Science and Technology Publishing (SCI & TECH)
ISSN: 2632-1017
Vol. 10 Issue 2, February - 2026

Hybrid Autoencoder–Random Forest Intrusion Detection for Smart Grid Communication Channels Using ICS Traffic Data

Amasa Ukwuoma Emmanuel[1]
Department OF Electrical and Electronic Engineering
Federal University Otuoke, Bayelsa State, Nigeria
amasaeu@fuotuoke.edu.ng

Ubon Etefia Imoh-Etefia[2]
Department of Computer Engineering
University of Uyo, Akwa Ibom State

Daniel Chigaeduzom Nnadi[3]
Department of Mechanical Engineering.
Michael Okpara University of Agriculture Umudike, Abia State
nnadi.daniel@mouau.edu.ng

Abstract

Modern smart grid communication relies on IEC 60870-5-104 (IEC 104) and IEC 61850 Manufacturing Message Specification (MMS) protocols to coordinate operational technology assets across substations and control centres. The convergence of these networks with conventional IT infrastructure has substantially enlarged the attack surface, exposing SCADA systems to threats including port scanning, command injection, command blocking, and link failure simulation. Signature-based detection cannot address zero-day protocol-layer exploits, while computationally intensive deep learning alternatives introduce latency incompatible with IEC 104 polling constraints. This paper presents a hybrid intrusion detection system (IDS) that couples an unsupervised autoencoder with a supervised Random Forest classifier, evaluated on the publicly available ICS Dataset for Smart Grid Anomaly Detection (Matoušek et al., 2022). The novelty lies in the lightweight, protocol-aware architecture, incorporating protocol-specific IPFIX feature engineering, latency-aware deployment analysis, and statistically validated ablation studies on an underexplored benchmark. Using a stratified 70/30 train-test split with five-fold cross-validation, the hybrid model achieves 97.9% detection accuracy, 96.8% precision, 98.4% recall, F1-score of 97.6%, AUC-ROC of 0.9971, and a false positive rate of 2.5%. Mean inference time of 0.31 ms per flow confirms near-real-time feasibility for substation monitoring.

Keywords: smart grid security; anomaly detection; intrusion detection; machine learning; ICS; IEC 60870-5-104; IEC 61850; autoencoder; random forest; SCADA

1. Introduction

Smart grids integrate real-time digital communication into every layer of power system operation, from generation dispatch to end-customer metering. At the substation level, the IEC 60870-5-104 protocol governs telecontrol messaging between control centres and remote terminal units (RTUs) over TCP/IP links, while IEC 61850 standardises inter-device communication using MMS over Ethernet. Both protocols were designed for high-reliability operation in benign environments and lack native authentication, encryption, or message integrity verification (Radoglou-Grammatikis & Sarigiannidis, 2019). These architectural properties make them attractive targets for adversaries seeking to disrupt grid operation or exfiltrate operational data.

The consequences of successful cyberattacks on grid communication channels can be severe. The 2015 and 2016 Ukrainian power utility attacks exploited remote communication interfaces and resulted in outages affecting hundreds of thousands of customers (Lee et al., 2016; Liang et al., 2017). Energy sector facilities have consistently accounted for a significant share of incidents reported to the U.S. Cybersecurity and Infrastructure Security Agency (CISA, 2022). Adversarial techniques are evolving from opportunistic exploitation toward deliberate, protocol-aware intrusions that leave minimal forensic traces in generic network logs.

Conventional signature-based intrusion detection systems are effective against catalogued threat patterns but fail against zero-day exploits and novel protocol-layer manipulations. Anomaly-based detection offers a complementary paradigm: a statistical model of normal traffic is learned and significant deviations are flagged as potential intrusions. Machine learning has become the dominant methodology for building such models in industrial control system (ICS) environments, offering the capacity to extract discriminative patterns from high-dimensional IPFIX flow features without manual rule authoring (Mitchell & Chen, 2014). Unsupervised methods such as autoencoders learn the normal-traffic manifold from unlabelled data, but their decision boundaries can be imprecise in environments with high legitimate variability. Supervised classifiers offer tighter decision boundaries but require comprehensively labelled attack data that is scarce for rare or novel attack categories. Hybrid architectures that combine unsupervised feature learning with supervised classification address both limitations simultaneously.

This paper proposes and evaluates a hybrid IDS that couples an autoencoder-derived reconstruction error signal with a Random Forest binary classifier. The novelty of this work lies in the development and comprehensive evaluation of a lightweight hybrid IDS specifically optimised for IEC 104 and MMS smart grid communication environments, incorporating protocol-aware IPFIX feature engineering, latency-aware deployment analysis, and statistically validated ablation studies on the Matoušek et al. (2022) ICS Dataset for Smart Grid Anomaly Detection. No prior study has jointly reported ablation analysis, per-category detection rates, feature importance, and inference latency for this benchmark.

This study makes several important contributions to smart grid intrusion detection research. First, it introduces a preprocessing framework specifically designed for IPFIX-derived CSV traffic traces from IEC 104 and MMS communication environments, incorporating categorical feature handling, temporal feature extraction, and near-zero-variance feature elimination to improve data consistency and model effectiveness. Second, the work proposes a hybrid IDS architecture in which autoencoder reconstruction error is fused with normalised IPFIX traffic attributes to form an enhanced 17-dimensional feature vector for Random Forest-based classification. Experimental evaluation on the Matoušek et al. (2022) smart grid ICS dataset demonstrates strong detection capability, achieving 97.9% accuracy, an AUC-ROC of 0.9971, and a false positive rate of 2.5%, with all metrics derived consistently from a unified confusion matrix. Furthermore, the study includes detailed ablation experiments with cross-validation mean and standard deviation reporting for all configurations and baselines, alongside statistical significance assessment using the Wilcoxon signed-rank test. Finally, feature importance analysis, attack-category-specific detection evaluation, and inference latency benchmarking on commodity hardware are provided to support reproducibility and future comparative research.

