PAMS: Um Arcabouço para Detecção de Intrusão em Internet das Coisas

Jan Martins Teixeira[1], Ian Vilar Bastos[2], Dalbert M. Mascarenhas[3], Igor Monteiro Moraes[1]

[1] Laboratório MídiaCom, IC/TCC/PGC – UFF
[2] Departamento de Eletrônica e Telecomunicações – UERJ
[3] CEFET/RJ, Petrópolis

jan.teixeira@id.uff.br, dalbert.mascarenhas@cefet-rj.br,
ian.bastos@eng.uerj.br, igor@ic.uff.br

Resumo. *A detecção de intrusão em Internet das Coisas baseada em aprendizado de máquina supervisionado enfrenta um gargalo crítico: o sobreajuste às assinaturas de ataque do domínio de treino, que impede a generalização para redes com perfis de protocolo distintos. Este trabalho investiga as causas dessa fragilidade e propõe o arcabouço Protocol-Aware Modeling System (PAMS), que modela exclusivamente o comportamento benigno e sinaliza como anomalia qualquer desvio desse padrão. O PAMS combina um Autoencoder Variacional e um Deep SVDD, cujos scores são integrados por rank-based fusion invariante à escala e calibrados por grupo de protocolo. Avaliado em quatro conjuntos do Canadian Institute for Cybersecurity sob regime de rodízio inter-domínio, o PAMS foi o único modelo a manter desempenho consistente nos três domínios de teste, superando modelos supervisionados e métodos clássicos de detecção de anomalia.*

Abstract. *Supervised machine learning-based intrusion detection for the Internet of Things faces a critical bottleneck: overfitting to attack signatures from the training domain, which prevents generalization to networks with distinct protocol profiles. This work investigates the causes of this fragility and proposes the Protocol-Aware Modeling System (PAMS), a framework that models exclusively benign behavior and flags any deviation from this pattern as an anomaly. PAMS combines a Variational Autoencoder and a Deep SVDD, whose scores are integrated via scale-invariant rank-based fusion and calibrated per protocol group using only training benign traffic. Evaluated on four Canadian Institute for Cybersecurity datasets under a cross-domain rotation regime, PAMS was the only model to maintain consistent performance across all three test domains, outperforming both supervised models and classical anomaly detection methods.*

1. Introdução

O crescimento do uso de dispositivos de Internet das Coisas (*Internet of Things* - IoT) em setores como saúde, manufatura e infraestrutura crítica [Rehman et al. 2025, Manzini et al. 2025, Miranda and Braga 2025], aumenta a preocupação com sua segurança, pois dispositivos comprometidos podem facilitar ataques em larga escala, vazamentos de dados e interrupções de serviço. De sensores industriais a eletrodomésticos

inteligentes, esses dispositivos geram fluxos de dados que podem, por exemplo, revelar padrões comportamentais valiosos para atacantes [Baena et al. 2025].

Sistemas de detecção de intrusão (*Intrusion Detection Systems* - IDSes) baseados em aprendizado de máquina (AM) ou aprendizado profundo (AP) emergem como ferramentas essenciais para identificar atividades maliciosas no tráfego de rede em IoT [Hozouri et al. 2025]. Esses sistemas dependem de que tais algoritmos sejam treinados em conjuntos de dados que contenham padrões de tráfego benigno e malicioso. No entanto, a implementação prática de IDSes em IoT enfrenta desafios significativos, particularmente em relação à qualidade e representação dos dados [Morshedi and Matinkhah 2025]. Embora o tráfego de rede em IoT possa apresentar uma predominância de instâncias benignas ou maliciosas, a classe dominante varia de acordo com o conjunto de dados e com as categorias específicas de ataque incluídas. Portanto, o grau e o tipo de desbalanceamento de classes são determinados pelo processo de coleta de dados e pelos cenários de ameaça representados. Consequentemente, a influência do desbalanceamento de classes no treinamento e na avaliação dos modelos é específica de cada conjunto de dados, e métricas de desempenho convencionais podem não capturar com precisão as reais capacidades de detecção dos IDS em diferentes ambientes de IoT e perfis de ataque [Suresh and Guttag 2021].

O Canadian Institute for Cybersecurity (CIC) desenvolveu diversos conjuntos de dados de referência, incluindo CICIoT2023 [Neto et al. 2023], CIC-IoT-DIAD [Rabbani et al. 2025], CICIDS2017 e CICIDS2018 [Sharafaldin et al. 2018], amplamente utilizados em pesquisas de detecção de intrusão. Embora sejam recursos valiosos, esses conjuntos herdam as características de desbalanceamento das redes reais e apresentam limitações adicionais: ferramentas proprietárias de coleta [Habibi Lashkari 2018] restringem a reprodutibilidade [Pekar and Jozsa 2024], a documentação incompleta obscurece as metodologias de ataque e conversões de formato de capturas brutas de pacotes para formatos estruturados limitam a flexibilidade analítica [Vitorino et al. 2025]. Essas restrições levantam questões sobre a representatividade dos conjuntos de dados e sua adequação para desenvolver modelos de detecção robustos e generalizáveis.

Este artigo investiga o impacto do desbalanceamento de classes na classificação de tráfego IoT com AM e AP, utilizando quatro conjuntos de dados do CIC. Demonstramos, por meio dos classificadores Random Forest (RF), Decision Tree (DT), XGBoost (XGB), Multi-Layer Perceptron (MLP), Gaussian Naive Bayes (GNB), Logistic Regression (LR), Deep SVDD (SVDD), Isolation Forest (IF), Elliptic Envelope (EE)[1] que em conjuntos de dados com distribuição assimétrica de classes, a acurácia revela-se uma métrica insuficiente, pois valores expressivos podem coexistir com um baixo poder de discriminação interclasse. Técnicas de sobreamostragem elevam o desempenho no treinamento, mas essa melhoria não se sustenta em validação entre domínios, evidenciando limitações na generalização.

Para mitigar esse problema, propomos o *PAMS (Protocol-Aware Multi-SVDD)*, uma arquitetura baseada em modelagem de normalidade que combina *Variational Autoencoders* (VAE) para extração de características latentes e *Deep SVDD* para delimitação de hiperesferas benignas. Ambos os componentes recebem como entrada um vetor fixo

[1]Todos os algoritmos foram utilizados a partir da biblioteca *Scikit-learn* [scikit-learn 2025], com exceção do Deep SVDD, cuja implementação foi baseada no trabalho [Ruff et al. 2018].

de features comuns aos quatro conjuntos de dados utilizados, garantindo compatibilidade estrutural independentemente do domínio de origem.

