HSAE: A Hybrid Unsupervised Autoencoder for Zero-Day Attack Detection in Realistic Environments

Fabiano C. da Silva[1], **Franklin A. M. Venceslau**[1],
Rafael R. de Souza[2], **José A. S. Monteiro**[2]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)

[2]CESAR.School – Recife - PE

{fcs3,famv}@cin.ufpe.br, rrs3@cesar.org.br, jasm@cesar.school

Abstract. ***Research Context:*** *Anomaly detection in computer networks is essential for cybersecurity, especially in complex environments like corporate and IoT networks that face emerging threats.* ***Scientific and/or Practical Problem:*** *Traditional Intrusion Detection Systems (IDS) struggle to identify zero-day attacks, creating significant security gaps. Signature-based approaches fail against unknown threats, while supervised models require large volumes of labeled data rarely available in practice. Unsupervised methods often produce high false positive rates (FPR), limiting their operational viability.* ***Proposed Solution and/or Analysis:*** *This work proposes HSAE (Hybrid Scoring Autoencoder), a lightweight hybrid deep learning model for network anomaly detection. The solution combines autoencoder reconstruction error with an auxiliary output trained using pseudo-labels, enabling the model to learn normality patterns exclusively from benign traffic.* ***Related IS Theory:*** *The research is grounded in anomaly detection theory via machine learning, integrating unsupervised approaches. The model uses Autoencoder (AE) architecture and its principles of data compression and reconstruction to identify behavioral deviations.* ***Research Method:*** *The study conducted quantitative and empirical evaluation, training and testing the model on public datasets CICIDS2017 and ToN_IoT. HSAE performance was compared to Variational Autoencoder (VAE), the state-of-the-art, using metrics including AUC, precision, recall, F1-Score, and FPR.* ***Summary of Results:*** *HSAE consistently outperformed VAE, achieving AUC of 0.94 in corporate scenarios and 0.86 in IoT ransomware detection. Results demonstrated the model's utility as a practical tool for high-precision anomaly identification with significantly lower false positive rates.* ***Contributions and Impact to IS area:*** *This work contributes to Information Systems by presenting an anomaly detection model tailored for realistic scenarios, offering practical approaches to managing zero-day threats. Its computational efficiency enables implementation in resource-constrained environments like IoT and industrial systems where security is critical.*

1. Introdução

A crescente complexidade e heterogeneidade dos ambientes computacionais, somadas à evolução contínua das ameaças cibernéticas, impõem desafios substanciais aos Sistemas de Detecção de Intrusões (IDS), sobretudo no que se refere à identificação de ataques *zero-day*. Esses ataques exploram vulnerabilidades desconhecidas em tempo de execução

e tornam ineficazes os mecanismos tradicionais de detecção baseados em assinaturas e modelos supervisionados que dependem de representações prévias dos padrões de ataque [Guo 2023, Ahmad et al. 2023]. O principal limitador das abordagens supervisionadas consiste na exigência de conjuntos de dados rotulados com alta representatividade, tanto em volume quanto em diversidade de classes. Em cenários reais, esses conjuntos geralmente se apresentam desbalanceados, desatualizados ou indisponíveis, o que compromete diretamente a capacidade dos modelos de generalizar frente a padrões maliciosos emergentes [Zoppi et al. 2021, Mbona and Eloff 2022].

Nesse contexto, pesquisadores exploram técnicas de aprendizado não supervisionado e semi-supervisionado têm sido exploradas como alternativas promissoras, sobretudo aquelas que modelam exclusivamente o tráfego benigno e identificam desvios por meio de métricas de dissimilaridade ou erro de reconstrução, como é o caso dos *Autoencoders, Clustering-Based Methods*, e detectores de anomalia por densidade [Zavrak and Iskefiyeli 2020, Oluwadare and ElSayed 2023]. Apesar disso, esses métodos frequentemente apresentam altas taxas de falsos positivos (FPR), devido à baixa capacidade discriminativa em contextos com elevada variabilidade de tráfego legítimo [Zoppi et al. 2021, Guo 2023]. Tal limitação reduz a confiabilidade operacional dos IDS, especialmente em ambientes críticos que exigem alta vazão de pacotes e múltiplos domínios de aplicação.

Paralelamente, observa-se o aumento do uso de arquiteturas profundas e modelos de *ensemble*, como *stacked generalization*, Redes Neurais Convolucionais (CNNs), redes recorrentes (LSTM) e, mais recentemente, modelos baseados em *Transformers*. Embora essas abordagens ofereçam avanços em termos de capacidade de modelagem e detecção, suas complexidades algorítmicas e custos computacionais elevados as tornam inadequadas para sistemas com restrições de recursos computacionais, como redes IoT, dispositivos de borda e aplicações em tempo real [Ahmad et al. 2023, Verkerken et al. 2023]. Essas limitações evidenciam a necessidade de modelos de detecção de anomalias que conciliem baixa dependência por dados rotulados, robustez à variação do tráfego legítimo, capacidade de detecção de ameaças *zero-day* e eficiência computacional.

Diante desse cenário, este artigo propõe uma arquitetura híbrida e computacionalmente eficiente, fundamentada em um *autoencoder* com saída auxiliar. O modelo é treinado unicamente com tráfego benigno e integra duas estratégias de detecção: (i) o erro de reconstrução, que captura desvios em relação ao comportamento normal aprendido; e (ii) uma função auxiliar de pontuação (e.g., *anomaly score regressor*) construída a partir de dados sintéticos ou pseudo-rótulos gerados via técnicas de *self-training*. A fusão dessas métricas refina a separação entre o comportamento normal e o anômalo, o que possibilita a definição de um limiar de decisão mais preciso para reduzir o impacto dos falsos positivos, sem comprometer a capacidade de detecção de ataques inéditos.

A proposta aborda três lacunas técnicas críticas identificadas na literatura: i) dependência de dados rotulados de alta qualidade, resolvida via modelagem autosupervisionada; ii) alta taxa de falsos positivos dos modelos não supervisionados, mitigada pela incorporação de sinais auxiliares supervisionados e iii) o alto custo computacional de abordagens profundas, contornado por uma arquitetura leve, passível de execução em tempo real e adaptável a cenários com recursos limitados.

As demais seções deste artigo estão organizadas da seguinte forma: a Seção 2 aborda tópicos relevantes sobre ataques *zero-day*; a Seção 3 traz os principais trabalhos relacionados; a Seção 4 descreve a proposta de modelo híbrido do HSAE; a Seção 5 aborda as etapas da Configuração dos Experimentos; a Seção 6 trata sobre a condução dos experimentos realizados; e a Seção 7 conclui e discute os trabalhos futuros.

2. Detecção de ataques *Zero-Day*

2.1. Ataques *Zero-Day* e seus desafios

Ataques *Zero-Day* exploram vulnerabilidades desconhecidas ou ainda não documentadas, sendo indetectáveis por sistemas tradicionais baseados em assinaturas [Guo 2023]. Em média, permanecem invisíveis por cerca de 312 dias, o que pode gerar impactos financeiros e operacionais severos [Guo 2023]. O surgimento constante de novos *malwares* reforça a necessidade de métodos mais eficazes para identificação proativa dessas ameaças [Yang et al. 2022]. Esse cenário crítico é corroborado por Ahmad et al. [Ahmad et al. 2023], que destacam a carência de soluções robustas capazes de lidar com essas ameaças em tempo real. Detectar ataques *Zero-Day* impõe desafios específicos, como a escassez de dados rotulados representativos para testes, alta variabilidade dos vetores de ataque e elevado risco de falsos positivos [Abdulganiyu et al. 2023]. Modelos baseados em premissas de distribuição estacionária entre treino e teste tendem a falhar nesses cenários, o que demanda abordagens mais robustas e adaptativas [Guo 2023].