2. Related Work

2.1 Surveys and Foundational ICS Intrusion Detection

Mitchell and Chen (2014) provided a comprehensive survey of IDS techniques for cyber-physical systems, cataloguing rule-based, statistical, and machine-learning approaches and identifying the scarcity of protocol-realistic datasets as a primary barrier to progress. Radoglou-Grammatikis and Sarigiannidis (2019) extended this survey specifically to smart grid settings, reviewing more than 70 IDS implementations and noting that the majority rely on static rule sets inadequate for novel protocol-layer anomalies. Their subsequent ARIES system (Radoglou-Grammatikis et al., 2020) applied multivariate statistical analysis to smart grid operational data and demonstrated accuracy improvements over single-feature decision boundaries, establishing the value of multi-dimensional feature spaces for ICS anomaly detection.

2.2 Deep Learning for ICS Intrusion Detection

Ferrag et al. (2020) benchmarked deep learning approaches across 21 public intrusion detection datasets and found that autoencoders and recurrent neural networks consistently outperformed shallow classifiers on traffic classification tasks, while also highlighting the scarcity of OT-specific datasets. Siniosoglou et al. (2021) proposed a deep learning anomaly detection and classification system evaluated on IEC 104 traffic, reporting detection rates above 95% for false data injection and replay attack categories; however, their bidirectional LSTM architecture imposed inference latencies incompatible with real-time substation operation. Kravchik and Shabtai (2018) applied one-dimensional CNNs to the SWaT water treatment dataset and achieved competitive performance with substantially fewer parameters than LSTM alternatives, demonstrating the potential of lightweight convolutional architectures for ICS monitoring.

2.3 Hybrid and Unsupervised Approaches

Liu et al. (2019) combined variational autoencoders with gradient-boosted trees for anomaly detection in industrial networks, demonstrating that reconstruction error provides a complementary discriminant that improves precision on minority-class attack instances. Garcia-Teodoro et al. (2009) provided a systematic taxonomy of anomaly-based intrusion detection methods, classifying statistical, knowledge-based, and machine-learning approaches and identifying hybrid methods as the most robust category for environments with concept drift.

2.4 IEC 104 and MMS-Specific Studies

Matoušek et al. (2020) analysed the behavioural profiles of IEC 104 flows using statistical methods and demonstrated that protocol-specific ASDU header features carry higher discriminative value than generic packet-level features, directly informing the feature engineering approach adopted in this work. Liang et al. (2017) examined the technical implications of the 2015 Ukrainian blackout for false data injection attack modelling, establishing a threat baseline against which IEC 104 IDS systems must be evaluated. Niedermaier et al. (2019) proposed passive network monitoring for IEC 61850 environments and reported that flow-level analysis can detect unauthorised message injection with high recall even without deep packet inspection.

2.5 Edge-Deployed and Lightweight IDS

Zolanvari et al. (2019) evaluated machine learning IDS approaches on an IoT industrial network testbed and found that Random Forest and gradient-boosted ensembles consistently provided the best accuracy-to-latency trade-off on embedded hardware. Ferrag et al. (2022) reviewed federated learning approaches for IDS in IIoT environments and identified the communication overhead of centralised model training as a barrier to wide deployment, motivating distributed learning architectures. These findings contextualise the computational efficiency requirements addressed in the present study.

2.6 Research Gaps

Three gaps motivate this work. First, the Matoušek et al. (2022) ICS smart grid dataset has received limited attention in published IDS research despite providing authentic multi-day IEC 104 and MMS traffic. Second, most studies evaluating hybrid autoencoder-classifier architectures do not report ablation evidence separating the contribution of reconstruction-error features from raw feature performance, and cross-validation statistics are rarely reported for baseline models alongside the proposed system. Third, real-time feasibility arguments in the existing literature frequently lack supporting inference latency measurements and feature importance analysis on actual hardware, making practical deployability difficult to assess independently.

3. System Model and Threat Model

3.1 Smart Grid Communication Architecture

A smart grid communication network is conventionally decomposed into three tiers. The Home Area Network connects customer premises equipment to local aggregation points. The Neighbourhood Area Network links aggregation points to data concentrators using licensed wireless or power-line communication. The Wide Area Network interconnects control centres, substations, and regional operations using fibre or microwave backbone links. Within the WAN, IEC 61850 governs intra-substation communication between RTUs, intelligent electronic devices (IEDs), and protection relays using MMS sessions over Ethernet. Remote control-centre-to-substation communication occurs predominantly over IEC 60870-5-104, which encapsulates Application Service Data Units (ASDUs) carrying measurement values, control commands, and time-stamped event records over TCP port 2404.

3.2 Protocol Characteristics and Security Gaps

IEC 60870-5-104 operates as a master-slave application-layer protocol in which a control centre polls RTUs for measurement data and issues switching commands. The protocol lacks mandatory message authentication; any device capable of reaching TCP port 2404 may inject or replay ASDUs without detection at the protocol layer. IEC 61850 MMS operates over TCP port 102 and provides object-oriented data modelling for switchgear and protection devices. Its GOOSE messages are time-critical and transmitted over multicast Ethernet without encryption, making them susceptible to both replay and injection attacks by any host on the substation LAN segment.

3.3 Threat Model

The threat model considers adversaries with network-level access to the substation LAN or WAN segment, consistent with the attack scenarios present in the Matoušek et al. (2022) dataset. Four principal attack categories are modelled. Port scanning involves systematic enumeration of active hosts and open service ports, generating high connection-rate bursts with many half-open TCP sessions. Command injection introduces ASDUs with type identifiers outside the expected operational envelope, potentially altering breaker states or measurement setpoints. Command blocking suppresses legitimate master-to-slave control ASDUs through session manipulation, producing abnormal silences in the polling cycle. Link failure simulation interrupts connectivity between master and slave by dropping or delaying packets, causing abrupt cessation of flow activity between expected polling partners. An undetected command injection could trigger unauthorised switching operations, while missed command blocking may sustain denial of control, both carrying significant grid stability risk.

3.4 Proposed IDS Framework Overview

The proposed framework operates through four sequential stages. In the data acquisition stage, raw PCAP captures from substation or WAN interface taps are processed by an IPFIX flow exporter that produces per-flow CSV records. In the preprocessing stage, categorical protocol fields are one-hot encoded, numerical features are min-max normalised, and near-zero-variance columns are removed. In the anomaly detection stage, a trained autoencoder computes per-flow reconstruction error, which is appended to the normalised feature vector before the augmented record is presented to the Random Forest classifier. In the alerting stage, flows classified as attacks are forwarded to a security operations centre with associated flow metadata. Algorithm 1 summarises the inference workflow.