As contribuições deste artigo são: (i) evidenciar como o desbalanceamento distorce a avaliação; e (ii) propor e validar o Framework PAMS, uma arquitetura não supervisionada baseada em modelagem de normalidade com contexto de serviço.

O restante deste trabalho está organizado da seguinte forma. A Seção 2 revisa trabalhos relacionados e o contexto dos conjuntos de dados. A Seção 3 descreve os dados utilizados e as etapas de pré-processamento. A Seção 4 apresenta a metodologia de avaliação e a arquitetura do framework PAMS. A Seção 5 detalha os resultados da avaliação experimental inter-domínio. Finalmente, a Seção 6 conclui o artigo com descobertas e direções futuras de pesquisa.

2. Trabalhos Relacionados

Diversos estudos [Bochie et al. 2020, Churcher et al. 2021, Lopes et al. 2022] abordam a classificação de tráfego de rede em IoT empregando algoritmos de AM para distinguir com precisão entre diferentes tipos de tráfego. Esses estudos geralmente dependem de conjuntos de dados que contêm registros de tráfego benigno e malicioso, que podem ser rotulados ou não. Existem diferentes conjuntos de dados desse tipo disponíveis na literatura [Bezerra et al. 2018, Koroniotis et al. 2018, Moustafa 2019b, Moustafa 2019a, Ferrag et al. 2022].

Silveira *et al.* [Silva et al. 2021] ressaltam o impacto do desbalanceamento de classes nos conjuntos CICIDS2017 e CICIDS2018 [Sharafaldin et al. 2018]. Para mitigar esse problema, aplicam-se técnicas de subamostragem, que reduzem a quantidade de amostras da classe majoritária, promovendo uma melhor representatividade dos dados.

Lopes *et al.* [Lopes et al. 2022] propõem o *Energy-Based Flow Classifier* (EFC), um classificador treinado exclusivamente com tráfego benigno, visando contornar a dificuldade de obtenção de amostras maliciosas representativas. Os resultados apresentados indicam que classificadores tradicionais de duas classes sofrem significativa degradação de desempenho em cenários inter-domínio, especialmente quando há mudanças na distribuição dos dados ou na arquitetura das *botnets*.

Pereira *et al.* [Pereira et al. 2021] destacam que o desbalanceamento de classes é um problema recorrente em tarefas de classificação, especialmente quando há grande disparidade entre o número de instâncias das classes. Embora técnicas de reamostragem, como sobreamostragem e subamostragem, sejam amplamente empregadas em cenários binários, multiclasse e multirrótulo, poucos estudos investigam esse problema no contexto da classificação hierárquica. Nesse sentido, os autores propõem métricas e estratégias específicas para medir e mitigar o desbalanceamento em bases hierárquicas, demonstrando que métodos de reamostragem adequadamente adaptados podem melhorar significativamente o desempenho dos classificadores.

Gao *et al* fornecem uma pesquisa abrangente sobre aprendizado com dados desbalanceados [Gao et al. 2025], organizando estratégias de mitigação em quatro categorias: rebalanceamento de dados, representação de características, ajustes de estratégia de treinamento e aprendizado conjunto. A pesquisa enfatiza que o desbalanceamento é um desafio estrutural em muitos domínios do mundo real como finanças, saúde e sistemas industri-

ais, nos quais classes minoritárias frequentemente representam menos de 1% dos dados. Para lidar com o desbalanceamento de classes, Chun e Lin, por exemplo, propõem um método de classificação de ataques que combina uma rede neural de múltiplas camadas com Redes Geradoras Adversariais (RGAs) para geração de dados [Chu and Lin 2023]. Essa abordagem melhora o desempenho em classes sub-representadas, atingindo acurácia, precisão, *recall* e F1-*score* acima de 90% em múltiplos conjuntos de dados. Há trabalhos que também avaliam se o impacto do desbalanceamento de dados afeta de maneira diferente os modelos de AM. Churcher *et al.*, por exemplo, mostram que RF e Redes Neurais Artificiais (RNAs) alcançam alto desempenho dependendo do balanceamento dos dados e do tipo de classificação, sugerindo sua adequação para integração em IDS em cenários IoT com múltiplos ataques [Churcher et al. 2021].

Entre os métodos de detecção de anomalias sem supervisão amplamente utilizados, destacam-se o *Isolation Forest* [Liu et al. 2008] e o *Elliptic Envelope* [Croux and Haesbroeck 1999]. O *Isolation Forest* opera por meio de particionamento aleatório recursivo do espaço de features: amostras anômalas, por serem estatisticamente raras e distantes da massa de dados, tendem a ser isoladas em partições mais rasas da árvore, recebendo scores de anomalia mais elevados. O *Elliptic Envelope*, por sua vez, ajusta uma distribuição gaussiana multivariada aos dados de treino e classifica como anômalos os pontos cuja distância de Mahalanobis ao centroide estimado excede um limiar determinado pela contaminação esperada. Ambos os métodos assumem, implicitamente, que os dados normais formam uma região compacta e unimodal no espaço de features.

A avaliação de detectores de anomalia baseados em aprendizado de uma classe (OCC) exige rigor metodológico para garantir comparações justas. Conforme apontado pela literatura recente [Hayashi et al. 2026] estratégias que utilizam exposição a outliers ou modelos pré-treinados com múltiplas classes, embora eficazes em cenários práticos, desvirtuam a essência do paradigma OCC e devem ser excluídas de experimentos que se propõem a avaliar exclusivamente esse tipo de abordagem.

Diferentemente de trabalhos focados em validação intra-domínio, submetemos os classificadores a uma avaliação externa, inter-domínios, utilizando quatro conjuntos de dados originários de topologias e infraestruturas distintas. Conforme detalhamos na Seção 3, testes estatísticos de homogeneidade e análise de portas de destino confirmam que esses conjuntos não partilham distribuições similares. A Tabela 1 evidencia essa transição tecnológica: enquanto o CICIDS2018 é dominado por tráfego de gerenciamento remoto 31,72% na porta 3389, RDP, o CIC-IoT-DIAD concentra 56,14% do tráfego na porta 8009, utilizada por protocolos de automação e dispositivos *Chromecast*, impondo ao modelo um desafio real de generalização.