2.2. *Autoencoders* híbridos para detecção de ataques *zero-day*

Autoencoders (AE) têm sido amplamente empregados em tarefas de detecção de anomalias devido à sua capacidade de aprender representações comprimidas dos dados e identificar desvios a partir do erro de reconstrução [Berahmand et al. 2024]. Entre suas variações, o *Variational Autoencoder* (VAE) destaca-se por sua natureza generativa, permitindo modelar distribuições probabilísticas dos dados e capturar padrões sutis de comportamento anômalo. De acordo com Berahmand et al. [Berahmand et al. 2024], o VAE integra a classe dos *autoencoders* mais promissores para aplicações não supervisionadas, sendo amplamente utilizado em contextos como segurança da informação e análise de tráfego de rede. No entanto, estudos recentes apontam que o desempenho do VAE pode ser impactado negativamente em ambientes realistas com tráfego altamente variado, dada a sensibilidade do modelo a diferentes tipos de ataque e padrões de dados [Guo 2023]. Ainda assim, segundo levantamento conduzido por Oluwadare e ElSayed [Oluwadare and ElSayed 2023], os *autoencoders* continuam figurando entre os algoritmos mais eficazes na detecção de ataques *zero-day*, especialmente pela capacidade de modelar padrões complexos sem a necessidade de rótulos.

2.2.1. Equal Error Rate (EER)

O Ponto de Igualdade de Erros (Equal Error Rate - EER) é uma métrica fundamental para a calibração de sistemas de classificação. No contexto deste trabalho, o EER é o critério utilizado para definir o limiar de decisão (τ^*) a ser aplicado sobre o *AnomalyScore* final gerado pelo HSAE. Ele estabelece o ponto operacional onde a Taxa de Falsa Aceitação (FAR) se iguala à Taxa de Falsa Rejeição (FRR). Matematicamente, essa condição

é expressa como FAR(τ^*) = FRR(τ^*), onde τ^* representa o limiar ótimo que satisfaz essa igualdade [Cheng and Wang 2004]. A principal vantagem do EER é fornecer um critério objetivo e padronizado para a calibração do limiar, eliminando a necessidade de ajustes empíricos e consolidando-se como um padrão para benchmarking de algoritmos. Valores menores de EER indicam um melhor desempenho discriminativo e, como métrica de avaliação, o próprio EER se correlaciona inversamente com a Área sob a Curva ROC (AUC). A Curva ROC (*Receiver Operating Characteristic*) ilustra a capacidade de um modelo em distinguir entre classes, e a AUC (*Area Under the Curve*) quantifica essa performance em um valor único, onde resultados mais altos indicam um melhor poder de discriminação [Rainio et al. 2024]. Contudo, suas limitações incluem a premissa de custos simétricos entre os tipos de erro e a sensibilidade a distribuições de dados que não sejam representativas das condições operacionais reais [Cheng and Wang 2004].

3. Trabalhos Relacionados

Nesta seção são revisadas técnicas de aprendizado de máquina (ML) e aprendizado profundo (DL) aplicadas à detecção de ataques, com ênfase especial na identificação de tráfego anômalo e ataques *zero-day* em redes IoT e sistemas críticos. Analisam-se o desempenho, limitações e aplicabilidade em cenários com tráfego dinâmico e realista.

O estudo conduzido por Zavrak e Iskefiyeli [Zavrak and Iskefiyeli 2020] propõe uma abordagem baseada em aprendizado profundo não supervisionado para detecção de tráfego anômalo a partir de dados de fluxo. A investigação compara três métodos — *Autoencoder* (AE), *Variational Autoencoder* (VAE) e *One-Class SVM* (OCSVM) — todos treinados exclusivamente com fluxos benignos e avaliados com base nas curvas ROC e na métrica AUC. Os resultados experimentais indicam que o VAE apresenta desempenho superior na maioria dos cenários analisados, especialmente na detecção de ataques com alta taxa de ocorrência, como DoS e DDoS. O estudo tem sido reconhecido na literatura como uma das referências relevantes no uso de *autoencoders* para a detecção de intrusões em redes, sendo citado em revisões sistemáticas recentes que discutem o papel de modelos generativos na segurança cibernética [Halvorsen et al. 2024] e que apresentam taxonomias atualizadas sobre sistemas de detecção de intrusões [Alkasassbeh and Al-Haj Baddar 2023]. Contudo, os autores não exploram mecanismos adaptativos de detecção nem estratégias de controle dinâmico de falsos positivos, o que pode comprometer a robustez da proposta em ambientes reais e dinâmicos.

Mbona e Eloff [Mbona and Eloff 2022] propõem uma estratégia para detectar ataques *zero-day* combinando a Lei de Benford com aprendizado semi-supervisionado. Eles realizam a seleção de atributos com base na distribuição de anomalias numéricas, e o modelo resultante utiliza OCSVM. Os resultados experimentais mostram um F1-score de 85% e um MCC (*Matthews Correlation Coefficient*) de 74%. Apesar desses índices, a proposta carece de comparações com outros classificadores, depende de parâmetros estáticos e não considera variações no tráfego nem otimizações de desempenho computacional — limitações que prejudicam sua aplicação em cenários com restrições de tempo e recursos.

No trabalho de Alrayes et al. [Alrayes et al. 2024], um sistema baseado em *Denoising Autoencoder* (DAE) é utilizado para identificar padrões anômalos por meio da reconstrução de fluxos corrompidos. A proposta é avaliada com os conjuntos CICIDS2017 e NSL-KDD, alcançando precisão de 99,991% e F1-score de 0,998 no primeiro dataset.

Embora obtenha ótimo desempenho, o modelo não incorpora ajuste adaptativo de sensibilidade nem avalia o impacto computacional em tempo real. Além disso, não considera ruído nem perda de pacotes, o que limita sua aplicabilidade em ambientes adversos.

Uma proposta mais estruturada é apresentada por Verkerken et al. [Verkerken et al. 2023], que introduzem uma arquitetura hierárquica multiestágio para a detecção de intrusões. Na primeira fase, aplicam filtragem leve com AE ou OCSVM; em seguida, classificam ataques conhecidos usando modelos supervisionados (por exemplo, Random Forest ou redes neurais profundas); por fim, detectam ataques zero-day por meio de análise de desvios com limiares ajustáveis. A abordagem alcança precisão balanceada de até 96% e *recall* de 78% para ataques *zero-day*, além de reduzir o consumo de largura de banda em até 69%. Contudo, exige configuração manual dos limiares e não relata experimentos sobre latência e custo computacional, o que dificulta avaliar sua viabilidade em tempo real.

Por fim, Lu et al. [Lu et al. 2024] exploram o uso de aprendizado por transferência em sistemas CBTC (*Communication-Based Train Control*), com uma arquitetura híbrida que combina 1D-CNNs otimizadas e camadas LSTM. Eles automatizam a extração de características espaciais e temporais, ajustam hiperparâmetros por simulated annealing e transferem conhecimento entre domínios. Os resultados mostram F1-score de 93,21% para ataques *zero-day* e 99,6% para ataques conhecidos. No entanto, o modelo exige fine-tuning com amostras de ataques novos, implicando dependência de conhecimento prévio. A complexidade arquitetural pode comprometer o tempo de resposta e aumentar o custo computacional, reduzindo a viabilidade em ambientes com recursos restritos.

Diferenciando-se das abordagens anteriores, este trabalho propõe uma arquitetura híbrida não supervisionada baseada em *Autoencoder*, voltada à detecção de anomalias e ataques *Zero-Day*. Integramos a aprendizagem não supervisionada a uma saída auxiliar, permitindo capturar o comportamento normal via reconstrução e refinar a classificação de anomalias através da aplicação de um limiar de decisão mais eficaz sobre um *score* combinado. A solução integra aprendizado não supervisionado com uma saída auxiliar, permitindo capturar o comportamento normal via reconstrução e ajustar a fronteira de decisão com suporte adicional. Técnicas como *Dropout* e *Batch Normalization* são incorporadas para aumentar a robustez frente a variações no tráfego, reduzindo especialmente a incidência de falsos positivos, mantendo baixo custo computacional e alta aplicabilidade em ambientes dinâmicos e críticos como redes IoT e infraestruturas industriais [Abdulganiyu et al. 2023, Hajj et al. 2021]. Os limiares de decisão estáticos foram substituídos por um mecanismo de otimização dinâmica baseado no Equal Error Rate. Este ajuste é aplicado a um score combinado que une ambas as saídas do modelo, eliminando a necessidade de configurações manuais e adaptando-se automaticamente às características dos dados. Por fim, a avaliação do HSAE é realizada com um conjunto abrangente de métricas, incluindo precisão, FPR, recall e F1-score, superando a limitação de trabalhos anteriores que se restringem à análise da AUC [Zavrak and Iskefiyeli 2020].