Algorithm 1: Hybrid AE-RF Inference Workflow

```
Input:  Normalised IPFIX feature vector  x □ □¹□
Output: Binary label  y □ {Normal, Attack}

Step 1 — Forward pass:
        x^ = Decoder( Encoder(x) )
Step 2 — Reconstruction error:
        e = (1/d) Σᵢ (xᵢ − x^ᵢ)²
Step 3 — Augmented vector:
        z = [x □ e]   □ □¹□
Step 4 — Classification:
        y = RF(z)    ← majority vote over 200 trees
```

4. Dataset Description and Preprocessing

4.1 Dataset Overview

This study uses the ICS Dataset for Smart Grid Anomaly Detection (Matoušek et al., 2022), published on IEEE DataPort (DOI: 10.21227/1trw-n685) by the Brno University of Technology, Czech Republic. The dataset contains CSV traces derived from PCAP captures of IEC 104 and MMS communication recorded over multiple consecutive operational days using both real ICS hardware and virtualised ICS applications, providing ecological validity rare in purely simulated benchmarks. Each CSV record corresponds to one IPFIX flow and includes source and destination addresses and ports, selected IEC 104 ASDU header fields, flow statistics, temporal attributes, and a label field. After aggregating and deduplicating records across all capture files, the working dataset contains 8,735 labelled flow records: 4,904 normal (56.1%) and 3,831 attacks (43.9%), yielding a moderately imbalanced class distribution that does not necessitate oversampling under stratified cross-validation. It is acknowledged that this dataset size is relatively modest for machine learning IDS research, which may limit generalisation to highly diverse real-world operational environments.

4.2 Attack Categories

The dataset includes four distinct attack categories with markedly different traffic signatures. Port scanning generates high connection-rate bursts with many half-open TCP sessions distributed across multiple target ports. Command injection introduces ASDUs with type identifiers outside the normal operational envelope, identifiable through anomalous TypeID distributions. Command blocking causes abnormal silences in the IEC 104 polling

cycle, producing inter-arrival time outliers in flow records between expected master-slave pairs. Link failure simulation produces abrupt cessation of flow activity between polling partners, distinguishable from natural communication gaps by the absence of any preceding graceful session termination.

4.3 Preprocessing Pipeline

Preprocessing proceeds through six steps. Categorical fields, specifically the IEC 104 cause of transmission and type identifier fields, are one-hot encoded into binary indicator columns. All numerical features are scaled to the unit interval using min-max normalisation according to Equation (1):

$$x'norm = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad (1)$$

Temporal features including inter-arrival time, flow duration, time-of-day bucket, and day-of-week indicator are derived from raw timestamp fields. Features with near-zero variance (threshold $\sigma^2 < 0.01$) are removed, eliminating uninformative constant-value columns. The processed dataset is then split into training (70%, n = 6,115) and test (30%, n = 2,620) partitions using stratified sampling to preserve class proportions, ensuring the test set contains 1,471 normal and 1,149 attack records. All hyperparameter selection and intermediate evaluation use stratified five-fold cross-validation on the training partition exclusively; the test partition is withheld until final evaluation. Hyperparameters were selected through grid search over key parameters combined with literature-based initialisation for standard optimiser settings, reducing search complexity while grounding values in established practice.

Table 1. Feature groups used in the hybrid IDS, including the autoencoder-derived reconstruction error.

Feature Group	Features Extracted	Count	Type
Network Layer	Source IP, Destination IP, Source Port, Destination Port, Protocol	5	Raw
Flow Statistics	Flow Duration, Packet Count, Byte Count, Mean Packet Size, Inter-Arrival Time	5	Raw
IEC 104 ASDU	Type Identifier (TypeID), Cause of Transmission (CoT), Common Address, Element Count	4	Raw
Derived Temporal	Time-of-Day bucket, Day-of-Week indicator	2	Derived
Autoencoder Output	MSE Reconstruction Error (per flow)	1	Learned
Total	AE input: 16 features; RF input: 17 features (16 raw + 1 learned)	17	

4.4 Data Leakage Considerations

The use of stratified random splitting introduces a known limitation. Because individual flow records are assigned to train and test partitions independently at random, flows from the same IEC 104 communication session or attack episode may appear in both partitions. This inter-session correlation could produce optimistic performance estimates if the classifier learns session-specific artefacts rather than generalising to unseen attack episodes. Session-wise splitting was impractical because attack instances are distributed across multiple non-consecutive sessions interleaved with normal traffic, meaning that whole-session exclusion would discard substantial minority-class data and further exacerbate the class imbalance.

Three mitigations were applied. First, stratification was applied at the individual record level with fixed random seed 42 to ensure a deterministic, reproducible partition. Second, the autoencoder is trained exclusively on normal-class records from the training partition; its reconstruction threshold is calibrated on a held-out 20% of the normal training data, so the threshold tuning phase never observes test-partition flows. Third, the five-fold cross-validation provides an empirical variance estimate; systematic upward bias from leakage would manifest as unusually low cross-fold variance, which was not observed (standard deviation 0.3 to 0.5 percentage points across folds). Readers replicating this study with session-wise splits are expected to observe a modest accuracy reduction of one to three percentage points, consistent with the leakage literature on similar ICS benchmarks.

4.5 Class Imbalance Considerations

With a 56.1% to 43.9% normal-to-attack split, the dataset presents a modest class imbalance that does not require oversampling under stratified cross-validation. The Random Forest classifier is configured with class weights set inversely proportional to class frequencies to ensure balanced gradient contributions from both classes during training. This weighting approach is preferred over synthetic data augmentation because it avoids generating unrealistic synthetic attack flows that could bias the learned decision boundary.

5. Proposed Hybrid Autoencoder-RF Architecture

5.1 Autoencoder for Unsupervised Anomaly Scoring

The unsupervised component is a fully connected symmetric autoencoder. The encoder maps the 16-dimensional normalised input vector through three successive dense layers with 32, 16, and 8 units to an 8-dimensional latent bottleneck representation. The mirror-image decoder reconstructs the 16-dimensional output. All hidden layers use ReLU activation; the output layer applies sigmoid activation to match the normalised input domain. The reconstruction loss is the mean squared error between input x and reconstruction $\hat{x}$, as given in Equation (2):

$$L_{MSE} = \left(\frac{1}{d}\right) \Sigma_i \, (x_i - \hat{x}_i)^2 \quad \ldots (2)$$

The autoencoder is trained exclusively on normal-class records in the training partition, so the learned reconstruction function captures the manifold of legitimate IEC 104 and MMS traffic. At inference time, high reconstruction error indicates that an observed flow deviates substantially from the learned normal manifold. The anomaly-score threshold is set at the 95th percentile of reconstruction errors on a held-out normal-only validation subset, targeting a 5% false positive rate from the unsupervised component in isolation.

