DISSERTAÇÃO DE MESTRADO PROFISSIONAL

DATASETS REALISTAS PARA NIDS BASEADOS EM FLUXOS: UMA PROPOSTA DE METODOLOGIA PARA CONSTRUÇÃO E AVALIAÇÃO DE DETECTORES SOB DERIVA TEMPORAL

Eduardo Rodrigues de Oliveira

Programa de Pós-Graduação Profissional em Engenharia Elétrica
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
FACULDADE DE TECNOLOGIA
UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO PROFISSIONAL

DATASETS REALISTAS PARA NIDS BASEADOS EM FLUXOS: UMA PROPOSTA DE METODOLOGIA PARA CONSTRUÇÃO E AVALIAÇÃO DE DETECTORES SOB DERIVA TEMPORAL

Eduardo Rodrigues de Oliveira

Dissertação de Mestrado Profissional submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia Elétrica

Banca Examinadora

Prof. João José Costa Gondim, Ph.D, FT/UnB _____
Orientador

Prof. Geraldo Pereira Rocha Filho, Ph.D, FT/UnB _____
Examinador interno

Prof. Iguatemi Eduardo da Fonseca, Ph.D, UFPB _____
Examinador externo

Prof. Fábio Lúcio L. de Mendonça, Ph.D, FT/UnB _____
Membro suplente

FICHA CATALOGRÁFICA

OLIVEIRA, EDUARDO. R.
DATASETS REALISTAS PARA NIDS BASEADOS EM FLUXOS: UMA PROPOSTA DE METODOLOGIA PARA CONSTRUÇÃO E AVALIAÇÃO DE DETECTORES SOB DERIVA TEMPORAL [Distrito Federal] 2026.
xvi, 58 p., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2026).
Dissertação de Mestrado Profissional - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. Detecção de Anomalias	2. NIDS
3. Datasets	4. Validação Temporal
I. ENE/FT/UnB	II. Título (série)

PUBLICAÇÃO: PPEE.MP.116

REFERÊNCIA BIBLIOGRÁFICA

OLIVEIRA, E. R. (2026). *DATASETS REALISTAS PARA NIDS BASEADOS EM FLUXOS: UMA PROPOSTA DE METODOLOGIA PARA CONSTRUÇÃO E AVALIAÇÃO DE DETECTORES SOB DERIVA TEMPORAL* . Dissertação de Mestrado Profissional, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 58 p.

Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

DEDICATÓRIA

Aos meus pais, que me deram o dom da vida e me ensinaram a não desistir dos meus sonhos; às minhas irmãs, pelo carinho, amizade e torcida constante.

À minha amada esposa, dedico cada linha deste trabalho. Guardamos conosco a certeza de que o inverno nunca falha em se tornar primavera. Obrigado pela nobreza de dividir comigo o peso das nossas batalhas e por me encorajar a seguir, mesmo quando o seu próprio caminhar era difícil. Juntos, cuidamos do nosso jardim, atravessamos o frio da adversidade e resgatamos o florescer da nossa primavera.

AGRADECIMENTOS

Antes de tudo, agradeço a Deus, que foi meu refúgio, minha força e minha luz nos momentos mais desafiadores desta caminhada.

Ao meu orientador, Prof. Dr. João José Costa Gondim, e ao meu co-orientador, Prof. Dr. Robson de Oliveira Albuquerque, expresso minha profunda reverência e gratidão. Agradeço pela orientação segura, paciência, competência técnica e pelos excelentes insights que foram cruciais para o desenvolvimento desta pesquisa.

Ao corpo docente do Programa de Pós-Graduação Profissional em Engenharia Elétrica (PPEE) da Universidade de Brasília, em especial, aos professores doutores William Ferreira Giozza, Georges Daniel Amvame Nze, Fábio Lúcio Lopes de Mendonça, Laerte Peotta de Melo, Geraldo Pereira Rocha Filho e Vinícius Pereira Rocha. Agradeço pelos ensinamentos transmitidos.

Aos colegas de classe com os quais pude interagir e trocar valorosas experiências.

Ao Sensei Kleber Bueno, agradeço por compreender minhas ausências nos treinos de Karate Uechi-Ryu durante o período em que precisei me concentrar nesta pesquisa.

E finalmente, aos amigos e irmãos do grupo "Turma da Onça", Carlos Sousa e Pedro Gontijo, obrigado pelo incentivo para ingressarmos e concluirmos juntos mais essa missão.

RESUMO

Muitos *datasets* usados na avaliação de NIDS baseados em aprendizado de máquina ainda combinam tráfego benigno pouco realista, rotulagem pouco auditável e partições estáticas que podem superestimar o desempenho dos detectores. Esta dissertação propõe e avalia, por meio de um estudo de caso controlado, uma metodologia reprodutível e auditável para construção de *datasets* realistas baseados em fluxos de rede, orientada à análise de detectores de anomalia sob deriva temporal. A metodologia integra coleta de tráfego benigno real em NetFlow/IPFIX, injeção controlada de ataques, extração e discretização de atributos, rotulagem assistida, separação temporal em T1, T2 e T3 e diagnóstico distribucional por PSI e Distância de Wasserstein. Como estudo de caso, aplicou-se o *Energy-based Flow Classifier* (EFC) em configuração *one-class*. Os resultados indicaram alto *recall* global e desempenho desigual entre *syn flood* e *aggressive scan*. Uma auditoria de proveniência revelou que o aumento agregado de falsos positivos em T3 decorreu majoritariamente de tráfego de retorno do ataque na fronteira de rotulagem, ao passo que a deriva T1–T3 entre períodos comerciais equivalentes, de magnitude moderada porém consistente, manteve-se evidenciada pelo diagnóstico distribucional sobre o tráfego benigno e pela taxa de falsos positivos depurada. A comparação com *Isolation Forest* e PCA mostrou que a ordenação dos modelos depende do regime de calibração, prospectivo ou retrospectivo. Os achados reforçam a utilidade da metodologia para revelar instabilidades temporais, diferenças por tipo de ataque e riscos do uso exclusivo de métricas agregadas, delimitando seu alcance ao contexto experimental avaliado.

ABSTRACT

Many datasets used to evaluate machine-learning-based NIDS still combine unrealistic benign traffic, weakly auditable labeling, and static partitions that may overestimate detector performance. This dissertation proposes and empirically assesses, through a controlled case study, a reproducible and auditable methodology for building realistic flow-based network datasets, oriented toward evaluating anomaly detectors under temporal drift. The methodology integrates real benign traffic collection using NetFlow/IPFIX, controlled attack injection, feature extraction and discretization, assisted labeling, temporal splitting into T1, T2, and T3, and distributional diagnosis using PSI and Wasserstein distance. As a case study, the *Energy-based Flow Classifier* (EFC) was evaluated in a *one-class* configuration. The results indicated high overall *recall* and uneven performance between *syn flood* and *aggressive scan*. A provenance audit revealed that the aggregate increase in false positives in T3 was mostly attributable to attack return traffic at the labeling boundary, whereas the T1–T3 drift between equivalent business-hour periods, moderate yet consistent in magnitude, remained evidenced by the distributional diagnosis over benign traffic and by the depurated false-positive rate. Comparisons with *Isolation Forest* and PCA showed that model ranking depends on the calibration regime, either prospective or retrospective. The findings highlight the usefulness of the methodology for exposing temporal instability, attack-wise performance differences, and risks associated with relying only on aggregate metrics, while delimiting its scope to the evaluated experimental context.