4. O modelo Híbrido proposto

4.1. Estrutura do HSAE

A proposta central deste estudo é construir um modelo híbrido não supervisionado denominado HSAE (*Hybrid Scoring Autoencoder*). Esse modelo consiste em um *autoencoder*

profundo com uma ramificação auxiliar responsável pela geração de uma pontuação de anomalia probabilística. A Figura 1 ilustra a arquitetura completa do modelo HSAE proposto, destacando o fluxo de dados desde a entrada até as duas saídas: a reconstrução e a pontuação de anomalia.

Figura 1. Arquitetura do HSAE com dupla saída

Formalmente, seja $x \in \mathbb{R}^d$ uma amostra de entrada. O encoder mapeia x em um espaço latente $z \in \mathbb{R}^k$, onde $k < d$, por meio da função $h_\psi : \mathbb{R}^d \to \mathbb{R}^k$. O decoder reconstrói $\hat{x} = g_\phi(z)$, onde $g_\phi : \mathbb{R}^k \to \mathbb{R}^d$. A saída auxiliar é uma função $\hat{y} = \sigma(Wz+b)$, com σ sendo a função sigmoide.

O modelo completo é, portanto, uma composição de funções parametrizadas:

$$f_\theta(x) = (g_\phi(h_\psi(x)), \sigma(W h_\psi(x) + b))$$

onde $\theta = \{\psi, \phi, W, b\}$.

A arquitetura do HSAE é composta por:

- *Encoder*: Dense(1024) $\to$ Dense(512) $\to$ Dense(256), com *LeakyReLU*, *BatchNormalization* e *Dropout*;
- *Decoder*: Dense(512) $\to$ Dense(1024) $\to$ Output;
- Saída Auxiliar: Dense(1, `sigmoid`).

Ou seja, temos um *encoder* composto por três camadas densas (1024, 512 e 256 unidades) com funções de ativação *LeakyReLU*, normalização em lote e regularização por *Dropout*, projetado para codificar representações compactas. O *decoder* reconstrói os dados a partir dessas representações utilizando camadas densas de 512 e 1024 unidades até a saída.

A concepção da arquitetura HSAE foi orientada pela busca de um método de detecção de anomalias que pudesse se beneficiar de uma análise multi-critério, especialmente em cenários com ataques zero-day. A seguir, detalha-se a fundamentação para a

adoção de uma saída dupla — o erro de reconstrução e o *anomaly score* — e como a combinação de ambos busca oferecer uma detecção mais robusta.

O *anomaly score* foi proposto como uma saída auxiliar que, na prática, atua como um classificador binário sobre o espaço latente — a representação comprimida e significativa dos dados gerada pelo *encoder*. A literatura aponta que essa representação latente pode ser utilizada como um extrator de características para outras tarefas, como a própria classificação [Bank et al. 2023]. Inspirado por essa capacidade, o *anomaly score* busca avaliar e classificar se a própria representação latente de uma amostra é consistente com os padrões de normalidade aprendidos.

A sua implementação utiliza uma função de ativação sigmoide, uma escolha fundamentada em suas propriedades matemáticas. Uma função de ativação sigmoide é não linear e diferenciável, requisitos para o funcionamento de redes MLP (*Multilayer Perceptron*) treinadas com retropropagação (*backpropagation*) [Narayan 1997]. Seus principais benefícios no contexto desta arquitetura são:

- **Interpretabilidade e Classificação:** A função sigmoide mapeia a saída para o intervalo [0, 1], o que permite que o resultado seja interpretado como uma pontuação de anomalia, análoga a uma probabilidade [Pratiwi et al. 2020]. Essa pontuação é a base para a classificação: valores próximos de 0 são associados à classe *"normal"*, enquanto valores próximos de 1 são associados à classe *"anômala"*.
- **Treinamento Direcionado:** No treinamento com dados exclusivamente benignos, a função de perda híbrida, por meio do componente de entropia cruzada binária (Binary Cross-Entropy – BCE) — uma perda clássica de classificação —, incentiva essa saída a se aproximar de zero [Pratiwi et al. 2020]. Com isso, o modelo é treinado não apenas para reconstruir dados normais, mas também para classificar a representação latente de tráfego benigno com um *score* mínimo.

Dessa forma, a proposta é que o *anomaly score* introduza um critério de detecção classificatório e complementar, focado na consistência da representação latente dos dados.

A eficácia da arquitetura híbrida é demonstrada na fase de teste, momento em que o modelo, já treinado com dados normais, confronta dados brutos que nunca viu, incluindo tanto tráfego normal quanto ataques. Ao receber uma nova amostra, o modelo a submete a duas avaliações simultâneas para decidir se é uma anomalia:

1. **Verificação Estrutural (via Reconstrução):** O modelo tenta reconstruir a amostra de entrada. O *erro de reconstrução* funciona como um sensor para **desvios estruturais**. Se uma amostra de ataque possui padrões, fluxos ou características que diferem da estrutura normal aprendida, o modelo falhará em recriá-la fielmente, gerando um erro alto.
2. **Verificação Representacional (via *Anomaly Score*):** O *encoder* mapeia a amostra para o espaço latente, e a saída sigmoide a classifica. O *anomaly score* atua como um sensor para desvios de representação. Mesmo que um ataque seja estruturalmente similar ao tráfego normal, sua representação interna no modelo pode ser atípica. A saída sigmoide, treinada para reconhecer apenas representações normais, irá sinalizar essa inconsistência com um *score* alto (próximo de 1).

A necessidade dessa dupla verificação no teste é justificada pela forma como o modelo é treinado. Se o treinamento dependesse apenas do *anomaly score*, o *encoder*

poderia ter aprendido um atalho: colapsar a informação, mapeando todos os tipos de tráfego normal para uma representação única e simplista. Isso destruiria a capacidade do modelo de generalizar, pois ele não teria a sensibilidade para notar, no teste, as nuances que distinguem um ataque sutil de um tráfego normal. A tarefa de reconstrução, portanto, atua como um regularizador estrutural durante o treino, forçando o encoder a criar um mapa rico e detalhado da normalidade, o que torna a verificação representacional na fase de teste muito mais confiável e precisa [Bank et al. 2023].

A fusão dos dois *scores* em um resultado combinado busca implementar uma estratégia de combinação de informações, visando criar um indicador de anomalia mais completo. Esta abordagem encontra respaldo na literatura, onde *autoencoders* são utilizados como uma forma de regularização para redes de classificação, combinando a perda de reconstrução com a perda de classificação em uma função de custo unificada [Bank et al. 2023]. Abordagens similares são vistas em *autoencoders* semi-supervisionados, que também integram diferentes tipos de perdas para alavancar tanto dados rotulados quanto não rotulados [Berahmand et al. 2024].

A adoção de uma ponderação igualitária (50/50) representa uma abordagem inicial equilibrada, que busca assegurar que ambos os mecanismos de detecção — estrutural e representacional — contribuam de maneira balanceada para a decisão final. Essa escolha procura evitar um viés prévio em favor de um tipo específico de anomalia, contribuindo para a capacidade de generalização do modelo.