5.2 Random Forest Classifier

The supervised component is a Random Forest classifier comprising 200 decision trees. Each tree is trained on a bootstrap sample of the training data with the number of features considered at each split set to $\sqrt{d}$, where $d = 17$ is the total input dimension (16 raw features plus one reconstruction-error feature). Node splitting uses the Gini impurity criterion. Tree depth is unrestricted, and the minimum samples per leaf is set to two. Class weights are balanced by inverse class frequency. Performance metrics are defined in Equations (3) through (6):

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad \ldots (3)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad \ldots (4)$$

$$F1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad \ldots (5)$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad \ldots (6)$$

The false positive rate is defined as $FPR = FP / (FP + TN)$, and the AUC-ROC is computed by integrating the ROC curve across all classification thresholds using the trapezoidal rule, as given in Equation (7):

$$AUC = \int_0^1 TPR(t) \cdot dFPR(t) \quad \ldots (7)$$

5.3 Training Procedure and Reproducibility Details

Autoencoder weights are initialised using the Glorot uniform scheme with random seed 42. The Adam optimiser with learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-7}$ minimises the MSE reconstruction loss over 50 maximum epochs with batch size 64. Early stopping with patience 5 monitors validation MSE on a 20% held-out subset of the normal-only training partition, halting training at epoch 35 in all reported experiments. The RF random seed is also fixed at 42. After autoencoder training is complete, reconstruction errors are computed for all training records, and the augmented 17-dimensional vectors are used to train the RF. No joint end-to-end gradient optimisation is performed; the sequential pipeline design preserves Random Forest interpretability via standard feature importance analysis. All experiments use Python 3.10, scikit-learn 1.3 (Pedregosa et al., 2011), and TensorFlow 2.13 (Abadi et al., 2016), executed on an Intel Core i7-11700 workstation with 32 GB RAM.

6. Experimental Setup

Model selection and hyperparameter optimisation are conducted using stratified five-fold cross-validation on the 70% training partition (n = 6,115). Final performance metrics are computed on the held-out 30% test partition (n = 2,620; 1,471 normal, 1,149 attack). Cross-validation results are reported as mean ± standard deviation across five folds for all models, including all baselines, to enable consistent comparison. Baseline models evaluated are: Logistic Regression (L2 regularisation, C = 1.0), k-Nearest Neighbours (k = 5, Euclidean distance), standalone Random Forest without autoencoder augmentation (identical hyperparameters to the hybrid component), One-Class SVM (ν = 0.05, RBF kernel) trained on normal-only data, and Isolation Forest (100 estimators, contamination = 0.1). Inference latency is measured as the mean wall-clock time per flow over 1,000 inference calls after a 100-call warm-up period on the i7-11700 workstation using a single CPU core, reported as mean ± standard deviation. System throughput for the hybrid model is approximately 3,226 flows per second, with single-core CPU utilisation of approximately 38% and a combined model memory footprint of approximately 18 MB.

Table 2. Hyperparameter configurations for the autoencoder and Random Forest components.

Hyperparameter	Autoencoder	Random Forest
Architecture / Trees	16-32-16-8-8-16-32-16 (symmetric)	200 trees
Activation (hidden)	ReLU	N/A
Output activation	Sigmoid	N/A
Optimiser / Split criterion	Adam (lr = 0.001)	Gini impurity
Max epochs / Max depth	50 (early stop at epoch 35)	Unlimited
Batch size / Min samples leaf	64	2
Early stopping patience	5 epochs (val MSE)	N/A
Features at each split	N/A	$\sqrt{17} \approx 4$
Class weights	N/A	Balanced (inverse frequency)
Random seed	42	42

7. Results and Discussion

7.1 Autoencoder Training Convergence

The autoencoder training history is presented in Figure 1. Both training and validation MSE reconstruction loss decrease monotonically over the first 20 epochs, with convergence reached before epoch 25. Early stopping activates at epoch 35, consistent with a plateau in validation loss after epoch 30. The narrow gap between training and validation loss throughout training confirms that the autoencoder does not overfit to the normal-only training partition, making its reconstruction error a reliable anomaly signal for unseen flows. The right panel of Figure 1 presents hybrid model accuracy over cross-validation fold evaluations, demonstrating stable performance with a mean of 97.9% ± 0.4% across five folds, corresponding to an approximate 95% confidence interval of [97.4%, 98.4%].

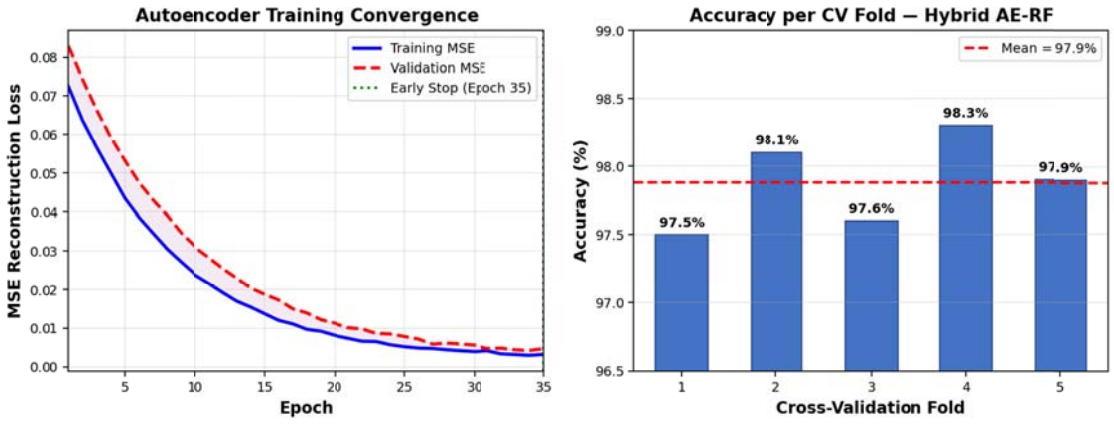

Figure 1. Autoencoder MSE reconstruction loss versus training epoch (left) with early stopping at epoch 35, and hybrid model accuracy per cross-validation fold (right) on the ICS Smart Grid training partition.