Para enfrentar esse cenário, este artigo propõe o PAMS, cuja arquitetura sustenta-se em três pilares fundamentais. Primeiro, ancoramos a segurança no tráfego benigno por meio de um *Ensemble* Híbrido. Este comitê utiliza um VAE [Kingma and Welling 2013] para comprimir fluxos normais em um espaço latente contínuo de baixa dimensão, combinado ao Deep SVDD para estabelecer uma fronteira geométrica de normalidade adaptativa. A combinação dos dois componentes é realizada por *rank-based fusion*: os scores são convertidos para postos percentuais antes da fusão, eliminando a sensibilidade à escala absoluta entre domínios. Segundo, introduzimos a ciência do protocolo na mode-

lagem; ao contrário de abordagens globais que tratam conexões heterogêneas de forma genérica [Lopes et al. 2022, Churcher et al. 2021], o PAMS intercepta a porta de destino como âncora semântica. Ao integrar o contexto do serviço agrupando o tráfego em categorias como *Web*, telemetria IoT, consultas DNS e administração remota, o modelo restringe a compressão latente e o limiar de decisão a cada protocolo individualmente, calibrado no percentil dos scores dos benignos do treino por grupo, com *fallback* global para protocolos não observados. Terceiro, validamos o modelo sob um regime exaustivo de rodízio inter-domínio entre as quatro bases.

Tabela 1. Principais portas de destino identificadas.

Conjuntos	Porta Principal	Freq. (%)	Segunda Porta	Freq. (%)
CICIDS2017	137 (NetBIOS)	1,49%	389 (LDAP)	1,18%
CICIDS2018	3389 (RDP)	31,72%	445 (SMB)	12,30%
CICIoT2023	32100 (IoT-API)	7,80%	8009 (AJP)	3,00%
CIC-IoT-DIAD	8009 (AJP)	56,14%	32100 (IoT-API)	5,73%

2.1. Conjuntos de Dados do CIC

Os conjuntos de dados disponibilizados pelo Canadian Institute for Cybersecurity (CIC) estão entre os mais citados na literatura. Por isso, são os conjuntos escolhidos para a avaliação deste artigo.

Sharafaldin *et al.* introduzem os conjuntos de dados CICIDS2017[2] e CICIDS2018[3], que abrangem uma variedade de ataques, como DoS, DDoS, força bruta, XSS, injeção de SQL, infiltração, varredura de portas e botnets [Sharafaldin et al. 2018]. Esses conjuntos totalmente rotulados fornecem mais de 80 características de tráfego de rede extraídas usando a ferramenta CICFlowMeter [Habibi Lashkari 2018].

Neto *et al.* apresentam o conjunto de dados CICIoT2023, construído a partir de uma topologia realista com 105 dispositivos IoT [Neto et al. 2023]. O conjunto reúne 33 tipos de ataques distribuídos em sete categorias e busca suprir limitações de bases anteriores, como a baixa diversidade de dispositivos e de cenários de ataques em IoT. Diferentemente de bases legadas, o CICIoT2023 é predominantemente composto por comunicação máquina-a-máquina, seu tráfego benigno reproduz o elevado volume de comunicação telemétrica, gerando fluxos contínuos de pacotes curtos e protocolos de descoberta ou publicação/assinatura, como MQTT, mDNS e SSDP. O estudo também avalia algoritmos de aprendizado de máquina para detecção de tráfego malicioso, destacando o potencial do conjunto no desenvolvimento de soluções práticas de segurança para IoT.

Rabbani *et al.* propõem um conjunto de dados para identificação de dispositivos IoT e detecção de anomalias que emprega uma extração de características híbrida [Rabbani et al. 2025]. O conjunto se apoia em duas metodologias:

1. Abordagem Baseada em Pacotes: Fornecendo 133 atributos derivados de características por pacote, incluindo métricas de *HTTPS*, *handshake*, *User-Agent*, *stream*, *channel* e *jitter*.

[2]A versão utilizada do dataset CICIDS2017 [chethuhn 2021] pode ser obtida no repositório Kaggle mantido por *chethuhn*

[3]A versão utilizada do dataset CICIDS2018 [SolarMainframe 2021] pode ser obtida no repositório Kaggle mantido por *SolarMainframe*

2. Abordagem Baseada em Fluxos: Fornece 84 atributos extraídos pela ferramenta própria, o CICFlowMeter [Habibi Lashkari 2018].

3. Distribuição de Dados e Pré-processamento

Esta seção apresenta uma análise da distribuição das classes nos conjuntos de dados utilizados, com o objetivo de caracterizar o grau de desbalanceamento presente e discutir suas implicações na avaliação de modelos de AM e AP para detecção de ataques. A Tabela 6 apresenta os valores absolutos das amostras por classe. Observa-se que todos os conjuntos analisados apresentam forte desbalanceamento entre classes. No CICIDS2017, por exemplo, a classe BENIGN representa aproximadamente 90% dos registros, enquanto ataques como BRUTE_FORCE e SQL_INJECTION correspondem a menos de 0,1%. No CICIDS2018, a distribuição é similar, com BENIGN atingindo cerca de 96% das amostras. O CICIoT2023 apresenta menor predominância da classe BENIGN aproximadamente 50%, neste conjunto os ataques DoS constituem quase 37% dos dados, e classes como SQL_INJECTION e XSS permanecem abaixo de 2%. No CIC-IoT-DIAD, DoS é a classe majoritária cerca de 80%, enquanto BENIGN representa apenas 16%, e ataques menores somam o restante.

Esses desequilíbrios evidenciam a necessidade de técnicas de balanceamento, pois podem impactar significativamente a capacidade de generalização dos algoritmos, favorecendo a predição das classes majoritárias em detrimento das minoritárias podendo gerar modelos com alta acurácia global, mas baixa capacidade de detectar as classes minoritárias. Assim, a avaliação de desempenho deve considerar métricas que reflitam adequadamente a distribuição dos rótulos, orientando a adoção de estratégias de pré-processamento e técnicas de modelagem que mitiguem o viés introduzido pelos dados.

O trabalho também se baseia na grande heterogeneidade entre os domínios analisados. Foram computados testes estatísticos de Kolmogorov-Smirnov e estatística de Levene [4] sobre subconjuntos de *atributos* entre os conjuntos. O resultado apontou p-valores nulos. Na prática, isso prova que validar algoritmos treinados no CIC frente à novos fluxos de IoT é um teste de generalização, exigindo real capacidade de abstração dos dados.

3.1. Limpeza de Dados e Geração de Dados

O pré-processamento prepara sistematicamente os dados antes de qualquer análise, garantindo sua qualidade e confiabilidade. Essa etapa compreende a limpeza de dados, remoção de valores nulos, duplicatas e inconsistências. Embora a remoção de valores atípicos (*outliers*) seja comum, neste trabalho opta-se por mantê-los devido ao objetivo de detectar ataques de rede.