4.2. Modelagem do mecanismo de detecção

O objetivo do modelo é minimizar uma função de custo híbrida, que incorpora tanto a reconstrução da entrada quanto a precisão da saída auxiliar. A função de perda total é definida como:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda \cdot \mathcal{L}_{cls}$$

onde $\lambda = 0{,}03$ é um hiperparâmetro de regularização. O fator de ponderação λ na função de perda total foi definido em 0,03 para priorizar a tarefa primária de reconstrução dos dados normais, permitindo que a perda de classificação auxiliar atuasse como um regularizador para aprimorar a discriminação do espaço latente sem sobrepujar o objetivo principal [Bank et al. 2023]. A perda de reconstrução quantifica o erro quadrático médio entre a entrada original e a sua respectiva saída, sendo dada por:

$$\mathcal{L}_{rec}(x, \hat{x}) = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2$$

A perda de classificação não supervisionada é modelada por entropia cruzada binária:

$$\mathcal{L}_{cls}(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Este acoplamento de perdas permite que o HSAE aprenda a reconstruir apenas padrões benignos, penalizando desvios pela saída auxiliar. Quanto menor esse valor, melhor a rede aprende a reconstruir padrões benignos, permitindo que desvios significativos na reconstrução indiquem possíveis anomalias.

5. Configuração dos Experimentos

A seguir, descrevem-se os procedimentos adotados para a configuração dos experimentos, abrangendo a preparação dos dados, definição das métricas, treinamento do modelo e metodologia de validação experimental da abordagem proposta, cuja estrutura e funcionamento foram apresentados na Seção 4.

5.1. Objetivo da validação

A validação do modelo HSAE busca verificar sua capacidade de generalizar a detecção de ataques *Zero-Day* com base apenas em tráfego benigno. Avalia-se sua eficácia em identificar padrões anômalos com alta precisão e baixa taxa de falsos positivos, mesmo em contextos operacionais variados. Para isso, realizam-se experimentos com conjuntos de dados heterogêneos, simulando cenários práticos e testando o desempenho do mecanismo de detecção híbrido baseado em erro de reconstrução e um *score* de camada auxiliar.

5.2. Métricas de avaliação

As métricas utilizadas para avaliar o desempenho do modelo incluem:

- Precisão: $\frac{TP}{TP+FP}$ — mede a proporção de acertos entre as predições positivas feitas pelo modelo.
- *Recall*: $\frac{TP}{TP+FN}$ — indica a capacidade do modelo em recuperar corretamente os casos positivos.
- F1-Score: $2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}}$ — representa a média harmônica entre precisão e *recall*, útil em cenários com classes desbalanceadas.
- AUC-ROC: área sob a curva ROC — avalia a habilidade do modelo em distinguir entre classes positivas e negativas.
- FPR: $\frac{FP}{FP+TN}$ — taxa de falsos positivos, ou seja, a proporção de negativos classificados erroneamente como positivos.

5.3. Metodologia

A Figura 2 apresenta a metodologia adotada nesta pesquisa para a avaliação do modelo híbrido proposto. Ela descreve desde a divisão dos dados utilizados, passando pelo pré-processamento necessário, treinamento do modelo híbrido proposto (HSAE) e a etapa final de avaliação do desempenho por meio de diversas métricas.

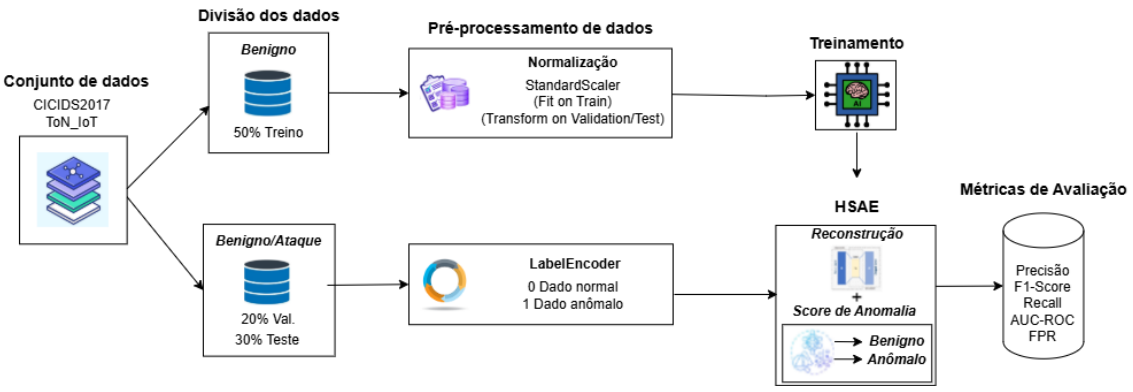


Figura 2. Visão geral do experimento para o modelo proposto.

Nosso objetivo é simular um cenário realista de detecção de anomalias, onde treinamos o modelo exclusivamente com dados benignos e, posteriormente, o expomos a tráfego misto, contendo instâncias normais e maliciosas. Para isso, dividimos os dados da seguinte forma: 50% do tráfego benigno foi destinado ao treinamento, 20% foram utilizados na validação e os 30% restantes compuseram o conjunto de teste. O limiar de decisão para o *Score* Combinado do modelo é definido via EER sobre o conjunto de validação e, então, aplicado ao teste.

Para os dados de treinamento, utilizamos a técnica de padronização `z-score` para o pré-processamento dos atributos, por meio do `StandardScaler`. Os parâmetros de média e desvio padrão foram calculados exclusivamente sobre o conjunto de treinamento, uma prática que evita o vazamento de informações (*data leakage*) entre as etapas. Esse conjunto de dados de treino, já normalizado e composto unicamente por tráfego benigno, foi então utilizado para treinar o modelo HSAE, permitindo que ele aprendesse a representar o padrão de normalidade da rede.

Para a etapa de avaliação, os conjuntos de validação e teste, que contêm tráfego misto (benigno e malicioso), foram transformados utilizando os mesmos parâmetros do `StandardScaler` ajustado no treino. Em seguida, para viabilizar o cálculo das métricas de desempenho, os rótulos desses conjuntos foram convertidos em formato binário (0 para benigno, 1 para anomalia) com o `LabelEncoder`. Esse procedimento permite uma avaliação quantitativa da capacidade do modelo em generalizar e identificar anomalias em dados não vistos.

5.4. Separação dos dados

Empregamos um paradigma não supervisionado, utilizando apenas o tráfego benigno durante o treinamento. Conduzimos o processo de rotulagem por meio do algoritmo `LabelEncoder`, que atribui os seguintes valores:

$$\text{BENIGNO} \rightarrow 0, \quad \text{ANÔMALO} \rightarrow 1$$

Seja $\mathcal{X} \subset \mathbb{R}^n$ o espaço de características extraídas a partir do tráfego de rede, e $\mathcal{Y} = \{0, 1\}$ o espaço de rótulos binários, em que $y_i = 0$ representa tráfego benigno e $y_i = 1$ representa tráfego malicioso. A base de dados original é composta por pares $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, nos quais x_i representa um vetor de atributos extraídos de um fluxo de rede, e y_i é o rótulo correspondente.

A divisão dos dados foi realizada por meio da função `train_test_split`, com a utilização do parâmetro random_state=42 em todas as chamadas da função, de forma a garantir a preservação da proporção das classes. O conjunto de treinamento $\mathcal{D}_{\text{train}}$ é definido como:

$$\mathcal{D}_{\text{train}} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid y_i = 0\}$$

ou seja, é composto exclusivamente por amostras benignas, permitindo ao modelo, neste caso, o *autoencoder* aprender a estrutura normal do tráfego de rede em ausência de ruído malicioso.

O conjunto de teste $\mathcal{D}_{\text{test}}$ é constituído por uma amostragem balanceada contendo exemplos de ambas as classes, permitindo avaliar a capacidade discriminativa do modelo:

$$\mathcal{D}_{\text{test}} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid y_i \in \{0, 1\} \text{ e } N_0 = N_1\}$$

onde N_0 e N_1 representam a cardinalidade dos subconjuntos de instâncias benignas e maliciosas, respectivamente, garantindo o balanceamento entre as classes no conjunto de teste.

Adicionalmente, utilizamos 20% dos dados contendo tráfego benigno e anômalo para compor o conjunto de validação durante o treinamento. Essa divisão permite aferir a capacidade de generalização do modelo em relação a dados benignos não observados no processo de otimização dos parâmetros, evitando, assim, o sobreajuste (*overfitting*) no conjunto de treinamento.

