Anusandhanvallari
Vol 2025, No.1
January 2025
ISSN 2229-3388

Autonomous Anomaly Edge Verification Via Symmetric Deep Autoencoders with Post-Hoc Interpretability for Zero-Day Threat Isolation

Aravind Chagantipati

Department of Computer Science and Engineering, Kennedy University

Abstract: Contemporary enterprise network environments are increasingly exposed to sophisticated zero-day exploits that bypass traditional signature-based security perimeters due to the lack of historical threat profiles. To address this paradigm shift, unsupervised machine learning presents a viable defense strategy by modeling the intrinsic statistical behavior of legitimate network traffic and identifying structural anomalies as security breaches. This paper introduces an unsupervised perimeter defense framework utilizing symmetric deep Autoencoder architectures. Evaluated on the standardized NSL-KDD benchmark, our model is trained exclusively on uncompromised network transactions to establish an optimized baseline behavior. During testing, hidden zero-day threats introduce distinct reconstruction discrepancies, yielding an optimal accuracy of 93.2% and an F1-score of 92.7%. Furthermore, to mitigate the opaque nature inherent to multi-layered neural configurations, we integrate a Local Interpretable Model-agnostic Explanations (LIME) framework. This post-hoc layer maps elevated reconstruction errors to granular, interpretable feature attributions, empowering analysts with immediate structural visibility into flagged vectors.

Keywords: Deep Autoencoders, Unsupervised Machine Learning, Network Intrusion Detection, Zero-Day Vulnerabilities, Explainable AI (XAI), Local Interpretable Model-agnostic Explanations (LIME), Reconstruction Discrepancy.

1. INTRODUCTION

The geometric expansion of modern cloud frameworks and decentralized corporate digital topologies has radically intensified the attack surface available to malicious actors. A critical vulnerability confronting modern Security Operations Centers (SOCs) stems from zero-day exploits—novel intrusion strategies engineered to weaponize unpatched system flaws. Because these security threats possess no previously recorded signatures, classical supervised machine learning paradigms are structurally incapable of detecting them, given their reliance on explicit historical boundaries and pre-labeled threat vectors.

Unsupervised learning circumvents this vulnerability by redefining the foundational objective of the classifier. Rather than attempting to characterize a volatile and expanding catalog of attacks, unsupervised methodologies map the mathematical distribution defining baseline network operations. Any significant departure from this normative space is flagged as an anomaly. Symmetric Deep Autoencoders (AEs) are exceptionally well-suited for this implementation, projecting high-dimensional traffic features into a constrained bottleneck layer before attempting reconstruction. Exposed to zero-day deviations, the network encounters heightened reconstruction errors, facilitating accurate threat isolation without requiring prior training annotations. However, the multi-layered latent representations operate as an uninterpretable "black box," masking the underlying structural features driving the alert. This paper implements an explainable, high-fidelity unsupervised pipeline, delivering robust threat identification on the NSL-KDD dataset coupled with actionable, post-hoc local feature attribution through LIME.

2. LITERATURE REVIEW AND TECHNICAL CONTEXT

Prior academic work regarding unsupervised network infrastructure surveillance has thoroughly documented the effectiveness of mathematical compression systems in isolating structural anomalies. Early iterations

frequently leveraged shallow architectures, such as Principal Component Analysis (PCA), to linearly project network statistics, categorizing feature vectors exhibiting high reconstruction variance as suspicious. Nonetheless, linear projections fail to capture the highly complex, multi-variable dependencies characteristic of modern data streams. The advent of deep neural Autoencoders effectively circumvented this barrier by employing deeply stacked bottleneck vectors to approximate complex, non-linear distributions. Despite these technical developments, contemporary network security frameworks persistently struggle to interpret high reconstruction loss metrics into clear, human-auditable feature metrics. The architecture proposed herein directly addresses this gap.