A transformação dos dados inclui a padronização de valores e codificação de variáveis categóricas, assegurando que as informações fornecidas aos algoritmos estejam adequadamente representadas. Tais procedimentos evitam problemas relacionados a diferentes escalas entre atributos e interpretações inadequadas das informações.

[4]A implementação das rotinas de pré-processamento, assim como a execução das estatísticas descritivas, encontra-se publicamente disponível em: `https://anonymous.4open.science/r/sbseg-C572/`.

Ao final do pré-processamento, os dados são divididos em conjuntos de treinamento e teste na proporção 80/20, utilizando amostragem estratificada, é fixada a semente aleatória *random_state=42* para garantir reprodutibilidade dos resultados. Para a avaliação de generalização inter-domínio, os modelos foram submetidos à totalidade dos dados dos conjuntos alvo.

No contexto de AP, investigaram-se quatro arquiteturas distintas aplicadas sobre os mesmos dados. Inicialmente, utilizou-se uma rede *Multi-Layer Perceptron* como baseline supervisionado, composta por camadas densas de 128 e 64 neurônios, otimizada com `Adam`, taxa de aprendizado, $lr = 0{,}001$ e função de perda *Binary Cross-Entropy*, servindo como referência consolidada. Em seguida, foram empregados VAE, com arquitetura simétrica. Encoder $128 \rightarrow 64$ e Decoder $64 \rightarrow 128$ e espaço latente de 16 dimensões, incorporando *BatchNormalization* e *Dropout* de 0,3 para maior estabilidade, com o objetivo de modelar a distribuição probabilística do tráfego benigno por meio da Divergência de Kullback-Leibler (KL).

Por fim, empregou-se o modelo Deep SVDD, baseado em uma rede *feed-forward* profunda, com $128 \rightarrow 64 \rightarrow 32$, treinada para minimizar a distância Euclidiana entre as representações latentes e um centro fixo, com uso de *Early Stopping* de 10 épocas, explorando uma abordagem geométrica na qual o tráfego benigno é compactado em uma hiperesfera, facilitando a detecção de anomalias fora desse limite.

3.2. Distribuição de Protocolos e Perfil do Tráfego

Compreender quais protocolos dominam o tráfego de rede e como os ataques se distribuem entre eles é essencial para interpretar o comportamento de um sistema de detecção de intrusão. Esta seção analisa a composição dos quatro conjuntos utilizados neste trabalho sob duas perspectivas: a frequência de cada protocolo e a proporção de tráfego benigno e malicioso dentro deles.

A Tabela 2 mostra a participação percentual de cada protocolo no volume total de cada conjunto. Em todos os cenários, DNS, HTTP e HTTPS concentram a maior parte do tráfego. Nos conjuntos de 2017 e 2018, eles respondem por aproximadamente 77% do total de fluxos, com o DNS aparecendo como o protocolo mais volumoso 38,22% e 28,83%, respectivamente. Já nos conjuntos voltados para IoT, a categoria "Outros/Dinâmicos", que agrupa portas não padronizadas, salta para 61,08% no CICIoT2023 e 72,82% no CIC-IoT-DIAD. Essa diferença reflete a diversidade de protocolos proprietários típicos de ambientes IoT, que escapam às classificações tradicionais de porta. O tráfego web composto por HTTP e HTTPS apresenta altos indices em todos os cenários, variando de 21,75% no CICIoT2023 a 38,88% no CICIDS2017. Protocolos como SSH, FTP, Telnet e MQTT são mais escassos aparecendo pouco mais de 0,5% do total.

3.3. Proporção de Tráfego Malicioso por Protocolo

A Tabela 3 apresenta a distribuição percentual interna de tráfego benigno com rótulo 0 e malicioso com rótulo 1, para cada protocolo nos quatro conjuntos analisados. Os percentuais foram calculados com precisão de até três casas decimais, tomando como denominador o total de amostras do respectivo protocolo em cada conjunto. Para a correta interpretação dos valores, é importante observar que a tabela descreve a composição relativa dentro de cada protocolo, e não sua participação no volume global do conjunto esta última será abordada na Tabela 4.

Tabela 2. Distribuição perceantual de protocolos em cada conjunto.

Protocolo	CICIDS2017	CICIDS2018	CICIoT2023	CIC-IoT-DIAD
DNS	38,22%	28,83%	16,63%	6,48%
HTTP	18,71%	15,44%	15,43%	16,02%
HTTPS	20,17%	14,32%	6,32%	3,79%
SSH	0,43%	0,07%	0,06%	–
FTP	0,21%	–	–	–
Telnet	–	0,05%	–	–
SSDP (UPnP)	–	–	0,39%	0,24%
mDNS	–	–	–	0,37%
Outros/Dinâmicos	22,22%	41,27%	61,08%	72,82%

Os protocolos que dominam o tráfego total não são, necessariamente, os mais visados por ataques. O HTTP destaca-se como o vetor de ataque predominante nos conjuntos tradicionais. No CICIDS2017, 49,74% das conexões HTTP são maliciosas, o que equivale a praticamente metade do tráfego desse protocolo. No CICIDS2018, essa proporção reduz-se para 21,48%. Em ambos os conjuntos, DNS, HTTPS e FTP permanecem integralmente benignos, o que indica que os cenários de ataque simulados concentraram-se na exploração de servidores web.

No CIC-IoT-DIAD, 92,70% do tráfego HTTP é malicioso, caracterizando o protocolo como um canal hostil. Adicionalmente, protocolos específicos de IoT, como MQTT e MQTTS, que se mostravam inexpressivos nos conjuntos anteriores, passam a apresentar proporções relevantes de ataques. No CICIoT2023, 9,57% do tráfego MQTT e 8,05% do MQTTS são maliciosos. O DNS, majoritariamente benigno nos demais conjuntos, exibe 7,74% de tráfego malicioso no CIC-IoT-DIAD, sugerindo a exploração de consultas de resolução de nomes como superfície de ataque.

Tabela 3. Distribuição percentual de tráfego.