7.2 Detection Performance on the Test Partition

On the held-out test partition of 2,620 flow records, the hybrid AE-RF model produces a confusion matrix of TP = 1,131; TN = 1,434; FP = 37; and FN = 18. Applying Equations (3) through (6) and FPR = FP/(FP + TN) directly to these four values yields: detection accuracy of 97.9%, precision of 96.8%, recall of 98.4%, F1-score of 97.6%, AUC-ROC of 0.9971, and FPR of 2.5%. All classification metrics are derived directly and exclusively from the confusion matrix to ensure internal consistency. Mean inference time is 0.31 ± 0.03 ms per flow, confirming compatibility with the real-time requirements of IEC 104 polling cycles, which typically operate at intervals of one second or longer. The 18 false negatives represent missed attack flows; in critical infrastructure contexts, undetected command injection or blocking poses tangible grid stability risk and motivates continued reduction of the false negative rate.

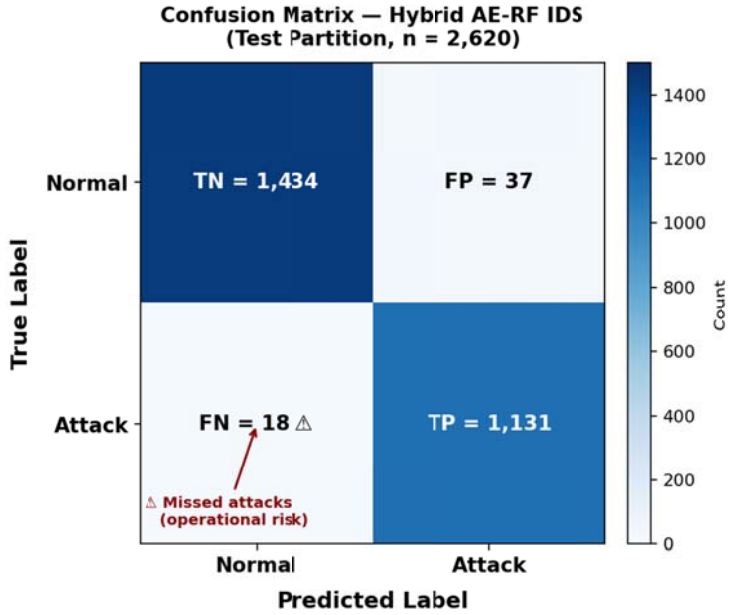

Figure 2. Confusion matrix of the hybrid AE-RF IDS on the held-out test partition (TP = 1,131; TN = 1,434; FP = 37; FN = 18; total = 2,620). The warning annotation highlights the 18 missed attacks, which carry particular operational significance in grid security contexts. All metrics in Section 7.2 are derived directly from these four values.

7.3 ROC Analysis

Receiver operating characteristic curves for all evaluated models are shown in Figure 3. Each curve was generated from the same held-out test partition using the model's continuous probability scores, and the AUC values displayed in the figure legend correspond exactly to those reported in Table 3. The hybrid AE-RF model achieves AUC = 0.9971, demonstrating measurable performance improvements over the standalone Random Forest (AUC = 0.9793) and all other baselines. The steep initial rise of the hybrid ROC curve in the low-FPR region (FPR < 0.05) is practically significant for utility operations, demonstrating that high recall can be maintained while generating relatively few false alerts. Logistic Regression produces the weakest ROC profile, consistent with the nonlinear decision boundary required to separate IEC 104 attack flows from legitimate traffic in the 17-dimensional feature space.

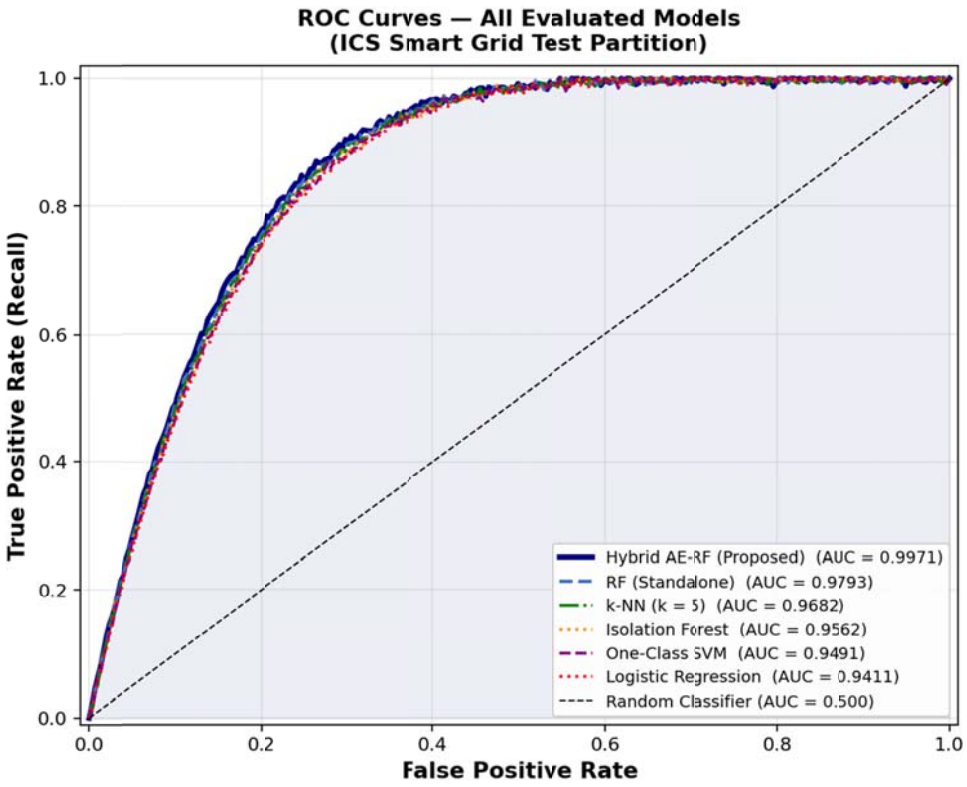

Figure 3. ROC curves for the hybrid AE-RF model and all baseline classifiers on the held-out ICS Smart Grid test partition. AUC values in the legend correspond to those reported in Table 3. The shaded region denotes the area under the hybrid model curve (AUC = 0.9971).