SUMÁRIO

1 INTRODUÇÃO 1
1.1 CONTEXTUALIZAÇÃO E MOTIVAÇÃO 1
1.2 PROBLEMA 3
1.3 OBJETIVOS 5
1.3.1 OBJETIVO GERAL 5
1.3.2 OBJETIVOS ESPECÍFICOS 5
1.4 JUSTIFICATIVA 6
1.5 CONTRIBUIÇÕES DESTE TRABALHO 7
1.6 ESTRUTURA DA DISSERTAÇÃO 7

2 FUNDAMENTAÇÃO TEÓRICA E ESTADO DA ARTE 9
2.1 APRENDIZADO DE MÁQUINA E DETECÇÃO DE ANOMALIAS EM NIDS BASEADOS EM FLUXO 9
2.1.1 ENERGY-BASED FLOW CLASSIFIER (EFC) 10
2.2 DERIVA TEMPORAL E DESALINHAMENTO DISTRIBUCIONAL 11
2.3 ESTADO DA ARTE EM *datasets* PARA NIDS 12
2.4 TRABALHOS CORRELATOS 13

3 METODOLOGIA 17
3.1 METODOLOGIA PROPOSTA 17
3.1.1 CONSTRUÇÃO E PREPARAÇÃO DO DATASET 17
3.1.2 AVALIAÇÃO PROSPECTIVA E DIAGNÓSTICO DISTRIBUCIONAL 22
3.2 ARQUITETURA 25

4 RESULTADOS EXPERIMENTAIS 28
4.1 CENÁRIO EXPERIMENTAL 28
4.2 CARACTERIZAÇÃO DO *dataset* GERADO 29
4.3 CONFIGURAÇÃO DO MODELO E PROTOCOLO TEMPORAL DE AVALIAÇÃO 31
4.4 RESULTADOS DE DETECÇÃO DO EFC 33
4.4.1 ANÁLISE ENERGÉTICA POR TIPO DE ATAQUE 34
4.5 SENSIBILIDADE DO *cutoff* 35
4.6 ANÁLISE DE ESTABILIDADE DISTRIBUCIONAL 36
4.7 PERFIL DOS FALSOS POSITIVOS 37
4.7.1 PROVENIÊNCIA DOS FALSOS POSITIVOS E TRÁFEGO DE RETORNO DO ATAQUE 38
4.8 INTERVALOS DE CONFIANÇA POR BOOTSTRAP 39
4.9 COMPARAÇÃO COM MODELOS *one-class* ALTERNATIVOS 40
4.9.1 *Isolation Forest* 40
4.9.2 PRINCIPAL COMPONENT ANALYSIS (PCA) 41

4.9.3 Tentativa controlada com One-Class SVM 42
4.10 Síntese comparativa dos modelos 42
4.11 Artefatos consolidados 43
4.12 Síntese dos resultados 44

5 Discussão 45
5.1 Protocolo temporal como eixo de validade experimental 45
5.2 Sensibilidade do *cutoff* e implicações operacionais 46
5.3 Desempenho por tipo de ataque 46
5.4 Perfil dos falsos positivos e interpretação operacional 47
5.5 Comparação com modelos *one-class* alternativos 47
5.6 Design da coleta e implicações para *datasets* realistas 48
5.7 Limitações e considerações operacionais 48

6 Considerações Finais 50
6.1 Trabalhos Futuros 52

REFERÊNCIAS BIBLIOGRÁFICAS 54

LISTA DE FIGURAS

1.1 Incidentes notificados ao CERT.br por categoria no período de janeiro a dezembro de 2025. Adaptado de (1). 2
1.2 Vetores iniciais de infecção identificados no *ENISA Threat Landscape 2025*. 2

3.1 Funcionamento geral do trabalho desde a observação da rede até a interpretação dos resultados. 17
3.2 Arquitetura operacional da captura do tráfego e da execução controlada dos ataques. 18
3.3 Preparação dos atributos e produção da representação utilizada pelos detectores. 20
3.4 Rotulagem assistida, controle de qualidade e organização temporal dos dados. 21
3.5 Protocolo de validação temporal. T1 define a normalidade, T2 seleciona o ponto operacional com informação historicamente disponível e T3 permanece fora da calibração para representar dados futuros. 23
3.6 Topologia do experimento proposto 26

4.1 Matrizes de confusão do EFC nos recortes temporais T2 e T3, no ponto de melhor F1 de cada janela. 34
4.2 Distribuição das energias atribuídas pelo EFC aos fluxos benignos e maliciosos em T2 e T3. 34
4.3 Distribuição de energia por classe (benigno, `syn_flood` e `aggressive_scan`) em T2 e T3, com o limiar de decisão (*cutoff*). 35
4.4 Relação entre *recall* e taxa de falsos positivos na análise de sensibilidade do *cutoff*. 36
4.5 Distribuição dos falsos positivos por faixa horária em T2 e T3. 38

LISTA DE TABELAS

2.1 Síntese comparativa de trabalhos correlatos segundo as dimensões de objetivo, abordagem e limitações. 15
2.2 Comparação qualitativa com trabalhos correlatos (2022–2026). 16

3.1 Detalhamento dos serviços conteinerizados. 25

4.1 Recortes temporais utilizados no experimento. 29
4.2 Distribuição de classes por recorte temporal. 29
4.3 Plano de ataques controlados executados no experimento. 30
4.4 Atributos utilizados pelo EFC no experimento. 32
4.5 Limiar de referência e pontos de melhor F1 no pacote reprodutível. 32
4.6 Desempenho prospectivo estrito do EFC. 33
4.7 Desempenho global do EFC nos pontos de melhor F1 por janela. 33
4.8 Desempenho do EFC por tipo de ataque. 33
4.9 Estatísticas de energia por classe e fração de fluxos abaixo do *cutoff*. 35
4.10 Principais atributos com maior deriva distribucional segundo o PSI. 36
4.11 Principais concentrações dos falsos positivos por dimensão de tráfego. 37
4.12 Proveniência dos falsos positivos do EFC no ponto de melhor F1 por janela. 39
4.13 Taxa de falsos positivos bruta e após remoção do tráfego do contexto de ataque. 39
4.14 Métricas globais do EFC com intervalos de confiança de 95% por *bootstrap* (regime retrospectivo em T3). 40
4.15 Comparação entre EFC e *Isolation Forest* no ponto de melhor F1 por janela (regime retrospectivo em T3). 40
4.16 Comparação entre EFC, *Isolation Forest* e PCA no ponto de melhor F1 por janela (regime retrospectivo em T3). 41
4.17 Comparação dos modelos em avaliação prospectiva estrita em T3 (*cutoff* de T2 aplicado a T3). 42
4.18 Recall por tipo de ataque e F1-score global por modelo (regime retrospectivo em T3). 43

LISTA DE SIGLAS

ACK	Acknowledgment
C2	Command and Control
CERT.br	Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil
CIDR	Classless Inter-Domain Routing
CPU	Central Processing Unit
CSIRT	Computer Security Incident Response Team
CSV	Comma-Separated Values
DARPA	Defense Advanced Research Projects Agency
DDoS	Distributed Denial of Service
DoS	Denial of Service
EFC	Energy-based Flow Classifier
ENISA	European Union Agency for Cybersecurity
F1-score	Medida harmônica entre precisão e *recall*
FIN	Finish
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FT	Faculdade de Tecnologia
GB	Gigabyte
IC95%	Intervalo de Confiança de 95%
IDS	Intrusion Detection System
IIoT	Industrial Internet of Things
IoT	Internet of Things
IP	Internet Protocol
IPFIX	Internet Protocol Flow Information Export
KDD	Knowledge Discovery and Data Mining
Mbps	Megabits por segundo
MITRE ATT&CK	Adversarial Tactics, Techniques, and Common Knowledge
MQTT	Message Queuing Telemetry Transport
NIDS	Network Intrusion Detection System
P95	Percentil 95
PCA	Principal Component Analysis
PCAP	Packet Capture
PPEE	Programa de Pós-Graduação Profissional em Engenharia Elétrica
PSI	Population Stability Index
PVE	Proxmox Virtual Environment
RAM	Random Access Memory

ROC	Receiver Operating Characteristic
RST	Reset
SDN	Software-Defined Networking
SHA-256	Secure Hash Algorithm 256-bit
SSH	Secure Shell
SVM	Support Vector Machine
SYN	Synchronize
TCP	Transmission Control Protocol
TLS	Transport Layer Security
TN	True Negative
TP	True Positive
TTPs	Tactics, Techniques and Procedures
UDP	User Datagram Protocol
UFPB	Universidade Federal da Paraíba
UnB	Universidade de Brasília
vCPU	Virtual Central Processing Unit
WCNPS	Workshop on Communication Networks and Power Systems

1 INTRODUÇÃO

Este capítulo estabelece o contexto da pesquisa ao situar a detecção de intrusão em rede no cenário contemporâneo de segurança cibernética. A partir desse enquadramento, explicita o problema investigado e os objetivos que orientam o estudo. Em seguida, apresenta a justificativa e as contribuições da dissertação, com ênfase na proposição de uma metodologia reprodutível e auditável para construção e validação temporal de *datasets* de fluxo orientados à detecção de anomalias. Ao final, descreve a organização dos capítulos subsequentes.