3. MATHEMATICAL FORMULATION AND ARCHITECTURAL DESIGN

The proposed unsupervised framework leverages a symmetrical deep neural pipeline partitioned into an encoding operator φ and a decoding reconstruction operator ψ. The encoder yields a compressed bottleneck vector $z \in \mathbb{R}^m$ from a high-dimensional input vector $x \in \mathbb{R}^d$ (where $m \ll d$), formally expressed as:

$$z = \sigma(W^e x + b^e) \quad (1)$$

where W^e and b^e designate the weight matrices and bias vectors of the encoder layers, respectively, and σ denotes the non-linear activation function. Subsequently, the decoder maps the latent state z back into the original input dimensions, generating a reconstructed vector $\hat{x} \in \mathbb{R}^d$:

$$\hat{x} = \sigma(W^d z + b^d) \quad (2)$$

where W^d and b^d represent the corresponding weight and bias parameters of the decoder pipeline. Optimization is performed exclusively on clean, uncompromised traffic records by minimizing the Mean Squared Error (MSE) objective function:

$$\&Lagr;(x, \hat{x}) = (1/d) \sum_{i=1}^{d} (x_i - \hat{x}_i)^2 \quad (3)$$

During operational inference, an anomalous connection entry characterized by unfamiliar structural patterns fails to align with the compressed representations, generating a heightened reconstruction loss exceeding a rigorous threshold: $\&Lagr;(x, \hat{x}) > \tau$. This condition immediately flags the transaction as an active zero-day anomaly.

4. EXPERIMENTAL METHODOLOGY AND PIPELINE EXECUTION

The unsupervised pipeline was configured and evaluated utilizing the verified normal partition of the benchmark NSL-KDD dataset. Data preprocessing involved standard feature transforms: multi-class categorical indicators were converted via one-hot encoding, and all numerical metrics were mapped to a uniform range of [0, 1] using min-max normalization. The neural layer topology was arranged symmetrically using a structured node density sequence of 41-24-12-24-41. Model optimization spanned 35 training epochs utilizing the Adam algorithm, configured with a batch constraint size of 256. The anomaly boundary value τ was set at the 95th percentile of the baseline validation reconstruction errors, deliberately suppressing false alerts during inference.

5. EMPIRICAL FINDINGS AND COMPARATIVE ANALYSIS

Following optimization, the deep Autoencoder was evaluated against a distinct test partition containing an assortment of sophisticated attack variants mixed with legitimate traffic. Table 5.1 documents the comprehensive quantitative results achieved by the unsupervised model.

Table 5.1: Quantitative Performance Matrix for the Unsupervised Autoencoder IDS Model

Model Classification Taxonomy	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Reconstruction Limit (τ)
Deep Autoencoder (AE) (MSE Boundary)	93.2	92.5	93.0	92.7	0.0142

The empirical metrics tabulated in Table 5.1 underscore the structural fidelity of the unsupervised design, securing a prominent 93.2% accuracy ceiling along with a balanced 92.7% F1-score. These findings confirm that symmetric bottleneck networks establish tightly bound normative envelopes, isolating zero-day attack variables effectively without requiring a priori exposure to malicious labels during training phases.

6. POST-HOC EXPLAINABILITY LAYER INTEGRATION

To resolve the inherent operational opacity characterizing the deep hidden layers of the Autoencoder, a post-hoc LIME framework was linked directly to the output of the reconstruction architecture. This integration treats the computed MSE reconstruction metric ***&Lagr;(x, x′)*** as the target continuous dependent variable for explanation. Upon the triggering of an alert via an elevated loss metric, LIME constructs synthesized perturbations around the localized instance to fit an interpretable surrogate linear estimator. For a correctly identified network probe vector, this breakdown layer clearly isolated that an anomalous value in the ***same_srv_rate*** parameter contributed a shift of

+0.38 toward the overall reconstruction error, granting security operators instantaneous visibility into the structural drivers behind the system alert.

7. CONCLUSION

This investigation validates that unsupervised anomaly perimeter tracking using deep symmetric Autoencoders provides a highly capable methodology for mitigating unknown zero-day exposures. By successfully coupling this system with local model-agnostic LIME feature explanations, we demonstrate that the historical "black box" bottleneck associated with neural networks can be thoroughly dismantled. This dual-layered strategy delivers exceptional detection fidelity alongside instantly human-verifiable feature attributions, facilitating the reliable deployment of unsupervised deep models inside high-stakes operational security infrastructures.

REFERENCES

[1] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pp. 37-50, 2012.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[3] M. Ring et al., "A survey of network-based intrusion detection data sets," *Computers & Security*, vol. 86, pp. 147-167, 2019.