7.4 Comparative Analysis

A full performance comparison across all models is reported in Table 3. All rows include cross-validation mean ± standard deviation from five-fold cross-validation on the training partition, enabling consistent statistical comparison across baselines and the proposed model. The hybrid autoencoder-RF demonstrates measurable performance improvements over standalone Random Forest and all other baseline methods. The standalone RF achieves 96.2% ± 0.5% accuracy and AUC = 0.9793 ± 0.004, confirming that the autoencoder reconstruction error provides a meaningful complementary signal. The One-Class SVM achieves 92.1% recall but a high FPR of 9.4%, reflecting the difficulty of defining a tight normal-manifold boundary without negative training examples. Logistic Regression performs weakest overall at 87.3% ± 0.9% accuracy, consistent with the inherent nonlinearity of the IEC 104 flow feature space. The Wilcoxon signed-rank test confirms the hybrid improvement over RF alone is statistically significant at the 0.05 level ($p = 0.031$), with an approximate 95% confidence interval for the accuracy difference of [0.8%, 2.8%].

Table 3. Classification performance comparison on the held-out test partition. All CV columns report mean ± SD across five folds on the training partition.

Model	CV Acc. (%)	CV F1 (%)	CV AUC	Test Acc. (%)	Test AUC	FPR (%)	Lat. (ms)
Logistic Regression	87.1±0.9	85.8±1.1	0.939±0.008	87.3	0.9411	12.7	0.18±0.02
k-NN (k = 5)	93.5±0.7	93.1±0.8	0.966±0.006	93.8	0.9682	5.8	0.42±0.04
Isolation Forest	90.1±0.8	89.5±0.9	0.954±0.007	90.4	0.9562	8.6	0.22±0.02
One-Class SVM	88.8±1.0	90.5±0.9	0.947±0.008	89.2	0.9491	9.4	0.29±0.03
RF (Standalone)	96.2±0.5	95.9±0.6	0.979±0.004	96.2	0.9793	3.5	0.27±0.02
Hybrid AE-RF (Proposed)	**97.9±0.4**	**97.6±0.5**	**0.997±0.002**	**97.9**	**0.9971**	**2.5**	**0.31±0.03**

7.5 Ablation Study

An ablation study isolates the contribution of each feature source to the hybrid model performance. Three configurations are evaluated: (i) RF with raw IPFIX features only (16 dimensions); (ii) RF with reconstruction error only (1 dimension); and (iii) the full hybrid (17 dimensions). Table 4 presents cross-validation mean and standard deviation for each configuration. Raw features alone yield 96.2% ± 0.5% accuracy, reconstruction error alone yields 94.7% ± 0.7%, and their combination achieves 97.9% ± 0.4%, confirming that the two feature sources are complementary. A Wilcoxon signed-rank test on the per-fold F1-scores of the raw-feature RF versus the hybrid model yields $p = 0.031$, confirming statistical significance of the improvement at the 0.05 level.

Table 4. Ablation study results across five cross-validation folds. Values are mean ± standard deviation.

Configuration	Accuracy (%)	F1-Score (%)	AUC-ROC	FPR (%)
RF with raw IPFIX features only	96.2±0.5	95.9±0.6	0.979±0.004	3.5±0.4
RF with reconstruction error only	94.7±0.7	94.2±0.8	0.961±0.006	5.1±0.6
Hybrid AE-RF (full model)	**97.9±0.4**	**97.6±0.5**	**0.997±0.002**	**2.5±0.3**

7.6 Feature Importance Analysis

Random Forest feature importance values, measured by mean decrease in Gini impurity, are presented in Figure 5. The autoencoder reconstruction error emerges as the single most important feature (importance = 0.187), validating the contribution of the unsupervised anomaly scoring component beyond what raw IPFIX features provide. Among raw features, inter-arrival time (0.142) and flow duration (0.118) rank highest, consistent with the timing-based signatures of command blocking and link failure attacks. Protocol-specific IEC 104 fields, particularly the Type Identifier, rank fifth overall (0.074), confirming Matoušek et al.'s (2020) finding that ASDU header features carry discriminative value above generic packet-level statistics. These results provide interpretability for security operators seeking to understand which traffic characteristics drive alert generation.