1.1 CONTEXTUALIZAÇÃO E MOTIVAÇÃO

Sistemas de Detecção de Intrusão em Rede (*Network Intrusion Detection Systems* – NIDS) ocupam posição estratégica na defesa de infraestruturas digitais contemporâneas. Em ambientes corporativos, acadêmicos, governamentais e industriais, tais sistemas constituem uma camada essencial de observação e resposta, complementando mecanismos preventivos tradicionais, como *firewalls*, listas de controle de acesso e ferramentas antimalware. A sua relevância cresce à medida que as redes se tornam mais heterogêneas, distribuídas e dependentes de aplicações críticas, ao mesmo tempo em que a superfície de ataque se expande com a adoção de computação em nuvem, dispositivos IoT/IIoT, serviços expostos à Internet e tráfego cada vez mais criptografado (2, 3, 4).

Esse cenário não é apenas teórico. No Brasil, o CERT.br mantém séries históricas de incidentes notificados voluntariamente por CSIRTs, administradores de redes e usuários, disponibilizando totais anuais desde 1999 e, para os anos mais recentes, detalhamento mensal, categorial e por origem de eventos (1). Os dados de janeiro a dezembro de 2025, sintetizados na Figura 1.1, mostram um predomínio expressivo de ocorrências classificadas como *Scan*, com 393.919 notificações, correspondendo à ampla maioria dos registros do período. Em contraste, categorias como *DoS* (26.418), *Fraude* (24.030) e *Web* (19.965) aparecem em patamar bem inferior, ao passo que *Invasão* (3.729) e *Outros* (2.824) possuem participação residual.

Esse perfil sugere que o padrão dos incidentes notificados permanece fortemente marcado por atividades de reconhecimento automatizado, tentativa de enumeração de serviços e sondagens em larga escala, as quais frequentemente antecedem ataques mais destrutivos ou mais direcionados. Para estudos de NIDS baseados em fluxo, tal evidência é particularmente importante, pois indica que mecanismos de detecção precisam ser sensíveis não apenas a intrusões confirmadas, mas também a desvios comportamentais associados às fases iniciais de reconhecimento e preparação do ataque.

No contexto europeu, o *ENISA Threat Landscape 2025* indica que o acesso inicial às campanhas maliciosas permaneceu fortemente concentrado em *phishing*, responsável por cerca de 60% dos vetores identificados, seguido por exploração de vulnerabilidades (21,3%), botnets (9,9%), aplicações maliciosas (8%) e ameaça interna (0,8%) (5). Embora essas categorias representem formas distintas de comprometimento

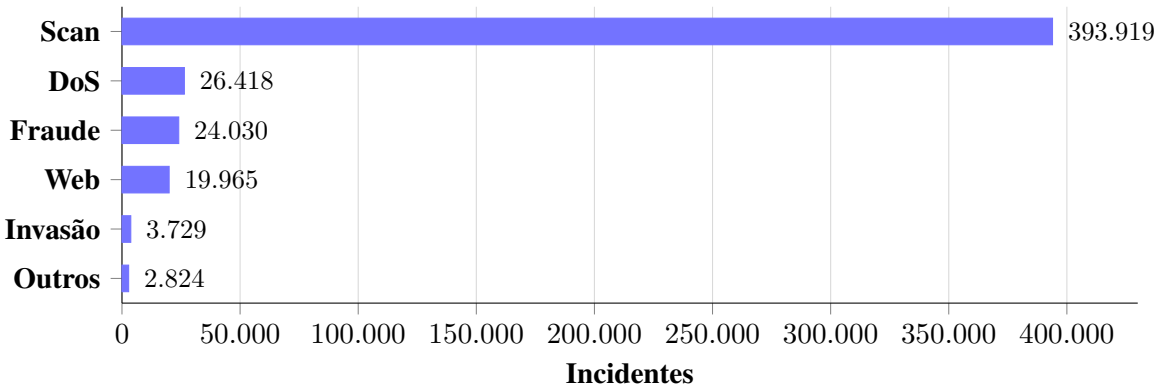


Figura 1.1: Incidentes notificados ao CERT.br por categoria no período de janeiro a dezembro de 2025. Adaptado de (1).

inicial, a centralidade do *phishing* sugere sua função recorrente como ponto de partida para cadeias mais amplas de intrusão. Em conjunto, esses dados reforçam a necessidade de mecanismos de detecção aptos a identificar padrões heterogêneos de comprometimento em nível de fluxo.

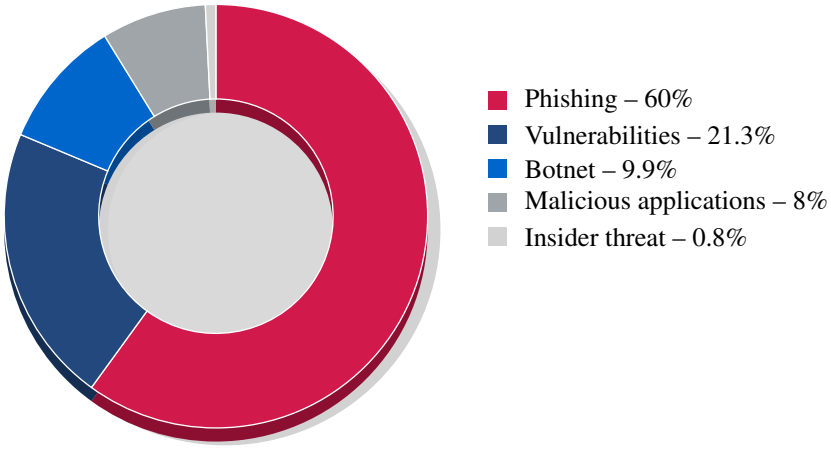


Figura 1.2: Vetores iniciais de infecção identificados no *ENISA Threat Landscape 2025*.

A distribuição dos vetores de acesso inicial apresentada na Figura 1.2 evidencia que as campanhas maliciosas podem explorar diferentes meios de comprometimento. À medida que técnicas, ferramentas e infraestruturas utilizadas pelos adversários se modificam, os padrões de tráfego associados a essas atividades também podem se alterar. Essa dinâmica reforça a necessidade de *datasets* capazes de acompanhar mudanças no comportamento da rede e de avaliar detectores em períodos temporalmente distintos. Sem essa preocupação, um conjunto de dados pode representar apenas condições específicas de sua coleta e produzir estimativas excessivamente otimistas sobre o desempenho futuro de um NIDS.

Historicamente, parte importante dos NIDS foi construída sob o paradigma de detecção por assinaturas. Essa abordagem continua relevante para reconhecer padrões conhecidos com baixa ambiguidade, sobretudo quando há inteligência prévia consolidada sobre indicadores de comprometimento, regras de exploração e artefatos observáveis. Ainda assim, a dependência de assinaturas impõe limitações estruturais. A base de regras exige atualização contínua, o que acarreta custo operacional e defasagem frente ao surgimento de novas campanhas. Além disso, pequenas variações em cadeias de ataque, protocolos, infraestrutura de comando e controle ou técnicas de ofuscação podem reduzir a eficácia de detectores baseados estritamente

em padrões estáticos. Soma-se a isso o avanço da criptografia fim a fim, que restringe o acesso ao conteúdo do *payload* e desloca a atenção para atributos laterais da comunicação e para o comportamento estatístico dos fluxos (2, 3, 6).

Nesse contexto, abordagens orientadas por dados, especialmente aquelas baseadas em aprendizado de máquina, passaram a ocupar papel central na literatura de NIDS. O uso de classificadores supervisionados, métodos não supervisionados, aprendizado profundo e estratégias híbridas ampliou a capacidade de identificar tráfego anômalo ou malicioso a partir de regularidades extraídas do comportamento observado na rede (2, 7, 8). Em vez de depender exclusivamente de regras previamente especificadas, essas abordagens procuram inferir, a partir dos dados, relações discriminativas entre tráfego benigno e tráfego malicioso. Esse movimento foi impulsionado tanto pela disponibilidade crescente de capacidade computacional quanto pela consolidação de *benchmarks* públicos, que viabilizaram a comparação sistemática entre algoritmos.