Protocolo	CICIDS2017		CICIDS2018		CICIoT2023		CIC-IoT-DIAD	
	0	1	0	1	0	1	0	1
DNS	100,00%	0,00%	100,00%	0,00%	97,04%	2,96%	92,26%	7,74%
FTP	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	93,10%	6,90%
HTTP	50,26%	49,74%	78,52%	21,48%	98,85%	1,15%	7,30%	92,70%
HTTPS	100,00%	0,00%	100,00%	0,00%	96,97%	3,03%	79,33%	20,67%
MQTT	–	–	–	–	90,43%	9,57%	92,65%	7,35%
MQTTS	–	–	–	–	91,95%	8,05%	92,86%	7,14%

A Tabela 4 complementa a análise anterior ao apresentar a participação percentual de cada par, protocolo e rótulo, no volume total do conjunto. Diferentemente da Tabela 3, que descreve a proporção interna de cada protocolo, esta tabela considera como denominador o conjunto completo de amostras do conjunto, incluindo todos os protocolos e rótulos. Assim, a soma de todas as células da tabela totaliza 100% para cada conjunto, e valores iguais a 0,00% indicam que a combinação correspondente está presente no conjunto, porém em quantidade inferior a 0,001% do total.

Protocolos como DNS e HTTPS, são majoritariamente benignos nos conjuntos de 2017 e 2018. Nesses cenários, o HTTP constitui o único protocolo que contribui com parcelas expressivas de tráfego malicioso: 9,31% do volume total no CICIDS2017 e 3,32% no CICIDS2018. No CIC-IoT-DIAD, observa-se uma inversão acentuada desse padrão:

14,85% de todo o conjunto é composto por tráfego HTTP malicioso, contra apenas 1,17% de HTTP benigno. O HTTPS responde por 0,78% de tráfego malicioso, e o DNS contribui com 0,50% de ataques. Embora esses últimos percentuais possam parecer reduzidos em termos absolutos, eles sinalizam uma diversificação relevante dos vetores de ataque em ambientes IoT.

Essa disparidade entre os conjuntos acarreta implicações práticas para o desenvolvimento de sistemas de detecção de intrusão. Um modelo treinado exclusivamente nos conjuntos de 2017 e 2018 tende a ignorar ataques que trafegam em protocolos como MQTT ou DNS, pois de acordo com a Tabela 4 estes jamais aparecem nessas bases. Por outro lado, um modelo treinado no CIC-IoT-DIAD pode estabelecer uma correlação entre HTTP e atividade maliciosa, dado que 92,70% das amostras desse protocolo são ataques.

Em nossos experimentos, na validação dos classificadores treinado no CICIDS2017 contra o conjunto CIC-IoT-DIAD, observou-se uma queda no F1-*Macro*. Isso demonstra que, como o modelo nunca observou ataques em MQTT ou DNS durante o treino, ele classifica quase todo esse tráfego como benigno por omissão. Por outro lado, o viés de coleta manifesta-se de forma oposta no treinamento com o CIC-IoT-DIAD. Devido à forte correlação estatística nesse dataset, onde 92,70% das amostras HTTP são maliciosas, o modelo supervisionado tende a rotular preventivamente qualquer fluxo HTTP como ameaça.

Tabela 4. Participação percentual de protocolo e rótulo.

Protocolo	CICIDS2017		CICIDS2018		CICIoT2023		CIC-IoT-DIAD	
	0	1	0	1	0	1	0	1
DNS	38,22%	0,00%	28,83%	0,00%	16,14%	0,49%	5,98%	0,50%
FTP	0,21%	0,00%	0,00%	0,00%	0,02%	0,00%	0,01%	0,00%
HTTP	9,40%	9,31%	12,12%	3,32%	15,25%	0,18%	1,17%	14,85%
HTTPS	20,17%	0,00%	14,32%	0,00%	6,13%	0,19%	3,01%	0,78%
MQTT	–	–	–	–	0,01%	0,00%	0,08%	0,01%
MQTTS	–	–	–	–	0,03%	0,00%	0,13%	0,01%

4. Metodologia e Arquitetura PAMS

A Figura 1 apresenta a metodologia adotada neste trabalho, que compreende a seleção e preparação de conjuntos de dados de detecção de intrusão em IoT, o treinamento de diferentes classificadores de AM e AP, avaliando os desempenhos sob distintos cenários de balanceamento de classes.

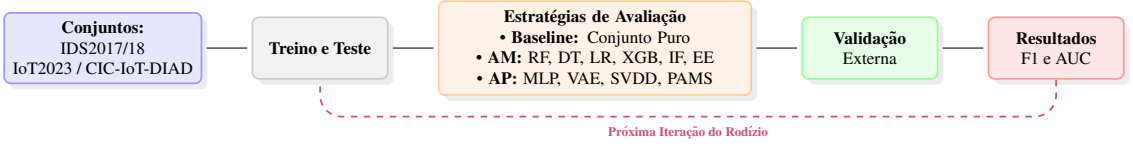

Figura 1. Fluxograma da metodologia experimental com ciclo de rodízio inter-domínio.

A preparação dos conjuntos de dados teve início a partir do CICIoT2023 *pcap*. A partir dos arquivos de captura de tráfego, foram extraídos atributos compatíveis com os

demais conjuntos usando a biblioteca PyShark [KimiNewt 2025]. Nem todos os atributos podem ser gerados diretamente, pois os conjuntos, utilizados neste trabalho, são gerados por meio de ferramentas próprias para a captura, resultando em atributos únicos que não podem ser obtidos diretamente com outros instrumentos de análise de tráfego. Diante desse cenário, duas estratégias podem ser adotadas: (i) utilizar as colunas dos conjuntos selecionados para gerar atributos sintéticos que simulem características reais do tráfego de rede; ou (ii) empregar métodos alternativos para reconstruir os fluxos de rede a partir das informações disponíveis. Neste trabalho, opta-se pela segunda abordagem.

Buscamos investigar o ganho informacional máximo proporcionado pelos atributos por meio do mapeamento de quatro conjuntos de dados, com o objetivo de identificar métricas comuns entre eles. Esse processo resultou na identificação de 59 atributos compartilhados. Para avaliar o ganho de informação, submetemos esses atributos a testes de correlação baseados em três critérios complementares: *SelectKBest* com pontuações da Análise de Variância (ANOVA), medições de Informação Mútua e *SelectFPR* com restrições baseadas na Taxa de Falsos Positivos[5]. Esse procedimento permitiu identificar um subconjunto de variáveis com alta pertinência informacional e aplicabilidade independente da topologia de rede, preservando atributos estruturais essenciais, como a porta de destino. Contudo, apesar da identificação desses subconjuntos, optamos por utilizar a totalidade dos 59 atributos, uma vez que a quantidade de atributos selecionados variou significativamente conforme o método empregado, como evidenciado nas matrizes de comparação apresentada na Tabela 5.