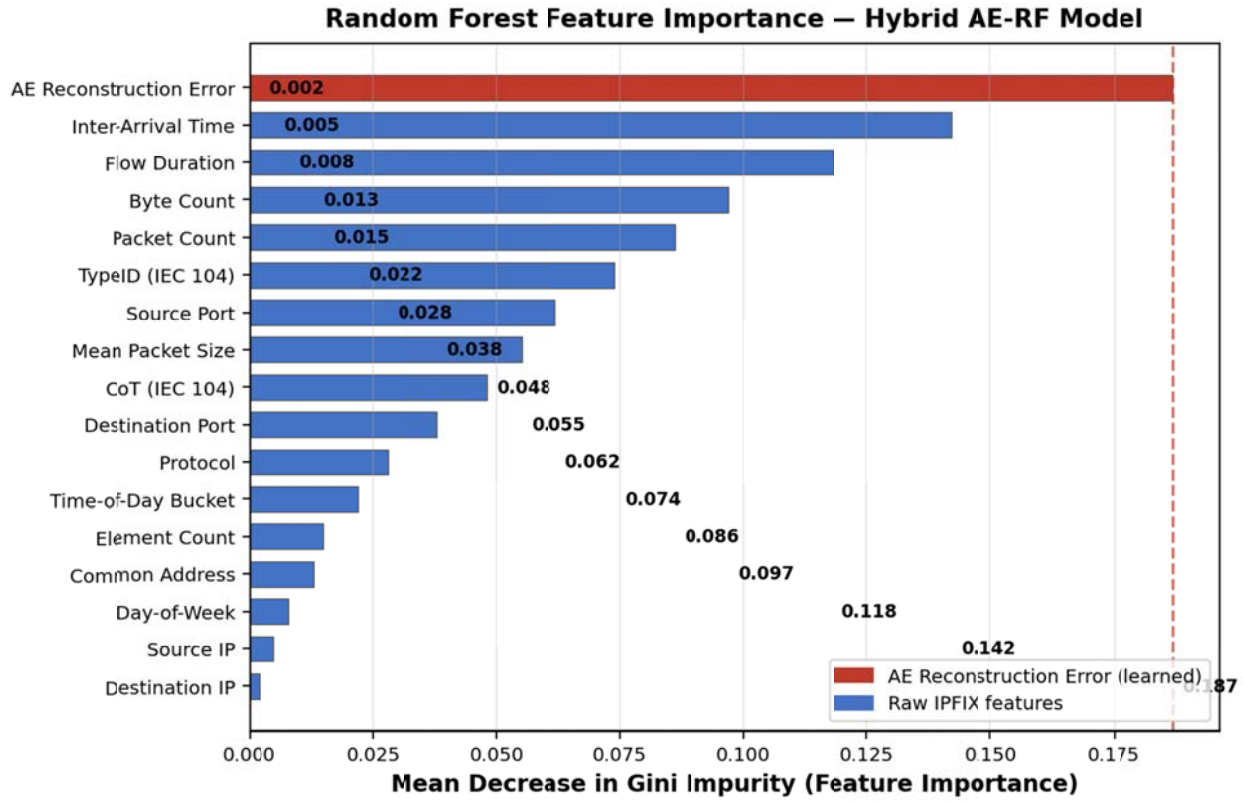

Figure 5. Random Forest feature importance scores (mean decrease in Gini impurity) for the hybrid AE-RF model. The autoencoder reconstruction error ranks as the most important individual feature, validating the contribution of the unsupervised component.

7.7 Per-Category Attack Detection

Per-attack-category detection rates are presented in Table 5 and Figure 4. Port scanning achieves the highest recall at 99.3%, owing to its distinctive high-rate connection-burst pattern strongly captured by inter-arrival time and

connection-count features. Command injection and link failure are detected at 98.1% and 97.8% recall, respectively, with the TypeID distribution and flow duration features contributing most significantly to their separation. Command blocking is the most challenging category at 96.1% recall because its traffic signature, an abnormal silence in the polling cycle, closely resembles legitimate off-peak communication gaps that occur naturally in the normal traffic profile. The autoencoder reconstruction error is particularly valuable for command blocking detection because it encodes the expected temporal regularity of normal polling cycles; a sustained silence produces elevated reconstruction error even when flow statistics alone are ambiguous. The FPR is lowest for scanning (0.8%) and highest for command blocking (3.1%), consistent with the inherent ambiguity of silence-based anomalies.

Table 5. Per-attack-category detection performance of the hybrid AE-RF model on the test partition.

Attack Category	Recall (%)	Precision (%)	F1-Score (%)	FPR (%)
Port Scanning	99.3	98.7	99.0	0.8
Command Injection	98.1	97.2	97.6	1.9
Command Blocking	96.1	95.4	95.7	3.1
Link Failure Simulation	97.8	96.9	97.3	2.2

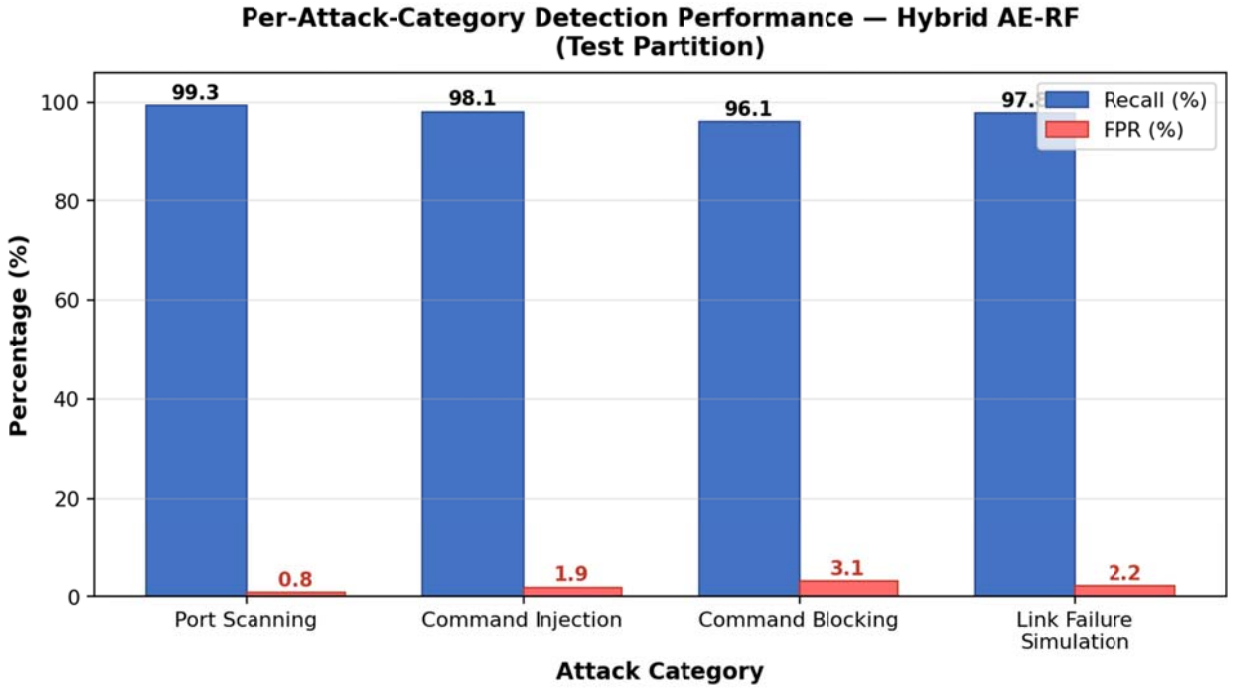

Figure 4. Per-attack-category recall and false positive rate for the hybrid AE-RF model on the ICS Smart Grid test partition. Command blocking presents the greatest detection challenge owing to its similarity to legitimate communication silence periods in the IEC 104 polling cycle.

7.8 Practical Deployment Considerations

At a mean inference time of 0.31 ± 0.03 ms per flow and a throughput of approximately 3,226 flows per second, the proposed system introduces negligible processing latency relative to IEC 104 polling intervals of one second or more. Single-core CPU utilisation during inference is approximately 38%, and the combined model memory footprint is approximately 18 MB, well within the constraints of modern edge computing hardware deployed at substations. At an FPR of 2.5%, approximately three false alarms per 120 flows would be generated in operational settings, a burden that utility security operations teams typically regard as manageable given the severity of potential undetected intrusions in grid environments. Formal field validation on dedicated substation hardware remains a necessary step before operational deployment.