5.5. Pré-processamento dos Dados

A padronização dos atributos foi realizada utilizando a transformação `z-score`, dada por:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma}$$

onde μ e σ são a média e o desvio padrão, respectivamente, calculados sobre o conjunto de treinamento $\mathcal{D}_{train}$. Essa transformação assegura que os dados possuam média zero e variância unitária, propriedade essencial para estabilizar o processo de otimização em redes neurais profundas.

Aqui aplicamos a padronização aos dados, utilizando o `StandardScaler` da biblioteca *Scikit-learn*. Primeiro, ele calcula a média e o desvio padrão de cada atributo no conjunto de treinamento e transforma os dados de treino para que tenham média zero e desvio padrão um. Em seguida, aplica a mesma transformação ao conjunto de teste, usando os parâmetros do treino, o que evita vazamento de dados e garante consistência na escala das variáveis.

5.6. Etapa de Treinamento

O modelo foi treinado por 150 épocas com batch size 128, utilizando o algoritmo de otimização Adam com taxa de aprendizado igual a 5×10^{-5}. A função de custo híbrida garante que o modelo seja sensível a desvios sutis na estrutura do tráfego. Durante a inferência, o modelo retorna a reconstrução $\hat{x}$ e a saída auxiliar $\hat{y}_{anomaly}$.

O erro de reconstrução é computado como:

$$\varepsilon(x) = \frac{1}{d} \sum_{i=1}^{d} |x_i - \hat{x}_i|$$

Para definir um limiar robusto de detecção τ o `AnomalyScore` é calculado para todas as amostras do conjunto de validação. O limiar ótimo de decisão $\hat{\tau}^*$ é então definido pelo Ponto de Igualdade de Erros (Equal Error Rate - EER), que corresponde ao ponto onde a taxa de falsos positivos (FPR) se iguala à taxa de falsos negativos (FNR), otimizando o equilíbrio entre os dois tipos de erro.

A pontuação final de anomalia é obtida pela média ponderada entre $\hat{y}_{\text{anomaly}}$ e $\varepsilon(x)$:

$$\text{AnomalyScore} = \frac{1}{2}(\hat{y}_{\text{anomaly}} + \varepsilon(x))$$

A decisão binária é então obtida comparando o `AnomalyScore` com o limiar $\hat{\tau}^*$ derivado do EER.

6. Experimentos Realizados

Para avaliar a eficácia do modelo HSAE na detecção de ataques *Zero-Day* em diferentes contextos operacionais, foram realizados experimentos com cenários realistas compostos por tráfego misto. A proposta foi testada quanto à robustez frente à variabilidade do tráfego legítimo e à capacidade de generalizar a detecção de padrões maliciosos inéditos. O código do modelo, contendo as principais funções, está disponível na plataforma *Github*[1].

6.1. Conjuntos de dados escolhidos

Para validar o desempenho do modelo proposto, utilizamos dois conjuntos de dados amplamente reconhecidos na literatura, cada um representando contextos operacionais distintos e apresentando características diversas do tráfego de rede. Utilizamos o primeiro conjunto de dados, o CICIDS2017 [Sharafaldin et al. 2018], desenvolvido pelo *Canadian Institute for Cybersecurity*, que simula um ambiente corporativo real, englobando tráfego legítimo e malicioso. Esse conjunto de dados é amplamente utilizado em pesquisas sobre sistemas de detecção de intrusões devido à diversidade de ataques modernos que apresenta, com destaque para os ataques de negação de serviço (DoS/DDoS). Escolhemos esses ataques por sua alta frequência em redes reais e pelo impacto significativo que exercem na disponibilidade dos serviços. De acordo com De Neira et al. [De Neira et al. 2023], os ataques DDoS figuram entre as principais ameaças cibernéticas globais, exigindo soluções que antecipem e mitiguem esses eventos volumétricos.

Adotamos o segundo conjunto de dados, o ToN_IoT [Moustafa 2021], desenvolvido pelo *Australian Centre for Cyber Security*, que representa ambientes modernos e complexos como casas inteligentes e redes industriais. Esse conjunto de dados inclui múltiplas fontes de dados — tráfego de rede, registros de sistemas e sensores — permitindo uma visão abrangente do comportamento da rede em contextos de Internet das Coisas (IoT) e IoT Industrial (IIoT). Embora o ToN_IoT contenha uma grande variedade de vetores de ataque (*Backdoor*, *Injection*, XSS, *Password Cracking*, *Man-in-the-Middle*, entre outros), foram selecionados para este trabalho apenas os ataques DoS, DDoS e *Ransomware*. Incluímos o ataque de *Ransomware* devido ao seu crescimento alarmante, sendo atualmente uma das ameaças mais sofisticadas e destrutivas, conforme destacado na literatura recente [Razaulla et al. 2023, Beaman et al. 2021]. Essa escolha metodológica alinha os experimentos com ameaças reais e críticas em redes operacionais modernas, conforme discutido no contexto do conjunto de dados CICIDS2017.

Utilizar esses dois conjuntos de dados possibilita avaliar a robustez da abordagem proposta em diversos contextos operacionais e tipos de ameaças, desde ambientes corporativos tradicionais (CICIDS2017) até infraestruturas modernas e heterogêneas (ToN_IoT). Essa diversidade é essencial para validar a generalização dos modelos de detecção de intrusões, conforme recomendado na literatura recente [Elouardi et al. 2024].

6.2. Resultados para o conjunto de dados CICIDS2017

A avaliação inicial do desempenho do modelo proposto foi baseada na métrica AUC (Área sob a Curva ROC), seguindo a abordagem do estudo de Zavrak e Iskefiyeli [Zavrak and Iskefiyeli 2020], que adotou exclusivamente essa métrica. Entre os modelos avaliados por aqueles autores, o *Variational Autoencoder* (VAE) apresentou os melhores resultados,

[1]https://github.com/fabianoinfosec/HSAE.git

sendo, por isso, adotado como modelo de referência nesta comparação. Para assegurar uma comparação justa e compatível com os experimentos realizados pelos autores, também foi adotado o mesmo conjunto de dados: o CICIDS2017.

Como apresentado anteriormente, o VAE continua sendo valorizado na literatura atual como uma solução eficaz para detecção de anomalias, especialmente em contextos relacionados à segurança da informação. Conforme Berahmand et al. [Berahmand et al. 2024], *autoencoders* como o VAE vêm sendo empregados com frequência em tarefas que envolvem a identificação de comportamentos irregulares em dados complexos. Sua capacidade de modelar distribuições probabilísticas permite detectar desvios sutis em relação ao padrão normal, característica fundamental para sistemas de detecção de intrusões (IDS). Nesse sentido, o VAE é classificado como um *autoencoder* generativo e figura entre as abordagens mais promissoras para cenários envolvendo tráfego anômalo, reforçando sua relevância e atualidade no campo. No entanto, nesse artigo, a descrição do processo de preparação dos dados não é suficientemente detalhada. Etapas essenciais, como os critérios adotados para normalização, estratégias de validação e o tratamento do desbalanceamento entre as classes, não são claramente especificadas, o que compromete a reprodutibilidade dos experimentos e dificulta comparações precisas com outras abordagens.

Além da métrica AUC adotada por Zavrak e Iskefiyeli [Zavrak and Iskefiyeli 2020], ampliamos a análise comparativa entre o VAE e o modelo HSAE, utilizando métricas adicionais relevantes, tais como precisão, *recall*, F1-score e taxa de falsos positivos (FPR). Como não dispúnhamos do código original, reimplementamos o VAE ajustado ao nosso cenário experimental, de modo a assegurar consistência metodológica em termos de pré-processamento, divisão dos dados e métricas de avaliação. Essa reimplementação preserva a estrutura central proposta por Zavrak e Iskefiyeli [Zavrak and Iskefiyeli 2020], com encoder, decoder e regularização do espaço latente via penalização KL. No entanto, foram adotadas adaptações práticas, como o uso do erro de reconstrução como *score* de detecção (em substituição à probabilidade de reconstrução) [Berahmand et al. 2024] [Yang et al. 2022] e treinamento com o otimizador Adam [Berahmand et al. 2024].