Tabela 5. Quantidade de atributos em comum entre os métodos – CICIDS2017

	Percentile	**SelectFwe**	**SelectFpr**	**Anova**	**MI_KBest**	**MI_Percentile**
Percentile	28.0	28.0	28.0	28.0	21.0	20.0
SelectFwe	28.0	50.0	50.0	30.0	27.0	25.0
SelectFpr	28.0	50.0	52.0	30.0	28.0	26.0
Anova	28.0	30.0	30.0	30.0	21.0	20.0
MI_KBest	21.0	27.0	28.0	21.0	30.0	28.0
MI_Percentile	20.0	25.0	26.0	20.0	28.0	28.0

Tabela 6. Distribuição dos rótulos nos conjuntos utilizados.

Rótulo	**CICIDS2017**	**CICIDS2018**	**CICIoT2023**	**CIC-IoT-DIAD**
BENIGN	2 273 097	13 484 708	463 350	398 330
DoS	231 073	461 912	345 588	1 642 744
BRUTE_FORCE	1 507	611	10 598	3 619
XSS	652	230	3 961	3 377
SQL_INJECTION	21	87	9 166	6 603

4.1. Abordagem Híbrida: Autoencoders Variacionais e Deep SVDD

Diferentemente de modelos supervisionados que aprendem fronteiras entre classes, o PAMS modela a normalidade a partir de tráfego exclusivamente benigno, combinando VAE e Deep SVDD. Ambos os componentes recebem como entrada um vetor fixo de

[5]Os resultados completos desses testes estão disponíveis no repositório `https://anonymous.4open.science/r/sbseg-C572/`

$D = 59$ features comuns aos quatro conjuntos de dados utilizados, garantindo compatibilidade estrutural independentemente do domínio de origem. O escalonamento é ajustado uma única vez sobre os benignos do conjunto de treino e aplicado sem re-ajuste a todos os conjuntos de teste, preservando a separação estrita entre treino e avaliação.

O VAE projeta o tráfego benigno em um espaço latente contínuo de dimensão $d = 16$, por meio de um encoder com camadas ocultas de 128 e 64 neurônios, seguido de um decoder simétrico que reconstrói o sinal na dimensão original. A função de custo pondera explicitamente o erro de reconstrução e a divergência KL:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}\left[\|x - \hat{x}\|^2\right] + \beta \cdot D_{\text{KL}}\big(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, I)\big)\,, \tag{1}$$

com $\beta = 1{,}0$. O score de anomalia do VAE é o erro de reconstrução quadrático médio por amostra:

$$s_{\text{VAE}}(x) = \frac{1}{D}\|x - \hat{x}\|^2. \tag{2}$$

O Deep SVDD é treinado de forma independente, também sobre os benignos do treino. Seu encoder reduz a entrada de $D = 59$ dimensões por camadas ocultas de 256, 128 e 64 neurônios, com inicialização `he_normal`, *Batch Normalization* e *Dropout* de $0{,}1$, chegando a um espaço latente de dimensão $d = 32$. O centro fixo $\mathbf{c}$ é inicializado como a média das projeções de uma subamostra dos benignos do treino, e o modelo minimiza a distância euclidiana ao quadrado entre as representações latentes e esse centro:

$$\mathcal{L}_{\text{SVDD}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{z}^{(i)} - \mathbf{c}\right\|^2. \tag{3}$$

O score de anomalia do SVDD é a distância euclidiana ao quadrado ao centro $\mathbf{c}$:

$$s_{\text{SVDD}}(x) = \|\mathbf{z} - \mathbf{c}\|^2. \tag{4}$$

4.2. Ensemble por Fusão de Postos

A combinação direta dos scores brutos dos dois componentes apresenta um problema estrutural: as magnitudes absolutas de s_{VAE} e s_{SVDD} variam consideravelmente entre domínios distintos. Um meta-aprendiz linear treinado sobre esses scores sobreajusta à escala do domínio de origem, como evidenciado por pesos da ordem de $w_{\text{VAE}} \approx 36{,}78$ e $w_{\text{SVDD}} \approx -3{,}08$ observados em experimentos preliminares. Para contornar esse problema, a fusão é realizada sobre postos percentuais. Cada score é convertido para seu posto relativo $\rho \in [0, 1]$ antes da combinação, tornando o ensemble invariante à escala absoluta:

$$s_{\text{ens}}(x) = w_{\text{VAE}} \cdot \rho_{\text{VAE}}(x) + w_{\text{SVDD}} \cdot \rho_{\text{SVDD}}(x), \tag{5}$$

onde $w_{\text{VAE}} + w_{\text{SVDD}} = 1$. Adicionalmente, o sentido de anomalia de cada componente é verificado automaticamente sobre os benignos do treino antes da fusão: se o posto médio

dos benignos exceder $0{,}5$, o score é invertido ($\rho \leftarrow 1 - \rho$), corrigindo eventuais inversões introduzidas pela heterogeneidade entre domínios.

Os pesos w_{VAE} e w_{SVDD}, com $w_{\text{VAE}} + w_{\text{SVDD}} = 1$, são definidos de forma simétrica ($w_{\text{VAE}} = w_{\text{SVDD}} = 0{,}5$), uma vez que a fusão por postos percentuais já elimina a necessidade de compensação de escala entre os dois componentes.

4.3. Calibração do Limiar de Decisão por Protocolo

O grupo de protocolo de cada fluxo é identificado a partir de sua porta de destino, mapeada para categorias semânticas — Web, telemetria IoT, DNS e administração remota. Para cada grupo g, o limiar de classificação τ_g é definido como o percentil 95 da distribuição de scores combinados dos benignos do treino pertencentes àquele grupo:

$$\tau_g = \text{percentil}_{95}(\{s_{\text{ens}}(x) \mid x \in \mathcal{B},\ \text{grupo}(x) = g\}), \tag{6}$$

onde $\mathcal{B}$ denota o conjunto de fluxos benignos do treino. Essa escolha implica uma taxa de falsos positivos esperada de 5% sobre a distribuição de treino. Grupos de protocolo não observados durante o treino recorrem ao limiar global τ_{global}, calculado sobre todos os benignos independentemente do grupo, garantindo que o sistema opere sem interrupção diante de protocolos inéditos.

Durante a inferência, um fluxo x com grupo de protocolo g é classificado como anômalo se e somente se:

$$s_{\text{ens}}(x) > \tau_g. \tag{7}$$

Essa separação estrita entre calibração realizada exclusivamente sobre os benignos do treino e avaliação assegura que nenhuma informação dos conjuntos de teste influencie o comportamento do modelo, refletindo o cenário real de implantação em produção onde o limiar é fixado antes do sistema entrar em operação.

