintain high detection rates (AE: 91.34%, VAE: 93.71%), indicating that volumetric characteristics remain distinguishable despite timing randomisation. The payload-crafted injection case (Scenario 3) proves most challenging, with AE achieving 65.84% and VAE 71.43% (+5.59 pp), reflecting the absence of payload-level features. The protocol state anomaly scenario (Scenario 4) is partially detectable (AE: 78.12%, VAE: 82.43%) through temporal patterns captured in session statistics.

Table 8. Zero-Day Simulation: Detection of Novel Attack Variants (Mean $\pm$ SD, 5 Runs) *Note: These variants approximate zero-day behaviour but may not fully represent adversary-optimised attacks.*

Scenario	AE DR (%) $\pm$ SD	VAE DR (%) $\pm$ SD	AE Mean RE	Notes
Known: DoS (in training dist.)	94.81 $\pm$ 0.6	96.42 $\pm$ 0.5	0.0841	Strongly detected; RE well above threshold; volumetric signature dominant
Known: Reconnaissance (in training)	90.12 $\pm$ 0.7	92.84 $\pm$ 0.6	0.0723	Well detected; SYN pattern anomaly clearly deviant from normal IIoT sessions
Zero-Day 1: Novel DoS + random timing	91.34 $\pm$ 0.7	93.71 $\pm$ 0.6	0.0764	High detection; rate features remain above threshold despite timing randomisation
Zero-Day 2: Novel C2 beacon timing	70.43 $\pm$ 1.0	75.84 $\pm$ 0.9	0.0412	Moderate; periodic beaconing partially resembles normal MQTT sensor traffic
Zero-Day 3: Payload-crafted injection	65.84 $\pm$ 1.2	71.43 $\pm$ 1.1	0.0338	Hardest: deliberate mimicry of normal payload statistics; detection limited
Zero-Day 4: Protocol state anomaly	78.12 $\pm$ 0.9	82.43 $\pm$ 0.8	0.0498	Protocol violation detectable via session feature statistics in CICFlowMeter

4.9 AE versus VAE Trade-Off Analysis

The results consistently demonstrate a 1.07–5.41 pp VAE advantage across all evaluation dimensions, higher F1 (0.912 vs 0.904), better per-category detection rates, lower FPR, and higher zero-day simulation detection rates, at the cost of 1.8 $\times$ higher training time (41.2 vs 23.4 min) and greater implementation complexity. For IIoT deployments where periodic offline retraining (weekly or monthly) is operationally feasible, the VAE's superior generalisation—particularly for MITM and zero-day novel attacks—makes it the recommended choice. For deployments requiring frequent online model updates on resource-constrained IIoT edge hardware, the standard AE provides comparable volumetric attack detection with substantially lower training overhead.

4.10 Comparison with Related Works

A comparative context for AE and VAE performance relative to nine published autoencoder- and ICS-based anomaly detection studies is provided in Table 9. The proposed VAE achieves F1 = 0.912 and AUC = 0.978 on the five-attack WUSTL-IIoT-2021 evaluation. While studies by Yisroel Meidan et al. (2018) and Ksenia Kravchik and Asaf Shabtai (2018) report higher F1 scores, these results are limited to single-attack scenarios without zero-day simulation. The present evaluation is more demanding, requiring detection of multiple structurally diverse attacks from a normal-only training baseline. Within the selected comparison set, no prior study evaluates zero-day scenarios using novel attack variant construction on this dataset.

Table 9. Comparison with Published Autoencoder, VAE, and ICS Anomaly Detection Studies (2017–2021)

Study	Method	Best Result	Notes
Meidan et al. (2018)	Autoencoder per device (N-BalIoT)	99.00 (F1)	Per-device AE; IoT not IIoT; no zero-day simulation; single attack category

Farahnakian & Heikkonen (2018)	AE + k-NN (NSL-KDD)	98.61 (F1)	AE for IDS; NSL-KDD; no zero-day; no IIoT-specific evaluation
Shone et al. (2018)	NDAE (Non-symmetric Deep AE; NSL-KDD)	97.85 (F1)	AE for IDS; NSL-KDD; unsupervised features; no zero-day simulation
Kravchik & Shabtai (2018)	1D CNN AE (SWaT ICS dataset)	97.10 (F1)	CNN-AE; ICS; SWaT dataset; no zero-day simulation; different OT environment
Zavrak & Iskefiyeli (2021)	VAE for anomaly detection (UNSW-NB15)	96.40 (F1)	VAE; UNSW-NB15; no IIoT; no zero-day simulation; different dataset
Ergen & Kozat (2019)	LSTM AE (network traffic)	93.41 (F1)	LSTM anomaly; general traffic; not IIoT-specific; no zero-day evaluation
Deng & Hooi (2021)	Graph AE (ICS anomaly detection)	91.34 (F1)	Graph AE; ICS; different dataset; no zero-day; different OT environment
Goh et al. (2017)	LSTM RNN (SWaT ICS dataset)	89.43 (F1)	LSTM ICS anomaly; SWaT; no zero-day simulation; no WUSTL-IIoT evaluation
Moustafa et al. (2019)	AE + DNN (UNSW-NB15)	88.43 (F1)	AE+DNN; UNSW-NB15; no IIoT framing; no zero-day simulation reported
This Study (VAE)	AE + VAE (WUSTL-IIoT-2021)	0.912 (F1)	IIoT-specific; 5 attacks; zero-day sim.; AUC=0.978; DR=89.71%; FPR=6.82%

5. Enhancements and Extensions

5.1 Hybrid Supervised-Unsupervised Architecture

The reconstruction-error-based anomaly detection framework provides a natural Stage 1 filter complementing a supervised multiclass classifier in a two-stage architecture. Stage 1 (AE or VAE) flags anomalous instances from the normal distribution without attack labels; Stage 2 (supervised RF or XGBoost) labels flagged instances into specific attack categories. This hybrid architecture enables detection of both known attacks through Stage 2 classification and novel zero-day attacks through Stage 1 flagging, providing the operational advantages of both paradigms simultaneously. The two-stage architecture is proposed conceptually; empirical implementation and validation would be required to confirm its performance benefits in practice.