Entre as diferentes formas de representação de tráfego, os NIDS baseados em fluxo se destacam como alternativa pragmática e escalável. Em vez de operar sobre cada pacote individualmente, tais sistemas utilizam registros agregados de comunicação, como NetFlow e IPFIX, preservando atributos como endereços de origem e destino, portas, protocolo, quantidade de bytes, quantidade de pacotes, direção e duração da sessão. Essa escolha oferece vantagens relevantes, como menor custo de armazenamento, viabilidade de monitoramento em larga escala, preservação de parte importante da semântica das comunicações e aplicabilidade mesmo quando o conteúdo da carga útil está criptografado (4, 9, 10). Em redes modernas, essa característica assume particular importância, uma vez que a inspeção profunda de pacotes em produção pode ser tecnicamente inviável, juridicamente sensível ou operacionalmente proibitiva.

No âmbito dos NIDS baseados em fluxo, é relevante distinguir abordagens supervisionadas, treinadas sobre exemplos de ambas as classes, de abordagens *one-class*, que modelam exclusivamente o perfil benigno e sinalizam como anômalo qualquer desvio significativo em relação a esse perfil. Esses detectores de anomalia são particularmente adequados a ambientes em que amostras maliciosas rotuladas são escassas ou rapidamente desatualizadas. Em contrapartida, a pureza da partição de treinamento torna-se requisito central, pois qualquer contaminação por tráfego malicioso no conjunto de treino compromete a calibração do detector, impondo restrições sobre o processo de construção do *dataset* que não se colocam no paradigma supervisionado convencional.

1.2 PROBLEMA

A literatura mostra que o sucesso experimental de modelos de NIDS baseados em aprendizado de máquina não garante, por si só, robustez em produção. Um dos principais motivos é o desalinhamento distribucional entre os dados usados em laboratório e o comportamento efetivo de redes operacionais. Em muitos estudos, treino e teste são realizados sobre subconjuntos derivados do mesmo *dataset*, frequentemente coletados no mesmo ambiente, no mesmo intervalo temporal e sob a mesma lógica de geração ou rotulagem. Embora esse procedimento facilite comparações, ele frequentemente preserva uma proximidade estatística entre amostras de treino e teste que raramente se verifica em implantação real (3, 7, 11).

Em consequência, métricas elevadas obtidas em *benchmarks* fechados podem superestimar a capacidade real de generalização do detector.

Essa limitação decorre da suposição de que os dados utilizados no treinamento permanecem representativos dos dados observados posteriormente em operação. Em NIDS baseados em fluxos, essa suposição é pressionada pela variabilidade natural do tráfego benigno e pela evolução contínua das atividades maliciosas. Redes reais incorporam dispositivos, aplicações, hábitos de uso, janelas de atividade e rotas de comunicação que se alteram ao longo do tempo. Os adversários também modificam técnicas, exploram vulnerabilidades emergentes e ajustam a intensidade e a forma de suas campanhas. Essas mudanças podem produzir deriva temporal e reduzir a aderência entre o perfil aprendido pelo detector e os fluxos observados em períodos futuros (3, 12).

Quando a deriva temporal não é explicitamente considerada, aumenta-se o risco de produzir avaliações excessivamente otimistas. Estudos recentes mostram que *datasets* amplamente utilizados, como UNSW-NB15, CIC-IDS2017 e TON_IoT, embora valiosos para fins de pesquisa, apresentam limitações relevantes, entre as quais se destacam a artificialidade dos cenários, a documentação incompleta, inconsistências de rotulagem, desequilíbrios severos entre classes e divergências estatísticas em relação ao tráfego de produção (3, 11, 13, 14, 15, 16). Nesse contexto, a dificuldade central já não reside apenas na disponibilidade de dados, mas na ausência de processos metodológicos consistentes capazes de produzir conjuntos de dados simultaneamente representativos, auditáveis e reutilizáveis. Por essa razão, a adoção de um *dataset* público não elimina a necessidade de examinar seu processo de construção, as hipóteses que o orientaram, as ferramentas empregadas, os critérios de rotulagem adotados e o grau de aderência de suas propriedades ao domínio de aplicação pretendido.

A rotulagem constitui uma dimensão particularmente crítica no domínio tratado nesta pesquisa. Em redes reais, a determinação de um *ground truth* confiável apresenta dificuldades inerentes ao próprio contexto de aplicação. Mesmo em situações nas quais o ataque é previamente planejado e executado sob condições controladas, a vinculação exata entre a janela temporal da ação maliciosa, os fluxos efetivamente observados e os efeitos laterais produzidos nem sempre se estabelece de maneira inequívoca. Em cenários não controlados, essa complexidade se acentua, uma vez que atividades anômalas podem ocorrer simultaneamente a comportamentos benignos raros, alterações legítimas de configuração, padrões incomuns de uso e artefatos introduzidos pelo processo de coleta. Revisões recentes assinalam que muitos *datasets* têm sido disponibilizados com rótulos pouco transparentes ou baseados em pressupostos fortes, nem sempre explicitados de modo suficiente para permitir sua verificação por parte do leitor ou a reprodução do experimento (3, 4, 17, 18). Diante desse quadro, propostas metodológicas mais recentes têm enfatizado a adoção de estratégias de rotulagem assistida, rastreável e auditável, sustentadas pela articulação entre planos de ataque, registros de ferramentas auxiliares, detectores complementares e, quando necessário, procedimentos de inspeção dirigida (18, 19, 20).

As limitações descritas até aqui indicam que a dificuldade central da área não reside apenas no desempenho isolado dos algoritmos de detecção, mas também na forma como os dados são produzidos, rotulados e avaliados. Quando o processo de construção do *dataset* não explicita, de maneira reprodutível e auditável, os critérios de coleta, rotulagem e separação temporal, os resultados obtidos tendem a superestimar a capacidade de generalização do detector em ambientes reais.

Embora estudos recentes tenham avançado na construção de *datasets* mais realistas, ainda não se identifica uma solução que integre coleta de tráfego benigno real, injeção controlada de ataques, rotulagem rastreável e avaliação temporal de detectores de anomalia baseados em fluxos. A lacuna está na articulação desses elementos em um processo que possa ser reproduzido, auditado e empregado para examinar o comportamento do detector diante de dados futuros. Esta dissertação aborda essa lacuna por meio de uma metodologia que organiza a construção do *dataset* e a avaliação temporal como partes do mesmo desenho experimental (3, 4, 18, 19).

A principal contribuição desta dissertação é a proposição de uma metodologia reprodutível e auditável para construir e avaliar *datasets* de NIDS baseados em fluxos sob deriva temporal. A metodologia integra coleta de tráfego benigno real, injeção controlada de ataques, extração e discretização de atributos, rotulagem assistida, separação cronológica em T1, T2 e T3 e diagnóstico distribucional por PSI e Distância de Wasserstein. O EFC é empregado como estudo de caso para demonstrar a aplicabilidade da metodologia e não como elemento do qual ela dependa.

1.3 OBJETIVOS

1.3.1 Objetivo geral

Propor e avaliar empiricamente, por meio de um estudo de caso com tráfego benigno real e ataques controlados, uma metodologia para a construção de *datasets* realistas baseados em fluxos de rede, orientada à avaliação de detectores de anomalia sob deriva temporal.

1.3.2 Objetivos específicos

- Desenvolver um *pipeline* que integre captura contínua de tráfego benigno real com injeção controlada de ataques;
- Implementar um mecanismo de rotulagem assistida e rastreável, correlacionando planos de ataque com registros de detectores auxiliares e artefatos de coleta;
- Estabelecer um protocolo de validação baseado em divisões temporais prospectivas, de modo a quantificar o impacto da deriva temporal sobre o desempenho do detector;
- Demonstrar a aplicação da metodologia proposta por meio de um estudo de caso com o *Energy-based Flow Classifier*, avaliando sua capacidade de revelar efeitos de deriva temporal, instabilidade do *cutoff* e variação de desempenho por tipo de ataque;
- Analisar, em caráter exploratório, o impacto do ajuste do ponto operacional do detector sobre a relação entre recall, precisão e taxa de falsos positivos em cenários prospectivos sujeitos a deslocamento temporal.