5. Avaliação Experimental

Para validar a resiliência dos modelos frente à mudança de domínios, adotou-se um esquema de rodízio inter-domínio. Diferentemente das abordagens tradicionais, que utilizam um único conjunto de dados para treinamento, validação e teste, a metodologia proposta expõe modelos previamente treinados a dados completamente não observados durante o treinamento. Esse procedimento permite avaliar de forma mais rigorosa o grau de generalização alcançado pelo PAMS, uma vez que o modelo treinado exclusivamente sobre os fluxos benignos do domínio de origem é aplicado diretamente aos domínios distintos sem qualquer retreinamento ou ajuste de parâmetros. O score de anomalia é produzido pela combinação do erro de reconstrução do VAE e da distância geométrica ao centro do SVDD, ambos calibrados sobre os benignos do treino, conforme descrito na Seção 4.

A avaliação também revela a complexidade do deslocamento que ocorre ao transitar de redes legadas para ambientes IoT. Arquiteturas tradicionais de detecção de anomalias frequentemente falham não pela ausência de capacidade representacional, mas pelo limiar *thresholds* de decisão estático. Quando um modelo é treinado com tráfego

IDS2017, ele estabelece uma restrição geométrica otimizada para conexões de longa duração e alta carga útil. No entanto, ao avaliar o tráfego telemétrico do *CICIoT2023*, o erro de reconstrução dos dados benignos é deslocado, gerando falsos positivos em massa. É nesse contexto de instabilidade geométrica que a ancoragem por protocolo do PAMS atua, permitindo que a hiperesfera do SVDD se readapte à semântica de cada serviço, ao invés de depender de um limiar global.

5.1. Análise Comparativa do Rodízio Inter-Domínio

A Tabela 7 apresenta o desempenho comparativo entre os modelos de base e o PAMS. Todos os algoritmos foram treinados exclusivamente no conjunto legado *CICIDS2017* e avaliados nos demais conjuntos, apresentando a acurácia e o F1-*Macro* como métricas.

Tabela 7. Desempenho de Generalização dos Modelos.

Modelo	2018			2023			DIAD		
	F1	Acc.	AUC-ROC	F1	Acc.	AUC-ROC	F1	Acc.	AUC-ROC
Random Forest	0,498	0,838	0,653	0,358	0,556	0,500	0,169	0,198	0,305
Decision Tree	0,805	0,981	0,519	0,505	0,926	0,499	0,206	0,224	0,492
G. Naive Bayes	0,382	0,394	0,441	0,330	0,402	0,532	0,082	0,083	—
MLP	0,661	0,859	0,845	0,511	0,598	0,723	0,236	0,241	0,604
XGBoost	0,598	0,857	0,902	0,358	0,557	0,539	0,189	0,212	0,803
SVDD	0,444	0,799	0,375	0,320	0,433	0,406	0,435	0,470	0,614
Isolation Forest	0,444	0,778	0,399	0,510	0,621	0,745	0,479	0,205	0,428
Elliptic Envelope	0,450	0,806	0,319	0,422	0,566	0,773	0,220	0,230	0,484
PAMS	**0,547**	**0,811**	**0,624**	**0,451**	**0,711**	**0,600**	**0,386**	**0,405**	**0,459**

Analisando os resultados da Tabela 7, os modelos supervisionados alcançam desempenho elevado no CICIDS2018, a DT atingindo F1-Macro de $0{,}805$, mas colapsam no CIC-IoT-DIAD, onde nenhum ultrapassa $0{,}237$, evidenciando sobreajuste às assinaturas do domínio de treino. Os métodos clássicos sem supervisão, IF e EE, apresentam maior estabilidade entre domínios, porém com F1-Macro inferior ao PAMS na maioria dos conjuntos; o IF supera o PAMS apenas no CIC-IoT-DIAD, enquanto o EE não ultrapassa $0{,}450$ em nenhum dos três conjuntos.

O PAMS é o único modelo que mantém desempenho consistente nos três domínios, com F1-Macro de $0{,}547$, $0{,}451$ e $0{,}386$ no CICIDS2018, CICIoT2023 e CIC-IoT-DIAD, respectivamente. A queda no CIC-IoT-DIAD reflete a severa divergência entre o tráfego corporativo legado do domínio de treino e a telemetria IoT do domínio de teste. No CICIoT2023, mesmo diante de protocolos como MQTT e mDNS ausentes no treino, o PAMS supera todos os modelos supervisionados e o EE, demonstrando que a modelagem da normalidade por grupo de protocolo confere maior robustez à mudança de domínio do que a memorização de assinaturas de ataque.

6. Conclusão

Este trabalho evidenciou que o desbalanceamento de classes em conjuntos de dados de IoT mascara falhas de generalização, onde métricas tradicionais como a acurácia ocultam o colapso de classificadores convencionais. A análise experimental sobre quatro conjuntos do CIC demonstrou nenhum modelo supervisionado ultrapassou F1-Macro de $0{,}237$ no CIC-IoT-DIAD.

A partir da análise do PAMS, demonstrou-se empiricamente que é possível alcançar generalização inter-domínio sem depender de exemplos de ataque durante o treinamento. O PAMS manteve desempenho consistente nos três domínios avaliados, com F1-Macro de $0{,}547$, $0{,}451$ e $0{,}386$ no CICIDS2018, CICIoT2023 e CIC-IoT-DIAD, respectivamente, superando todos os modelos supervisionados nos três conjuntos e os métodos clássicos sem supervisão em dois dos três, sendo superado apenas pelo *Isolation Forest* no CIC-IoT-DIAD. Observou-se que a ancoragem semântica por porta de destino, agrupando fluxos em categorias como Web, telemetria IoT, DNS e administração remota, favorece a abstração de padrões intrínsecos a cada protocolo em vez da memorização de assinaturas específicas do domínio de treino, propriedade particularmente relevante em ambientes IoT reais, onde novos dispositivos e protocolos são introduzidos continuamente.

Como trabalhos futuros, propõe-se a otimização automática dos hiperparâmetros do ensemble, incluindo os pesos de fusão e a dimensão do espaço latente, bem como a investigação de limiares adaptativos que se ajustem dinamicamente à deriva estatística do tráfego ao longo do tempo. Propõe-se também a integração de agentes baseados em Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) capazes de auxiliar na interpretação semântica dos grupos de protocolo e na classificação do tráfego em tempo real, ampliando a capacidade explicativa do sistema em cenários de produção.