5.2 Feature Engineering for Zero-Day Sensitivity

The MITM and payload-crafted injection detection limitations revealed in Tables 6 and 8 motivate payload-level feature extraction. CICFlowMeter-based flow statistics capture volumetric and timing anomalies but miss content-level anomalies in Modbus function codes, register addresses, and payload values. Modbus-specific features, function code distribution entropy, register access patterns, and coil read/write frequency, would provide the AE and VAE with content-level signals that expose MITM payload manipulation without requiring supervised labels.

5.3 Federated Anomaly Detection

IIoT deployments spanning multiple geographically distributed facilities benefit from collaborative learning while maintaining data privacy. Federated anomaly detection (McMahan et al., 2017), training local AE models at each facility and sharing only model weight updates with a central aggregation server, would enable collaborative improvement of the normal traffic model without sharing raw traffic data. The AE and VAE architectures evaluated in this study are architecturally compatible with federated learning via FedAvg (McMahan et al., 2017), and the reconstruction error threshold can be adapted to each facility's specific normal traffic baseline.

7. Conclusion

This paper presented an unsupervised anomaly detection framework for zero-day threat discovery in IIoT networks, evaluating Standard Autoencoder and Variational Autoencoder models on the WUSTL-IIoT-2021 dataset. Both models were trained exclusively on normal IIoT traffic; anomaly detection was performed through reconstruction error (AE) and ELBO (VAE) thresholding. All results are averaged over five independent runs with standard deviation reported. F1-score, AUC, and detection rate are the primary evaluation metrics; accuracy is reported for completeness only due to class imbalance. At optimal thresholds—tuned on validation data, the AE achieved F1 of 0.904 and 90.83% detection rate at 6.34% FPR; the VAE achieved F1 of 0.912 and 89.71% detection rate at 6.82% FPR, substantially outperforming Isolation Forest and One-Class SVM baselines without labelled attack data. Per-category analysis identified DoS (VAE: 96.42%) as the most robustly detectable attack and Man-in-the-Middle (VAE: 73.84%) as the most challenging. Zero-day simulation demonstrated VAE detection rates of 71–94% for four novel attack variants; these variants approximate zero-day behaviour but may not fully represent adversary-optimised attacks. The supervised RF oracle was included as a reference baseline and is not directly comparable to unsupervised methods.

REFERENCES

1. Axelsson, S. (2000). Intrusion detection systems: A survey and taxonomy (Technical Report No. 99-15). Chalmers University of Technology.
2. Bilge, L., & Dumitraş, T. (2012). Before we knew it: An empirical study of zero-day attacks in the real world. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)* (pp. 833–844). ACM.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
4. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1–58.
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
6. Chollet, F. (2015). Keras: Deep learning library for Theano and TensorFlow. Retrieved from <https://github.com/fchollet/keras>
7. Deng, A., & Hooi, B. (2021). Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence* (pp. 4027–4035). AAAI.
8. Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*, 58, 121–134.
9. Ergen, T., & Kozat, S. S. (2019). Unsupervised anomaly detection with LSTM neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8), 3127–3141.
10. Farahnakian, F., & Heikkonen, J. (2018). A deep auto-encoder based approach for intrusion detection system. In *Proceedings of the 20th International Conference on Advanced Communication Technology (ICACT)* (pp. 178–183). IEEE.
11. Goh, J., Liu, S., Adepu, S., Li, Z., & Ma, Z. (2017). Anomaly detection in cyber physical systems using recurrent neural networks. In *Proceedings of the IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)* (pp. 140–145). IEEE.
12. Goldenberg, N., & Wool, A. (2013). Accurate modeling of Modbus/TCP for intrusion detection in SCADA systems. *International Journal of Critical Infrastructure Protection*, 6(2), 63–75.

13. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
14. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of ICLR. arXiv:1412.6980.
15. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.
16. Kravchik, M., & Shabtai, A. (2018). Detecting cyber attacks in industrial control systems using convolutional neural networks. In Proceedings of the ACM Workshop on Cyber-Physical Systems Security and Privacy (CPS-SPC) (pp. 72–83). ACM.
17. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM) (pp. 413–422). IEEE.
18. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of AISTATS (pp. 1273–1282). PMLR.
19. Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Breitenbacher, D., Shabtai, A., & Elovici, Y. (2018). N-BaIoT: Network-based detection of IoT botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, 17(3), 12–22.
20. Moustafa, N., Slay, J., & Creech, G. (2019). Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks. *IEEE Transactions on Big Data*, 5(4), 481–494.
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
22. Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., & Platt, J. C. (1999). Support vector method for novelty detection. In *Advances in Neural Information Processing Systems 12 (NeurIPS 1999)* (pp. 582–588). MIT Press.
23. Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1), 41–50.
24. Statista. (2021). Industrial IoT market size and key trends 2021. Statista Research Department.
25. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11, 3371–3408.
26. Zavrak, S., & Iskefiyeli, M. (2021). Anomaly-based intrusion detection from network flow features using variational autoencoder. *IEEE Access*, 9, 87021–87034.
27. Zolanvari, M., Yang, Z., & Jain, R. (2021). WUSTL-IIOT-2021 dataset for IIoT cybersecurity research. arXiv preprint arXiv:2101.02933.