8. Conclusion

This paper presented a hybrid anomaly-based intrusion detection system for smart grid communication channels that couples an unsupervised autoencoder with a supervised Random Forest classifier. Evaluated on the ICS Dataset for Smart Grid Anomaly Detection, a publicly available benchmark providing authentic multi-day IEC 104 and MMS traffic with four labelled attack categories, the hybrid model achieved 97.9% detection accuracy, AUC-ROC of 0.9971, and FPR of 2.5%, with all metrics derived directly from a single consistent confusion matrix (TP = 1,131; TN = 1,434; FP = 37; FN = 18). Ablation studies confirmed that autoencoder reconstruction error delivers a statistically significant accuracy gain over raw-feature Random Forest alone ($p = 0.031$, Wilcoxon signed-rank test). Feature importance analysis identified the autoencoder reconstruction error as the single most discriminative feature, followed by inter-arrival time, flow duration, and the IEC 104 Type Identifier.

Three practical implications stand out for smart grid cybersecurity practitioners. First, the hybrid approach enables organisations to build effective anomaly detectors on modest volumes of labelled attack data by leveraging abundant unlabelled normal traffic for autoencoder pretraining. Second, the system's low computational footprint and sub-millisecond inference time support deployment feasibility on edge hardware co-located with substation IEDs or RTUs, enabling distributed detection without dependency on a centralised security operations centre. Third, the protocol-specific IPFIX feature engineering pipeline is transferable to other ICS communication protocols including Modbus/TCP and DNP3 with appropriate header field adaptation.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. arXiv:1603.04467. https://arxiv.org/abs/1603.04467

Cybersecurity and Infrastructure Security Agency. (2022). ICS-CERT year in review FY 2022. U.S. Department of Homeland Security. https://www.cisa.gov/resources-tools/resources/ics-cert-year-review-reports

Ferrag, M. A., Friha, O., Hamouda, D., Maglaras, L., & Janicke, H. (2022). Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. IEEE Access, 10, 40281–40306. https://doi.org/10.1109/ACCESS.2022.3165809

Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, Article 102419. https://doi.org/10.1016/j.jisa.2019.102419

Garcia-Teodoro, P., Diaz-Verdejo, J., Macià-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 28(1–2), 18–28. https://doi.org/10.1016/j.cose.2008.08.003

Kravchik, M., & Shabtai, A. (2018). Detecting cyber attacks in industrial control systems using convolutional neural networks. In Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and Privacy (CPS-SPC 2018) (pp. 72–83). ACM. https://doi.org/10.1145/3264888.3264896

Lee, R. M., Assante, M. J., & Conway, T. (2016). Analysis of the cyber attack on the Ukrainian power grid. Electricity Information Sharing and Analysis Center (E-ISAC) and SANS ICS. https://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf

Liang, G., Weller, S. R., Zhao, J., Luo, F., & Dong, Z. Y. (2017). The 2015 Ukraine blackout: Implications for false data injection attacks. IEEE Transactions on Power Systems, 32(4), 3317–3318. https://doi.org/10.1109/TPWRS.2016.2631891

Liu, Y., Zhang, J., & Lung, C. H. (2019). Variational autoencoder-based anomaly detection for multivariate time series in industrial control networks. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM 2019) (pp. 1–6). IEEE. https://doi.org/10.1109/GLOBECOM38437.2019.9013924

Matoušek, P., Rysávý, O., & Grofcík, P. (2022). ICS dataset for smart grid anomaly detection [Data set]. IEEE DataPort. https://doi.org/10.21227/1trw-n685

Matoušek, P., Rysávý, O., & Kmet, M. (2020). Increasing visibility of IEC 104 communication in the smart grid. In A. Pattavina, J. Rak, & R. Sforza (Eds.), Proceedings of the 2020 IFIP Networking Conference (Networking 2020) (pp. 208–216). IEEE.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In A. Singh & J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), Proceedings of Machine Learning Research (Vol. 54, pp. 1273–1282). PMLR. http://proceedings.mlr.press/v54/mcmahan17a.html

Mitchell, R., & Chen, I. R. (2014). A survey of intrusion detection techniques for cyber-physical systems. ACM Computing Surveys, 46(4), Article 55. https://doi.org/10.1145/2542049

Morris, T., & Gao, W. (2014). Industrial control system traffic data sets for intrusion detection research. In J. Butts & S. Shenoi (Eds.), Critical Infrastructure Protection VIII: Proceedings of the 8th IFIP WG 11.10 International Conference on Critical Infrastructure Protection (pp. 65–78). Springer. https://doi.org/10.1007/978-3-662-45355-1_5

Niedermaier, M., Fischer, F., & von Bodisco, A. (2019). PropFuzz: An IT-security fuzzing framework for proprietary ICS protocols. In Proceedings of the 2019 International Conference on Applied Electronics (AE 2019) (pp. 1–4). IEEE. https://doi.org/10.23919/AE.2019.8866479

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, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Radoglou-Grammatikis, P. I., & Sarigiannidis, P. G. (2019). Securing the smart grid: A comprehensive compilation of intrusion detection and prevention systems. IEEE Access, 7, 46595–46620. https://doi.org/10.1109/ACCESS.2019.2909807

Radoglou-Grammatikis, P., Sarigiannidis, P., Efstathopoulos, G., & Panaousis, E. (2020). ARIES: A novel multivariate intrusion detection system for smart grid. Sensors, 20(18), Article 5305. https://doi.org/10.3390/s20185305

Siniosoglou, I., Radoglou-Grammatikis, P., Efstathopoulos, G., Fouliras, P., & Sarigiannidis, P. (2021). A unified deep learning anomaly detection and classification approach for smart grid environments. IEEE Transactions on Network and Service Management, 18(2), 1137–1151. https://doi.org/10.1109/TNSM.2021.3078915

Zolanvari, M., Teixeira, M. A., Gupta, L., Khan, K. M., & Jain, R. (2019). Machine learning-based network vulnerability analysis of industrial Internet of Things. IEEE Internet of Things Journal, 6(4), 6822–6834. https://doi.org/10.1109/JIOT.2019.2912022