1.4 JUSTIFICATIVA

A justificativa desta pesquisa decorre da necessidade de enfrentar uma limitação recorrente na literatura de NIDS baseados em aprendizado de máquina, relacionada menos à ausência de algoritmos promissores do que à fragilidade metodológica dos processos de construção e avaliação dos dados. Em muitos casos, resultados expressivos são obtidos sobre *benchmarks* cuja composição não reflete adequadamente as propriedades estatísticas do tráfego benigno real, cujos rótulos não são plenamente auditáveis e cujos procedimentos de separação entre treino e teste não incorporam, de forma explícita, a dimensão temporal. Nessas condições, métricas favoráveis podem não traduzir robustez efetiva quando o modelo é submetido a cenários operacionais sujeitos a deriva temporal.

Essa constatação impõe a necessidade de deslocar o foco analítico do algoritmo isolado para o processo de construção e validação dos dados. A questão relevante, portanto, não se limita à identificação do "melhor classificador", mas envolve a formulação de uma metodologia capaz de integrar coleta contínua de tráfego benigno real, injeção controlada de ataques, extração consistente de atributos de fluxo, rotulagem assistida e validação temporal prospectiva. Tal orientação encontra respaldo em iniciativas recentes voltadas ao aumento do realismo experimental, como HIKARI-2021, que enfatiza a incorporação de tráfego criptografado e a transparência do processo de construção do conjunto de dados, e ID2T, que busca injetar ataques preservando propriedades estatísticas do tráfego de fundo (6, 19). Na mesma direção, também se observam esforços voltados à avaliação em janelas temporais distintas, como MAWIFlow, que explicita a degradação de desempenho ao longo do tempo em cenários mais realistas (3, 21).

É nesse quadro metodológico que se insere a presente pesquisa. Em vez de tomar o classificador como finalidade autônoma, adota-se o *Energy-based Flow Classifier* (EFC) como estudo de caso para operacionalizar a metodologia de construção de *datasets* de fluxo proposta nesta dissertação. O EFC se mostra particularmente apropriado para esse propósito porque opera em regime *one-class*, modelando exclusivamente o tráfego benigno com base em estatística inversa inspirada no modelo de Potts e classificando novos fluxos a partir de um valor de energia associado ao grau de conformidade com o perfil benigno aprendido (2). Além de reduzir a dependência de grandes volumes de amostras maliciosas rotuladas na etapa de treinamento, o método oferece interpretabilidade e apresenta menor complexidade estrutural do que parte das abordagens profundas empregadas em NIDS.

A relevância da proposta também se manifesta no plano metodológico e experimental. Ao organizar um *pipeline* ponta a ponta para coleta, ataque controlado, extração, discretização, rotulagem assistida e avaliação temporal, busca-se contribuir para o fortalecimento da validade experimental em NIDS e para a produção de artefatos reutilizáveis em estudos futuros. Desse modo, a justificativa do trabalho não se apoia apenas na necessidade de avaliar um detector específico, mas na conveniência científica de estabelecer um processo mais transparente, replicável e aderente às condições observadas em redes reais.

1.5 CONTRIBUIÇÕES DESTE TRABALHO

A pesquisa desenvolvida durante o programa de mestrado resultou na produção de um artigo (22) científico que consolida o núcleo metodológico e experimental desta dissertação:

- **Proposta metodológica e avaliação experimental:** E. R. de Oliveira, R. O. Albuquerque and J. J. C. Gondim, "Methodology for Building Realistic Network Intrusion Detection Datasets: A Case Study with the Energy-Based Flow Classifier,"2025 Workshop on Communication Networks and Power Systems (WCNPS), Brasilia, Brazil, 2025, pp. 1-6, doi: 10.1109/WCNPS69127.2025.11295886.

A contribuição central desse artigo consiste na proposição de um *pipeline* para construção de *datasets* de fluxo, integrando coleta contínua de tráfego benigno real, ataques controlados, rotulagem assistida e separações temporais prospectivas. Na avaliação realizada com o *Energy-based Flow Classifier*, observou-se aumento da taxa agregada de falsos positivos em janelas futuras, mesmo com *recall* elevado, o que reforça a necessidade de validações temporalmente não sobrepostas. A análise aprofundada conduzida nesta dissertação (Seção 4.7.1) refina essa leitura, ao mostrar que parte expressiva desse aumento decorre de tráfego de retorno do ataque na fronteira de rotulagem, enquanto a deriva do tráfego benigno é sustentada pelo diagnóstico distribucional.

No âmbito desta dissertação, as principais contribuições podem ser sintetizadas em quatro pontos. Primeiro, a formulação de uma metodologia auditável para construção e validação temporal de *datasets* de fluxo aplicados à detecção de intrusão. Segundo, a organização de um *pipeline* compatível com NetFlow/IPFIX, capaz de integrar coleta, extração, discretização e rotulagem assistida com rastreabilidade do processo. Terceiro, a evidência empírica de que o tráfego benigno se desloca ao longo do tempo de forma independente do efeito dos ataques, deslocamento caracterizado pelo diagnóstico distribucional (PSI e Distância de Wasserstein) e por uma elevação modesta da taxa de falsos positivos benigna sob limiar de referência fixo, verificável em horizonte temporal de curto prazo e sem alteração nos tipos de ataque avaliados, com os limites de generalização inerentes a uma coleta de 83 horas em rede única. Por fim, a análise de sensibilidade do *cutoff* e a avaliação estratificada por tipo de ataque indicam que, nos pontos analisados, reduções de falsos positivos tendem a vir acompanhadas de perda de *recall*, e que a não detecção de certas classes pelo EFC não foi revertida pela recalibração do limiar. Essa distinção ajuda a separar limitações de ponto operacional de limitações associadas à representação do modelo ou aos atributos de fluxo disponíveis.

1.6 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação está organizada em seis capítulos, além das referências bibliográficas e dos apêndices. A estrutura foi definida de modo a conduzir a discussão desde a contextualização do problema e a fundamentação teórica até a proposição metodológica, a arquitetura experimental, a avaliação dos resultados, a discussão dos achados e as considerações finais do estudo, conforme descrito a seguir:

- **Capítulo 1: Introdução** apresenta a contextualização do tema, a motivação da pesquisa, o problema investigado, os objetivos, a justificativa e as contribuições esperadas, além de descrever a organização da dissertação.

- **Capítulo 2: Fundamentação teórica e estado da arte** estabelece os conceitos necessários à compreensão do estudo, abordando detecção de intrusão baseada em fluxos, estratégias *one-class*, fundamentos do *Energy-based Flow Classifier* (EFC), deriva temporal, desalinhamento distribucional e limitações dos principais *datasets* utilizados em NIDS.

- **Capítulo 3: Metodologia** descreve a abordagem adotada para construção e validação temporal dos *datasets*, detalhando o *pipeline* proposto, os estágios de captura de tráfego e simulação de ataques, a extração e discretização de atributos, o processo de rotulagem assistida, a organização dos recortes temporais e o protocolo de avaliação prospectiva com diagnóstico distribucional. Apresenta ainda a arquitetura experimental utilizada para operacionalizar a metodologia proposta, incluindo a organização dos serviços conteinerizados, o ambiente de coleta, o host atacante, o coletor NetFlow/IPFIX, o processamento dos fluxos e os artefatos produzidos pelo experimento.

- **Capítulo 4: Resultados Experimentais** apresenta o cenário experimental, o *dataset* gerado, a configuração do EFC, os resultados obtidos nos recortes temporais T2 e T3, a análise energética por tipo de ataque, a análise de sensibilidade do *cutoff*, o diagnóstico distribucional por PSI e Distância de Wasserstein, o perfil dos falsos positivos, os intervalos de confiança por *bootstrap* e os comparativos com modelos alternativos, como *Isolation Forest*, PCA e One-Class SVM.

- **Capítulo 5: Discussão** analisa os resultados à luz do problema de pesquisa e da literatura correlata, discutindo as implicações metodológicas e operacionais da proposta, o impacto da deriva temporal sobre detectores *one-class*, a influência do ponto operacional, a diferença de desempenho entre tipos de ataque, a concentração dos falsos positivos, o risco de falsos negativos em ataques mais discretos e as limitações do estudo.