Tais escolhas são compatíveis com diretrizes amplamente aceitas na literatura para implementações práticas de *autoencoders* variacionais, que destacam a flexibilidade desses modelos quanto à função de ativação, tipo de perda e estratégias de limiar baseadas em percentis do erro de reconstrução [Yang et al. 2022]. Essas modificações foram necessárias para permitir uma comparação justa com o HSAE, sem comprometer os princípios fundamentais do modelo variacional. A arquitetura implementada para o VAE inclui um *encoder* com camadas *Dense* (512 e 256 neurônios, LeakyReLU), normalização por lotes e *dropout*, espaço latente com dimensão 64 e um decoder simétrico. A função de perda combina MSE (*Mean Squared Error*) com penalização KL-*divergence*, e o treinamento foi conduzido por 150 épocas. A detecção de anomalias baseou-se no erro de reconstrução, com o limiar definido pelo percentil 75 dos erros do tráfego benigno [Yang et al. 2022].

A Tabela 1 apresenta os resultados obtidos para cada tipo de ataque de negação de serviço (DoS e DDoS), considerando o conjunto de dados CICIDS2017, com os melhores valores de cada métrica destacados em negrito para cada um dos dois modelos comparados. Essa abordagem garante uma comparação justa, reprodutível e metodologicamente

alinhada com os objetivos da proposta.

Tabela 1. Comparação de desempenho para o conjunto de dados *CIC-IDS2017*.

Ataque	Modelo	Precisão	FPR	Recall	F1-Score	AUC
DDoS	VAE	65%	34,76%	**66%**	65%	0.75
	HSAE	**91%**	**6,36%**	64%	**75%**	**0.82**
DoS GoldenEye	VAE	87%	13,02%	**86%**	**86%**	0.92
	HSAE	**90%**	**8,82%**	77%	83%	0.92
DoS Hulk	VAE	80%	20,10%	**80%**	80%	0.91
	HSAE	**90%**	**8,30%**	73%	**81%**	**0.94**
DoS SlowHTTPTest	VAE	91%	8,95%	**88%**	**90%**	0.94
	HSAE	91%	**8,73%**	86%	88%	0.94
DoS Slowloris	VAE	76%	**24,21%**	78%	77%	0.87
	HSAE	**77%**	27,40%	**90%**	**83%**	**0.89**

Os resultados revelam que o modelo HSAE apresenta desempenho superior em diversas métricas relevantes. Por exemplo, no caso do ataque DDoS, o HSAE alcançou 91% de precisão, superando os 65% do VAE, com uma melhoria significativa no F1-score (75% contra 65%). Apesar do *recall* ter tido uma leve inferioridade (64% contra 66%), o HSAE apresentou uma boa redução na FPR (6,36% contra 34,76%), demonstrando excelente controle de falsos positivos. A métrica AUC do HSAE apresentou valor de 0,82, superando o VAE que obteve 0,75, refletindo uma maior capacidade do HSAE em classificar corretamente o tráfego malicioso e benigno.

No ataque DoS GoldenEye, ambos os modelos apresentaram desempenho equilibrado, com o HSAE obtendo 90% de precisão contra 87% do VAE. O F1-score do HSAE foi ligeiramente inferior (83% vs 86%), assim como o *recall* (77% vs 86%). A FPR do HSAE foi de 8,82%, comparada aos 13,02% do VAE, evidenciando melhor controle de falsos positivos. A métrica AUC manteve-se equivalente em 0,92 para ambos os modelos, demonstrando capacidade similar de discriminação entre as classes neste tipo específico de ataque.

Em relação ao ataque DoS Hulk, o HSAE apresentou ganhos relevantes, com F1-*score* de 81% superando os 80% do VAE, precisão de 90% contra 80%. O *recall* apresentou leve redução (73% vs 80%), porém a FPR do HSAE foi bastante inferior (8,30% vs 20,10%), demonstrando melhor controle de falsos positivos. A métrica AUC do HSAE foi superior (0,94 contra 0,91 do VAE), indicando melhor desempenho geral na discriminação entre tráfego benigno e malicioso para este tipo de ataque.

Nos ataques DoS Slowhttptest, o HSAE apresentou resultados superiores com 91% de precisão, mesmo valor do VAE, mas com F1-score inferior (88% vs 90%). O recall do HSAE foi de 86%, ligeiramente inferior aos 88% do VAE, enquanto a FPR ficou em 8,73% contra 8,95% do VAE. A métrica AUC manteve-se equivalente em 0,94 para ambos os modelos, demonstrando capacidade similar de classificação. No caso do DoS Slowloris, os resultados apresentaram maior variabilidade. O HSAE obteve 77% de precisão, superior aos 76% do VAE, com F1-score superior (83% vs 77%) e um recall melhor (90% vs 78%). A FPR do HSAE foi de 27,40%, superior aos 24,21% do VAE,

indicando maior taxa de falsos positivos neste cenário específico. A métrica AUC do HSAE foi de 0,89, superior aos 0,87 do VAE.

As curvas ROC apresentadas na Figura 3 ilustram a separação entre tráfego benigno e malicioso nos diferentes ataques, evidenciando visualmente o desempenho dos modelos analisados. Essa distinção permite uma comparação direta entre as abordagens, destacando o desempenho geral superior do HSAE em termos de capacidade de discriminação.

Esses resultados indicam que o HSAE é um modelo mais conservador e robusto, priorizando a redução de falsos positivos e a manutenção de alta precisão — características desejáveis em redes sensíveis como ambientes industriais ou infraestruturas críticas, onde alarmes falsos podem comprometer a operação. Embora o *recall* apresente, em alguns casos, valores similares aos do VAE, os ganhos em precisão, F1-score, FPR e, especialmente, na maioria dos casos analisados, pela métrica AUC, evidenciam a eficácia do HSAE na detecção de anomalias. Portanto, com o *dataset CICIDS2017*, o modelo HSAE demonstrou-se mais eficiente em termos gerais, validando sua adoção frente a abordagens tradicionais baseadas exclusivamente em reconstrução. Os resultados obtidos reforçam o potencial do HSAE para aplicações reais, onde o custo de falsos positivos seja elevado.

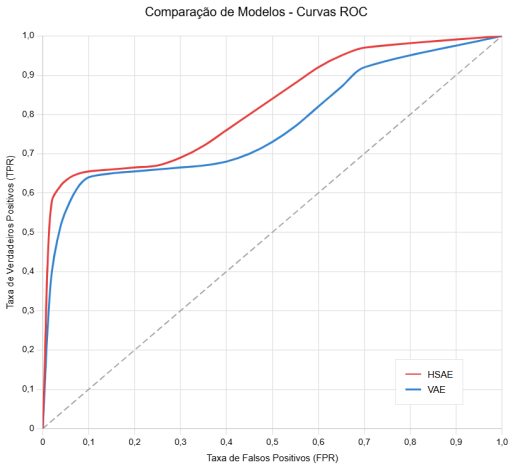
6.3. Resultados para o conjunto de dados ToN_IoT

Foram realizados testes utilizando o *ToN_IoT*, um conjunto de dados que reflete padrões de tráfego oriundos de dispositivos IoT em ambientes industriais e residenciais, abrangendo uma variedade de vetores de ataque e condições realistas de operação. A Tabela 3 apresenta os resultados obtidos para três classes de ataques distintas, comparando o desempenho do modelo proposto HSAE com o modelo de referência VAE, por meio de métricas consagradas em sistemas de detecção de intrusão, tais como: Precisão, *False Positive Rate* (FPR), F1-*Score*, *Recall* e Área sob a Curva ROC (AUC).

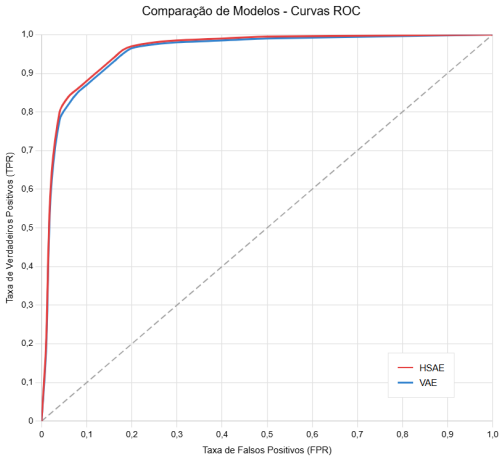
Tabela 2. Comparação de desempenho para o conjunto de dados *ToN_IoT*.