7. Agradecimentos

Este trabalho é parcialmente financiado pelo CNPq (408255/2023-4 e 407304/2025-8), CAPES (88887.987121/2024-00 e 88881.987122/2024-01), COFECUB (Ma1074/25), FAPERJ e RNP e faz parte do INCT de Redes de Comunicação e Internet das Coisas Inteligentes (ICoNIoT), financiado pelo CNPq (405940/2022-0) e pela CAPES (88887.954253/2024-00).

Referências

Baena, E., Yang, H., Koutsonikolas, D., and Haque, I. (2025). A comprehensive survey on smart home iot fingerprinting: From detection to prevention and practical deployment.

Bezerra, V. H., da Costa, V. G. T., Martins, R. A., Junior, S. B., Miani, R. S., and Zarpelão, B. B. (2018). Providing iot host-based datasets for intrusion detection research. In *Anais do XVIII Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSEG)*, pages 15–28, Porto Alegre, RS, Brasil. SBC.

Bochie, K., Gonzalez, E., Giserman, L., Campista, M., and Costa, L. (2020). Detecção de ataques a redes iot usando técnicas de aprendizado de máquina e aprendizado profundo. In *Anais do XX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSEG)*, pages 257–270, Porto Alegre, RS, Brasil. SBC.

chethuhn (2021). Network intrusion dataset. `https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset`. Acesso em: 23-11-2025.

Chu, H.-C. and Lin, Y.-J. (2023). Improving the iot attack classification mechanism with data augmentation for generative adversarial networks. *Applied Sciences*, 13(23).

Churcher, A., Ullah, R., Ahmad, J., ur Rehman, S., Masood, F., Gogate, M., Alqahtani, F., Nour, B., and Buchanan, W. J. (2021). An experimental analysis of attack classification using machine learning in iot networks. *Sensors*, 21(2).

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2):161–190.

Ferrag, M. A., Friha, O., Hamouda, D., Maglaras, L., and Janicke, H. (2022). Edge-iiotset: A new comprehensive realistic cyber security dataset of iot and iiot applications: Centralized and federated learning.

Gao, X., Xie, D., Zhang, Y., Wang, Z., He, C., Yin, H., and Zhang, W. (2025). A comprehensive survey on imbalanced data learning. `https://arxiv.org/abs/2502.08960`. arXiv preprint arXiv:2502.08960.

Habibi Lashkari, A. (2018). Cicflowmeter-v4.0: A network traffic bi-flow generator and analyser for anomaly detection.

Hayashi, T., Cimr, D., Fujita, H., and Cimler, R. (2026). Critical review for one-class classification: Recent advances and reality behind them. *WIREs Data Mining and Knowledge Discovery*, 16(1).

Hozouri, A., Mirzaei, A., and Effatparvar, M. (2025). A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. *Discover Artificial Intelligence*, 5(1):314.

KimiNewt (2025). Pyshark: A python wrapper for tshark, allowing python packet analysis using wireshark dissectors. `https://github.com/KimiNewt/pyshark`.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koroniotis, N., Moustafa, N., Sitnikova, E., and Turnbull, B. (2018). Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. Ieee.

Lopes, D., Marotta, M., Ladeira, M., and Gondim, J. (2022). Detecção de botnets baseada na análise de fluxos de rede utilizando estatística inversa. In *Anais do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 182–195, Porto Alegre, RS, Brasil. SBC.

Manzini, S. C., Fernandes, B. d. S., de Souza, N. B. P., and Servare Junior, M. W. J. (2025). Revisão da literatura nacional sobre o uso da inteligência artificial no contexto do lean manufacturing. *South American Development Society Journal*, 11(32):331–350.

Miranda, B. and Braga, J. (2025). Iot e a gestão de recursos hídricos na agricultura de precisão: Uma revisão sistemática. In *Anais da XIII Escola Regional de Informática de Goiás*, pages 195–204, Porto Alegre, RS, Brasil. SBC.

Morshedi, R. and Matinkhah, S. M. (2025). A comprehensive review of deep learning techniques for anomaly detection in iot networks: Methods, challenges, and datasets. *Engineering Reports*, 7(9):e70415.

Moustafa, N. (2019a). The bot-iot dataset.

Moustafa, N. (2019b). Ton$_i otdatasets$.

Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., and Ghorbani, A. A. (2023). Ciciot2023: A real-time dataset and benchmark for large-scale attacks in iot environments. *Sensors*, 23(13).

Pekar, A. and Jozsa, R. (2024). Evaluating ml-based anomaly detection across datasets of varied integrity: A case study. *Computer Networks*, 251:110617.

Pereira, R., Costa, Y., and Jr., C. S. (2021). Lidando com o desbalanceamento em problemas de classificação hierárquica com reamostragem de dados. In *Anais Estendidos do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 13–18, Porto Alegre, RS, Brasil. SBC.

Rabbani, M., Gui, J., Nejati, F., Zhou, Z., Kaniyamattam, A., Mirani, M., Piya, G., Opushnyev, I., Lu, R., and Ghorbani, A. A. (2025). Device identification and anomaly detection in iot environments. *IEEE Internet of Things Journal*, 12(10):13625–13643.

Rehman, A. U., Lu, S., Bin Heyat, M. B., Iqbal, M. S., Parveen, S., Bin Hayat, M. A., Akhtar, F., Ashraf, M. A., Khan, O., Pomary, D., and Sawan, M. (2025). Internet of things in healthcare research: Trends, innovations, security considerations, challenges and future strategy. *International Journal of Intelligent Systems*, 2025(1):8546245.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In *International Conference on Machine Learning (ICML)*, pages 4393–4402. PMLR.

scikit-learn (2025). scikit-learn: Machine learning in python. `https://github.com/scikit-learn/scikit-learn`.

Sharafaldin, I., Lashkari, A. H., and Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *International Conference on Information Systems Security and Privacy*.

Silva, B. R., Silveira, R. J., Silva Neto, M. G. d., Cortez, P. C., and Gomes, D. G. (2021). A comparative analysis of undersampling techniques for network intrusion detection systems design. *Journal of Communication and Information Systems*, 36(1):31–43.

SolarMainframe (2021). Ids intrusion dataset (csv). `https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv`. Acesso em: 23-11-2025.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

Vitorino, J., Pinto, D., Maia, E., Amorim, I., and Praça, I. (2025). Revisiting network traffic analysis: Compatible network flows for ml models.