- **Capítulo 6: Considerações Finais** sintetiza os principais achados da pesquisa, retoma o objetivo geral e os objetivos específicos, apresenta as contribuições alcançadas e indica possibilidades de continuidade, com destaque para a ampliação da diversidade temporal e comportamental do *dataset*, o enriquecimento dos atributos, o aprimoramento da rotulagem assistida, a exploração de versões multiclasse e *open-set* do EFC e a evolução do *pipeline* para contextos operacionais contínuos.

2 FUNDAMENTAÇÃO TEÓRICA E ESTADO DA ARTE

Este capítulo apresenta a base conceitual e o panorama do estado da arte que sustentam a pesquisa. Inicialmente, são discutidos os fundamentos da detecção de intrusão baseada em fluxo, as estratégias *one-class*, os princípios do *Energy-based Flow Classifier* e o problema da deriva temporal em NIDS baseados em aprendizado de máquina. Em seguida, são examinados criticamente os principais *datasets* e trabalhos correlatos da literatura, com o objetivo de delimitar as contribuições mais próximas da proposta desta dissertação e explicitar o ponto em que ela se diferencia.

2.1 APRENDIZADO DE MÁQUINA E DETECÇÃO DE ANOMALIAS EM NIDS BASEADOS EM FLUXO

NIDS podem ser classificados, em linhas gerais, em sistemas baseados em assinaturas e sistemas baseados em anomalias. Os primeiros dependem de padrões previamente conhecidos de ataque; os segundos modelam o comportamento esperado do tráfego e sinalizam desvios significativos em relação a esse padrão (23, 3). Em redes contemporâneas, a detecção baseada em anomalia é especialmente atraente porque reduz a dependência de bases de assinaturas atualizadas e preserva a possibilidade de identificar ameaças não observadas durante o treinamento. Em contrapartida, tende a enfrentar o problema clássico de *false alarms*, sobretudo quando o tráfego benigno é diverso e evolui ao longo do tempo (24, 25).

No âmbito do monitoramento, a literatura distingue a análise em nível de pacote e a análise em nível de fluxo. A inspeção profunda de pacotes oferece elevada granularidade, mas implica alto custo computacional, maior sensibilidade a políticas de privacidade e baixa aderência a ambientes com tráfego cifrado. Em oposição, a análise baseada em fluxo opera sobre registros agregados (*source/destination* IP, portas, protocolo, bytes, pacotes, duração, direção, entre outros), permitindo observabilidade suficiente para classificar comunicações em larga escala (2, 9, 10). Por essa razão, ela é particularmente apropriada para NIDS operacionais, para monitoramento em redes SDN e para cenários IIoT/IoT, nos quais dispositivos e enlaces possuem restrições relevantes (8, 26, 27).

A literatura recente também evidencia a relevância de estratégias *one-class* e semi-supervisionadas em cenários nos quais a obtenção de tráfego malicioso rotulado é cara, incompleta ou inviável (25, 28). Autoencoders, *One-Class SVM*, *Isolation Forest* e *Local Outlier Factor* figuram entre os exemplos mais recorrentes (12, 25, 29). Embora dispensem, em maior ou menor grau, conjuntos extensos de ataques rotulados, essas abordagens podem sofrer sob mudanças de domínio, reconstruir amostras anômalas com erro reduzido ou produzir taxas elevadas de falso positivo quando o perfil benigno não é caracterizado de forma adequada (3, 12).

2.1.1 Energy-based Flow Classifier (EFC)

O EFC foi proposto por (2) como um classificador de fluxos baseado em anomalia, construído a partir de estatística inversa e inspirado no modelo inverso de Potts. A origem desse formalismo remonta à física estatística, em que o modelo de Potts descreve spins interagentes em uma rede cristalina, por meio de campos locais e acoplamentos entre pares de nós. Na formulação do EFC, essa intuição é reutilizada no domínio de tráfego de rede, em que cada fluxo é representado como uma configuração discreta de atributos, e a inferência busca reconstruir, a partir de amostras benignas observadas, o modelo estatístico da normalidade. Em sua formulação original, o algoritmo infere esse modelo com base em exemplos benignos previamente rotulados e, a partir dele, passa a avaliar novos fluxos segundo uma medida de energia associada ao seu grau de compatibilidade com a distribuição aprendida no treinamento (2, 30).

Nessa perspectiva, o EFC se insere no contexto de classificação *one-class* e de detecção baseada em anomalia, uma vez que sua etapa de aprendizagem é fundamentada exclusivamente na classe benigna. Diferentemente de classificadores supervisionados convencionais, que exigem amostras representativas tanto de tráfego benigno quanto de tráfego malicioso, o EFC requer apenas exemplos benignos para construir o modelo, contornando a dependência de amostras maliciosas rotuladas e contribuindo para sua adaptabilidade a diferentes domínios (2, 31).

Em termos formais, para um vetor discreto de atributos $\mathbf{x} = (x_1, \ldots, x_N)$, a energia de um fluxo pode ser expressa por

$$\mathcal{H}(\mathbf{x}) = -\sum_{i<j} e_{ij}(x_i, x_j) - \sum_{i} h_i(x_i), \tag{2.1}$$

em que $h_i(x_i)$ representa o campo local associado ao estado assumido pelo atributo i, e $e_{ij}(x_i, x_j)$ representa o acoplamento entre os atributos i e j, conforme a formulação original do EFC (2). O índice $i < j$ na primeira soma percorre todos os pares não ordenados de atributos distintos, evitando a dupla contagem dos acoplamentos. De forma intuitiva, a energia de um fluxo corresponde à soma negativa dos campos locais e acoplamentos associados aos valores de seus atributos; assim, fluxos mais compatíveis com a distribuição benigna aprendida tendem a apresentar energias mais baixas, enquanto fluxos mais improváveis sob esse modelo tendem a apresentar energias mais elevadas.

Na etapa de detecção, calcula-se a energia de novos fluxos e compara-se esse valor a um limiar de energia (*cutoff*). A classificação é realizada assumindo que fluxos com energia inferior ao limiar são benignos, enquanto fluxos com energia maior ou igual ao limiar são classificados como maliciosos. Nos experimentos reportados pela autora, foi adotado um limiar estático definido como o 95º percentil da distribuição de energias dos fluxos benignos utilizados no treinamento (2).

Uma propriedade particularmente relevante do EFC é a interpretabilidade. Como o modelo é definido por campos locais e acoplamentos discretos, ele produz uma representação transparente do processo de decisão, em contraste com classificadores de caixa-preta (2). Além disso, os experimentos originais mostraram maior robustez a mudanças de domínio quando comparado a classificadores supervisionados clássicos em *benchmarks* como CIDDS-001, CICIDS2017 e CICDDoS2019 (2). Aplicações posteriores reforçaram esse comportamento na detecção de botnets (31), enquanto outros trabalhos exploraram sua aplicação em tráfego IoT/MQTT e em cenários com ataques desconhecidos (32). Mais recentemente, a

linha evoluiu para classificação aberta e multiclasse, estendendo o EFC para reconhecer ataques desconhecidos com baixa complexidade temporal (33).

2.2 DERIVA TEMPORAL E DESALINHAMENTO DISTRIBUCIONAL

O pressuposto de que dados de treinamento e dados futuros são amostras de uma mesma distribuição raramente se sustenta no domínio da segurança de redes. Diferentemente de aplicações em contextos mais estáveis, o tráfego de rede é influenciado por fatores operacionais, tecnológicos e adversariais que se modificam continuamente. A literatura destaca, nesse sentido, três fontes principais de instabilidade, a saber, a variabilidade intra-rede, a variabilidade entre redes distintas e a própria natureza adversarial do domínio (3). Em termos práticos, isso significa que o comportamento aprendido a partir de um recorte de dados pode rapidamente perder aderência quando o modelo é exposto a novos contextos de uso.

Em redes reais, novos dispositivos são incorporados, aplicações são introduzidas ou descontinuadas, protocolos evoluem, janelas de atividade seguem ciclos diurnos e semanais, políticas de acesso são alteradas e padrões de consumo se reorganizam ao longo do tempo. Paralelamente, atores maliciosos ajustam TTPs, exploram vulnerabilidades emergentes, alteram infraestrutura de ataque e buscam adaptar seus comportamentos para contornar mecanismos de defesa. Sob essas condições, o tráfego benigno futuro pode se deslocar para regiões do espaço de atributos antes pouco frequentes, ao mesmo tempo em que novas formas de tráfego malicioso podem surgir fora do perfil previamente observado. O resultado é um processo contínuo de desalinhamento distribucional entre o ambiente de treinamento e o ambiente de operação.