Ataque	Modelo	Precisão	FPR	*Recall*	*F1-Score*	AUC
DDoS	VAE	70%	25,69%	*75%*	*72%*	0.79
	HSAE	**85%**	**14,77%**	***84%***	***85%***	**0.91**
DoS	VAE	69%	26,34%	*73%*	*71%*	0.76
	HSAE	**91%**	**22,11%**	***78%***	***84%***	**0.77**
Ransomware	VAE	67%	28,09%	*71%*	*70%*	0.76
	HSAE	**86%**	**13,93%**	***85%***	***85%***	**0.86**

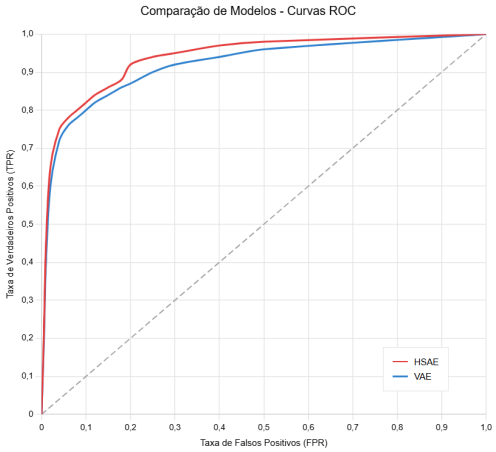
No contexto do ataque DDoS, o modelo HSAE demonstrou ganhos consideráveis em relação ao VAE. A precisão aumentou de 70% para 85%, evidenciando que o processo de codificação-decodificação variacional apresenta limitações na captura completa de padrões característicos do dataset ToN_IoT. A taxa de falsos positivos (FPR) apresentou redução de 25,69% para 14,77%, demonstrando que ambos os modelos alcançam patamares operacionalmente viáveis. Porém, o HSAE mantém vantagem significativa em ambientes com baixa tolerância a alarmes indevidos. A taxa de falsos negativos (FNR) reduziu de 25,46% para 15,55%, indicando que o VAE apresenta maior propensão à não identificação de ataques legítimos comparado ao HSAE. O F1-score aumentou de 72%



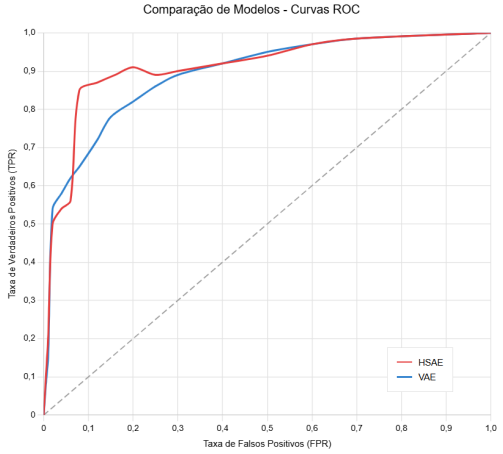
(a) Curva ROC para DDoS



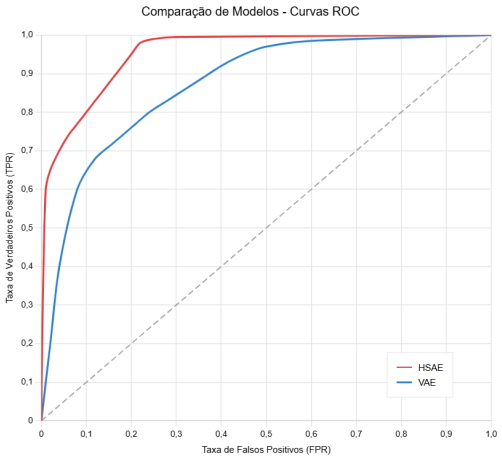
(b) Curva ROC para DoS Slowhttptest



(c) Curva ROC para DoS GoldenEye



(d) Curva ROC para DoS Slowloris



(e) Curva ROC para DoS Hulk

Figura 3. Comparação das curvas ROC para os modelos HSAE e o VAE usando o dataset CICIDS2017.

para 85%, e o *recall* evoluiu de 75% para 84%, evidenciando que a arquitetura híbrida proporciona cobertura superior na identificação dos ataques. O valor da AUC passou de 0,79 para 0,91, representando um salto qualitativo na capacidade discriminativa, sugerindo que a arquitetura híbrida do HSAE é particularmente efetiva na separação entre tráfego benigno e malicioso.

Para ataques DoS, observou-se comportamento similar com nuances interessantes, onde a precisão elevou-se de 69% para 91%, demonstrando superioridade consistente do HSAE. O FPR diminuiu de 26,34% para 22,11%, indicando que ambos os modelos operam em faixas de falsos positivos relativamente controladas, com o HSAE apresentando vantagem operacional. O *recall* apresentou aumento de 73% para 78%, evidenciando que o HSAE proporciona cobertura superior na detecção de ataques. O F1-*score* progrediu de 71% para 84%, e a AUC demonstrou evolução de 0,76 para 0,77, confirmando o aprimoramento da capacidade de distinção entre tráfego legítimo e malicioso, embora a diferença seja menor neste cenário específico, possivelmente devido às características menos complexas dos ataques DoS em comparação com DDoS.

No cenário de ataques Ransomware, o HSAE manteve a tendência de superioridade, com diferenças proporcionalmente menores, evidenciando que o VAE demonstra competência considerável neste domínio específico. A precisão aumentou de 67% para 86%, e o FPR reduziu de 28,09% para 13,93%. Isso indica que ambos os modelos alcançam níveis de desempenho relevantes para aplicações práticas, com o HSAE se destacando por apresentar resultados superiores. O *recall* evoluiu de 71% para 85%, acompanhado pelo F1-score que progrediu de 70% para 85%. A AUC apresentou melhoria de 0,76 para 0,86, sugerindo que as características específicas dos ataques ransomware são adequadamente capturadas por ambas as arquiteturas, com vantagem maior para o modelo híbrido.

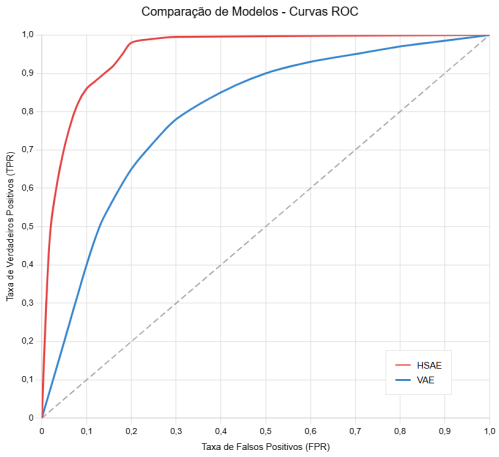
As curvas ROC apresentadas na Figura 4 ilustram a separação entre tráfego benigno e malicioso nos diferentes ataques, evidenciando visualmente o desempenho dos modelos analisados. A linha vermelha representa o modelo 1, correspondente ao HSAE (*Hybrid Scoring Autoencoder*). Já a linha azul representa o modelo 2, correspondente ao VAE (*Variational Autoencoder*).

Os melhores resultados do HSAE em todos os tipos de ataques avaliados sugerem que as limitações observadas do VAE estão relacionadas às características próprias do dataset ToN_IoT, que trabalha com dados de tráfego de rede IoT com complexidade temporal e espacial específica, onde a arquitetura híbrida do HSAE demonstra maior adequação para este tipo de dados.