Esse fenômeno pode se manifestar, de forma geral, em duas dimensões complementares. A primeira corresponde ao *data drift*, no qual a distribuição marginal dos atributos muda ao longo do tempo, ainda que a relação entre atributos e rótulos permaneça relativamente estável. A segunda diz respeito ao *concept drift*, no qual se altera a própria relação entre os atributos observados e a interpretação semântica da classe, de modo que padrões antes associados ao benigno ou ao malicioso deixam de ter o mesmo significado operacional (24, 3). Em segurança de redes, essas duas formas de deriva frequentemente coexistem, o que torna insuficiente a avaliação baseada apenas em partições aleatórias extraídas de um mesmo *dataset* estático.

Para detectores *one-class*, o impacto da deriva temporal tende a ser particularmente sensível. Como o modelo é ajustado exclusivamente a partir do tráfego benigno disponível no treinamento, qualquer deslocamento do benigno futuro pode aumentar sua energia, sua distância ou seu erro de reconstrução, dependendo do método empregado. Nesses casos, o detector pode continuar sensível a diversas atividades maliciosas, mas passar a penalizar porções crescentes do tráfego legítimo, elevando a taxa de falsos positivos. O problema é operacionalmente relevante porque degrada a confiabilidade do sistema, amplia o custo de triagem manual e dificulta a distinção entre mudanças legítimas do ambiente e anomalias efetivas.

A literatura recente reforça esse diagnóstico. Trabalhos voltados à avaliação em cenários mais realistas mostram que modelos treinados em cortes temporais estáticos tendem a sofrer degradação quando aplicados a janelas futuras, mesmo na ausência de alterações radicais no cenário de rede. O benchmark MAWIFlow, por exemplo, evidencia que a passagem do tempo afeta de forma mensurável a robustez de

detectores de anomalia, o que reforça a necessidade de validação temporal explícita e de protocolos experimentais que preservem a natureza prospectiva do uso real (21). Desse modo, a consideração da deriva temporal deixa de ser apenas uma preocupação metodológica adicional e passa a constituir requisito central para avaliações mais realistas de NIDS baseados em aprendizado de máquina.

A literatura de aprendizado de máquina oferece arcabouços consolidados para detecção e adaptação a *concept drift*, com revisões abrangentes que organizam as estratégias em re-treinamento periódico, janelas deslizantes e detectores explícitos de mudança distribucional (34). No contexto de NIDS, essa dimensão é particularmente relevante, pois tanto o tráfego benigno quanto os padrões de ataque evoluem continuamente, tornando insuficientes as abordagens estáticas baseadas em treinamento único. Antes de adotar mecanismos de adaptação, contudo, é necessário dispor de protocolos capazes de evidenciar e quantificar o deslocamento distribucional ao longo do tempo, requisito que reforça a pertinência da validação temporal e do diagnóstico distribucional empregados neste trabalho.

2.3 ESTADO DA ARTE EM *DATASETS* PARA NIDS

A avaliação de NIDS baseados em aprendizado de máquina depende fortemente da qualidade, da atualidade e da representatividade dos *datasets* empregados. Por essa razão, as pesquisas recentes têm dedicado atenção não apenas ao desenvolvimento de novos classificadores, mas também à análise crítica dos conjuntos de dados disponíveis. Em termos metodológicos, um *dataset* útil para NIDS não deve ser avaliado apenas pelo volume de amostras ou pela quantidade de ataques contemplados, mas também por propriedades como realismo do tráfego benigno, diversidade de cenários, granularidade dos dados, consistência dos atributos, transparência do processo de rotulagem, cobertura de tráfego criptografado e validade temporal do protocolo experimental (3, 10, 17). Esta seção apresenta o estado da arte em *datasets* para NIDS com ênfase nesses critérios.

A literatura oferece ampla variedade de *datasets* para avaliação de NIDS, mas poucos atendem simultaneamente a requisitos de atualidade, diversidade, documentação, criptografia, representatividade do benigno e transparência do processo de construção. O DARPA 1998/1999 e o KDD Cup 99 tiveram papel histórico na consolidação do campo, porém são amplamente reconhecidos como obsoletos e pouco representativos do tráfego contemporâneo (35, 36). O UNSW-NB15 foi concebido como alternativa mais atualizada, combinando tráfego normal moderno e ataques sintetizados em laboratório (13, 37). Ainda assim, permanece sujeito a limitações típicas de cenários montados, entre as quais se destacam a normalidade relativamente controlada, a dependência de ferramentas específicas de extração e a possível distância em relação às condições de produção.

Nos anos seguintes, o foco deslocou-se progressivamente para *datasets* mais ricos em diversidade de cenários e de fontes de dados. O CIC-IDS2017 representou avanço importante ao combinar múltiplos dias de tráfego benigno com diferentes famílias de ataque, tornando-se *benchmark* dominante em estudos recentes (14). Entretanto, análises posteriores apontaram problemas de geração, duplicações e inconsistências de rótulo em conjuntos da família CIC, o que reforça a necessidade de cautela metodológica na interpretação de resultados (4, 38). O TON_IoT, por sua vez, ampliou a heterogeneidade de fontes ao integrar

telemetria IoT/IIoT e dados de rede, mas não resolve integralmente a lacuna de tráfego benigno realista e de avaliação temporal prospectiva (15, 16, 39). Em paralelo, trabalhos como os de (9, 10) chamaram atenção para a necessidade de padronização mínima dos atributos de fluxo, uma vez que diferenças de granularidade, semântica e processo de extração dificultam a comparação entre estudos e a reprodutibilidade experimental.

Outra linha importante do estado da arte diz respeito aos *datasets* que procuram equilibrar realismo, documentação e controle experimental. O HIKARI-2021 é representativo desse esforço ao combinar tráfego benigno real com ataques sintéticos criptografados e ao explicitar requisitos de conteúdo e de processo para a construção do conjunto de dados (6). De modo semelhante, a literatura sobre geração sintética e injeção controlada de ataques passou a enfatizar que a utilidade do *dataset* depende não apenas da presença de amostras maliciosas, mas da capacidade de preservar propriedades estatísticas do tráfego de fundo e de documentar o processo de geração de forma auditável (3, 19). Ainda assim, mesmo quando essas iniciativas avançam em transparência, a componente maliciosa frequentemente permanece sintética e a validação costuma ser conduzida em cenários temporais relativamente fechados.

Estudos comparativos recentes mostram que a diferença estatística entre *benchmarks* sintéticos e ambientes reais continua expressiva. (11) demonstram que o benigno de *datasets* sintéticos difere substancialmente do benigno coletado em redes de produção, o que ajuda a explicar o fraco transporte de desempenho do laboratório para a operação. Trabalhos mais recentes, como o MAWIFlow, reforçam esse diagnóstico ao propor *benchmarks* baseados em tráfego real e avaliação em janelas temporais distintas, evidenciando a degradação de desempenho sob deriva temporal (21). Em direção semelhante, (40) destacam a relevância de coleções obtidas em ambiente real, com mistura de tráfego cifrado e não cifrado, tanto para análise de aplicações quanto para segurança.

Por fim, revisões amplas do domínio indicam que o desafio contemporâneo já não se resume à escassez de dados. (3), ao analisarem 89 *datasets*, mostram que o problema central reside também no desconhecimento das propriedades dos conjuntos disponíveis e na ausência de melhores práticas consolidadas para sua criação, seleção e uso. Em outras palavras, o estado da arte em *datasets* para NIDS revela uma transição importante: da mera disponibilização de conjuntos de dados para a exigência de processos metodológicos mais transparentes, auditáveis e alinhados às condições observadas em redes reais. É precisamente nesse ponto que se insere a proposta desta dissertação.

2.4 TRABALHOS CORRELATOS

Os trabalhos mais diretamente relacionados a esta dissertação concentram-se em três frentes complementares, quais sejam, a construção de *datasets* mais realistas, a injeção controlada de ataques e o aprimoramento dos processos de rotulagem. Embora essas linhas tenham produzido avanços relevantes, a literatura ainda carece de uma articulação sistemática entre coleta realista, rotulagem assistida, validação temporal prospectiva e avaliação de detectores *one-class* sob deriva temporal.