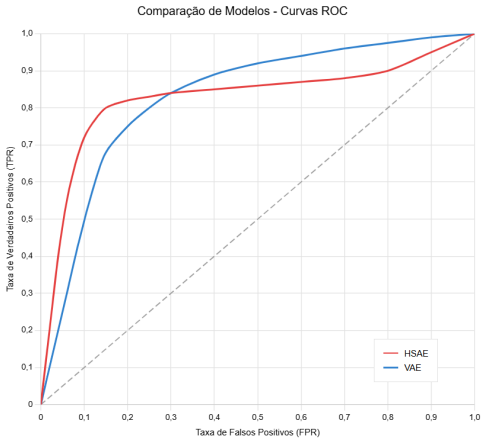
7. Conclusão

Este trabalho apresentou o HSAE, uma arquitetura híbrida não supervisionada baseada em *autoencoder*, construída para lidar com a detecção de ataques *Zero-Day* em cenários realistas e heterogêneos, mantendo baixos índices de falsos positivos e alta precisão. Ao combinar o erro de reconstrução com uma saída auxiliar, a abordagem proposta mostrou-se eficaz na identificação de tráfego anômalo, mesmo na ausência de dados maliciosos previamente rotulados no processo de treinamento.

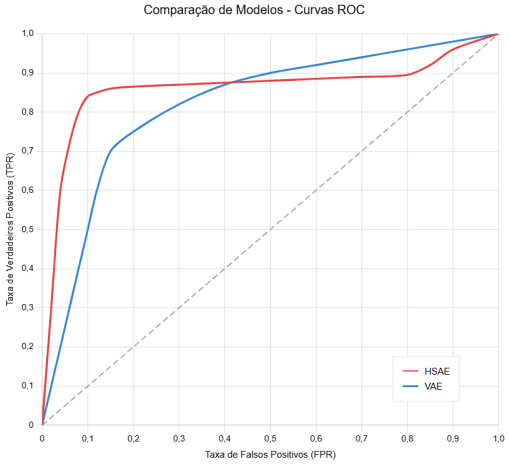
A avaliação empírica, conduzida em dois cenários distintos com os conjuntos de dados CICIDS2017 e ToN_IoT, demonstrou que o HSAE supera o modelo de referên-



(a) Curva ROC para DDoS



(b) Curva ROC para DoS



(c) Curva ROC para Ramsoware

Figura 4. Comparação das curvas ROC para os modelos HSAE e VAE usando o dataset ToN_IoT.

cia VAE em métricas-chave como AUC-ROC, precisão, F1-score e FPR. Tais resultados evidenciam não apenas a robustez do modelo frente à variabilidade do tráfego legítimo, mas também sua capacidade de generalização em contextos heterogêneos, incluindo redes corporativas e ambientes críticos baseados em IoT e IIoT. Além disso, a proposta introduz uma metodologia de pré-processamento transparente e reprodutível, que contribui para a mitigação de vieses e falhas comuns em experimentos com aprendizado não supervisionado. Essa abordagem, aliada à leveza computacional da arquitetura, torna o HSAE uma alternativa promissora para aplicações em tempo real e com restrições de recursos.

Para estudos futuros, pretendemos aprimorar ainda mais o desempenho do modelo. Uma das possibilidades que estamos explorando é o uso de técnicas baseadas em *stacked ensemble*, que podem contribuir para tornar o modelo mais sensível sem comprometer a precisão. Também planejamos adaptar a arquitetura para viabilizar a detecção em tempo real, alinhando a proposta às necessidades de ambientes operacionais críticos. Por

fim, uma etapa importante será a incorporação de mecanismos de resposta automática, com a implementação de políticas de mitigação baseadas nos ataques detectados, abrindo caminho para um sistema mais proativo e adaptativo no contexto da defesa cibernética.

Agradecimentos

Este trabalho foi financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP (Proc. 2018/23098-0) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Proc. 162441/2021-5).

Referências

Abdulganiyu, O. H., Ait Tchakoucht, T., and Saheed, Y. K. (2023). A systematic literature review for network intrusion detection system (IDS). *International journal of information security*, 22(5):1125–1162.

Ahmad, R., Alsmadi, I., Alhamdani, W., and Tawalbeh, L. (2023). Zero-day attack detection: a systematic literature review. *Artificial Intelligence Review*, 56(10):10733–10811.

Alkasassbeh, M. and Al-Haj Baddar, S. (2023). Intrusion detection systems: A state-of-the-art taxonomy and survey. *Arabian Journal for Science and Engineering*, 48(8):10021–10064.

Alrayes, F. S., Zakariah, M., Amin, S. U., Khan, Z. I., and Helal, M. (2024). Intrusion detection in IoT systems using denoising autoencoder. *IEEE Access*.

Bank, D., Koenigstein, N., and Giryes, R. (2023). Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374.

Beaman, C., Barkworth, A., Akande, T. D., Hakak, S., and Khan, M. K. (2021). Ransomware: Recent advances, analysis, challenges and future research directions. *Computers & security*, 111:102490.

Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., and Xu, Y. (2024). Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review*, 57(2):28.

Cheng, J.-M. and Wang, H.-C. (2004). A method of estimating the equal error rate for automatic speaker verification. In *2004 International Symposium on Chinese Spoken Language Processing*, pages 285–288. IEEE.

De Neira, A. B., Kantarci, B., and Nogueira, M. (2023). Distributed denial of service attack prediction: Challenges, open issues and opportunities. *Computer Networks*, 222:109553.

Guo, Y. (2023). A review of machine learning-based zero-day attack detection: Challenges and future directions. *Computer communications*, 198:175–185.

Hajj, S., El Sibai, R., Bou Abdo, J., Demerjian, J., Makhoul, A., and Guyeux, C. (2021). Anomaly-based intrusion detection systems: The requirements, methods, measurements, and datasets. *Transactions on Emerging Telecommunications Technologies*, 32(4):e4240.

Halvorsen, J., Izurieta, C., Cai, H., and Gebremedhin, A. (2024). Applying generative machine learning to intrusion detection: A systematic mapping study and review. *ACM Computing Surveys*, 56(10):1–33.

Lu, H., Zhao, Y., Song, Y., Yang, Y., He, G., Yu, H., and Ren, Y. (2024). A transfer learning-based intrusion detection system for zero-day attack in communication-based train control system. *Cluster Computing*, 27(6):8477–8492.

Mbona, I. and Eloff, J. H. (2022). Detecting zero-day intrusion attacks using semi-supervised machine learning approaches. *IEEE Access*, 10:69822–69838.

Moustafa, N. (2021). A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. *Sustainable Cities and Society*, 72:102994.

Narayan, S. (1997). The generalized sigmoid activation function: Competitive supervised learning. *Information sciences*, 99(1-2):69–82.

Oluwadare, S. and ElSayed, Z. (2023). A survey of unsupervised learning algorithms for zero-day attacks in intrusion detection systems. In *The International FLAIRS Conference Proceedings*, volume 36.

Pratiwi, H., Windarto, A. P., Susliansyah, S., Aria, R. R., Susilowati, S., Rahayu, L. K., Fitriani, Y., Merdekawati, A., and Rahadjeng, I. R. (2020). Sigmoid activation function in selecting the best model of artificial neural networks. In *Journal of physics: conference series*, volume 1471, page 012010. IOP Publishing.

Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

Razaulla, S., Fachkha, C., Markarian, C., Gawanmeh, A., Mansoor, W., Fung, B. C., and Assi, C. (2023). The age of ransomware: A survey on the evolution, taxonomy, and research directions. *IEEE Access*, 11:40698–40723.

Sharafaldin, I., Lashkari, A. H., Ghorbani, A. A., et al. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018)*.

Verkerken, M., D'hooge, L., Sudyana, D., Lin, Y.-D., Wauters, T., Volckaert, B., and De Turck, F. (2023). A novel multi-stage approach for hierarchical intrusion detection. *IEEE Transactions on Network and Service Management*, 20(3):3915–3929.

Yang, Z., Liu, X., Li, T., Wu, D., Wang, J., Zhao, Y., and Han, H. (2022). A systematic literature review of methods and datasets for anomaly-based network intrusion detection. *Computers & Security*, 116:102675.

Zavrak, S. and Iskefiyeli, M. (2020). Anomaly-based intrusion detection from network flow features using variational autoencoder. *IEEE Access*, 8:108346–108358.

Zoppi, T., Ceccarelli, A., and Bondavalli, A. (2021). Unsupervised algorithms to detect zero-day attacks: Strategy and application. *IEEE Access*, 9:90603–90615.