Estudos pioneiros, como o de (41), já defendiam a necessidade de uma abordagem sistemática para criação de *benchmarks* de intrusão. Em seguida, pesquisas como as de (42), (43) e (44) exploraram diferentes

estratégias para produzir tráfego mais representativo, parametrizável e aderente a contextos específicos. Nessa mesma linha, o ID2T e seus desdobramentos consolidaram a ideia de injeção sintética sobre tráfego de fundo com preocupação explícita em preservar propriedades estatísticas do ambiente e evitar padrões artificiais conhecidos (19, 45, 46). O HIKARI-2021, por sua vez, ampliou essa discussão ao enfatizar a inclusão de tráfego criptografado e a necessidade de documentar não apenas o conteúdo do *dataset*, mas também o processo de sua construção (6).

No campo da rotulagem, a literatura tem ressaltado que a utilidade de um *dataset* depende tanto da qualidade dos dados quanto da transparência do processo de anotação (18). Nesse contexto, propostas baseadas em rotulagem autônoma, auto-treinada, ativa e interativa buscaram reduzir o custo humano sem comprometer a consistência dos rótulos (47, 48, 49, 50). Apesar desses avanços, tais abordagens tendem a tratar a rotulagem como problema isolado, sem articulá-la de forma explícita com a temporalidade da avaliação, com a construção do *dataset* em ambiente realista e com a validação de detectores *one-class* sob deriva temporal.

Com o objetivo de sintetizar esse panorama, a Tabela 2.1 organiza os trabalhos correlatos segundo objetivo, abordagem e limitações, em paralelo com as dimensões centrais discutidas no artigo que fundamenta esta dissertação, a saber, coleta realista, rotulagem, temporalidade, generalização, formas de dados e objetivo do modelo (22).

A síntese apresentada na Tabela 2.1 mostra que os estudos recentes, embora relevantes, tendem a atender apenas parcialmente às dimensões consideradas centrais para esta pesquisa. Edge-IIoT, RT-IoT2022 e FLNET2023 priorizam coleta realista e generalização em contextos específicos, mas não conferem centralidade à rotulagem assistida auditável nem à validação temporal prospectiva. HIKARI-2021 avança ao incorporar tráfego criptografado e explicitar requisitos de processo, mas mantém componente maliciosa sintética e não estrutura a avaliação em recortes temporais prospectivos. MAWIFlow, por sua vez, representa avanço importante na consideração da temporalidade, embora permaneça dependente de rótulos herdados do processo original. Já Goldschmidt et al. oferecem uma contribuição de natureza analítica, fundamental para o diagnóstico das limitações da área, mas não se configuram como proposta metodológica de geração de dados.

Para reduzir a subjetividade da comparação qualitativa, a análise foi estruturada em seis critérios. O critério RC corresponde à coleta realista de tráfego e indica a presença de tráfego benigno real ou representativo do ambiente operacional. O critério LB corresponde à rotulagem auditável e indica a existência de procedimento explícito, rastreável e verificável de atribuição de rótulos. O critério TMP corresponde à avaliação temporal e identifica trabalhos que adotam separação cronológica dos dados ou análise por janelas distintas. O critério GEN corresponde à evidência de generalização e registra avaliações entre janelas, ambientes, silos ou *datasets*. O critério DF corresponde aos formatos ou representações dos dados e indica o suporte ou a disponibilidade de pacotes, fluxos ou atributos derivados. O critério MO corresponde ao objetivo de modelagem ou detecção e indica a declaração explícita de tarefas como detecção de anomalias, classificação binária, classificação multiclasse ou avaliação *open-set*. A marcação positiva foi atribuída apenas quando o critério correspondente estava claramente descrito no trabalho analisado. Nos casos de ausência, ambiguidade ou tratamento apenas indireto, adotou-se marcação negativa.

Tabela 2.1: Síntese comparativa de trabalhos correlatos segundo as dimensões de objetivo, abordagem e limitações.

Estudo	Objetivo	Abordagem	Limitações
Edge-IIoT / Edge-IIoT-2022(51)	Disponibilizar *dataset* IoT/IIoT realista para IDS.	Testbed multicamadas com dispositivos IoT/IIoT, 14 ataques e atributos para classificação binária e multiclasse.	Não enfatiza rotulagem assistida nem validação temporal prospectiva; foco maior na avaliação de classificadores.
RT-IoT2022(52)	Gerar tráfego IoT para avaliação de modelos leves.	Coleta em ambiente IoT real, ataques DDoS e SSH *brute force*, conversão por CICFlowMeter e uso em modelos de anomalia.	Escopo restrito a IoT; não trata rotulagem auditável nem separação temporal prospectiva.
HIKARI-2021(6)	Disponibilizar um *dataset* com tráfego benigno real e ataques sintéticos criptografados, com ênfase em requisitos de construção e documentação do processo.	Geração de tráfego em ambiente real, incorporação de ataques sintéticos sobre comunicações criptografadas e publicação dos procedimentos de construção e dos requisitos do *dataset*.	Embora avance em transparência de processo e incorpore tráfego criptografado, não enfatiza rotulagem assistida auditável nem validação temporal prospectiva; a componente maliciosa permanece sintética.
FLNET2023(53)	Oferecer *dataset* realista para NIDS em aprendizado federado.	Construção orientada a silos distribuídos e avaliação de generalização em ambiente federado.	Ênfase no cenário federado; não destaca rotulagem assistida, temporalidade ou *pipeline* auditável.
Koumar et al. (2023)(54)	Melhorar a classificação de tráfego por enriquecimento temporal.	Proposição de 69 atributos derivados de séries temporais de fluxo e avaliação em múltiplos *datasets*.	Não constrói *dataset*; não aborda rotulagem auditável nem coleta temporal.
MAWIFlow (2026)(21)	Fornecer *benchmark* realista para avaliação sob deriva temporal.	Pipeline reprodutível a partir do MAWILab, com conversão para fluxos e avaliação em janelas temporais distintas.	Não enfatiza rotulagem assistida; depende de rótulos herdados do processo original.
Wang et al. (2025)(40)	Disponibilizar *dataset* real para análise e classificação de tráfego.	Coleta *in situ*, com tráfego criptografado e não criptografado, processamento e publicação em PCAP/CSV.	Foco em análise e classificação de tráfego, não em NIDS; não prioriza rotulagem auditável nem validação temporal.
Goldschmidt et al. (2025)(3)	Sistematizar o estado da arte dos *datasets* de NIDS.	Revisão sistemática de 89 *datasets*, comparação por propriedades e discussão de limitações recorrentes.	Não propõe *pipeline* de geração de dados; contribuição é analítica e orientativa.

A Tabela 2.2 sintetiza a aderência dos trabalhos recentes a esses seis critérios. A comparação evidencia que os conjuntos de dados e arcabouços existentes tendem a atender apenas parte das dimensões analisadas. Entre os trabalhos externos analisados, não se identificou proposta que combine simultaneamente coleta realista de tráfego benigno, rotulagem explícita e auditável, validação temporal, evidência de generalização, suporte a múltiplas representações dos dados e objetivo declarado de detecção ou modelagem. Essa lacuna, observada a partir dos critérios adotados, reforça a necessidade de metodologias que integrem, em um mesmo processo experimental, realismo operacional, rastreabilidade dos rótulos, avaliação prospectiva e documentação suficiente para auditoria e reprodução.

Como a última linha da tabela sintetiza a proposta desenvolvida nesta dissertação, sua marcação deve ser lida como aplicação dos critérios definidos aos artefatos e procedimentos apresentados no próprio estudo, segundo os mesmos parâmetros adotados para os trabalhos externos. No critério GEN, adotou-se marcação parcial ($\checkmark^{*}$) para tornar explícito que a evidência de generalização se restringe à avaliação em janelas temporais distintas, T1, T2 e T3, e não implica validação em ambientes, redes ou cenários independentes, forma de generalização que uma coleta única não cobre. Os demais critérios da linha

Tabela 2.2: Comparação qualitativa com trabalhos correlatos (2022–2026).

Referência	**RC**	**LB**	**TMP**	**GEN**	**DF**	**MO**
Edge-IIoT (2025)(51, 55)	✓	×	×	✓	✓	×
RT-IoT (2022)(52)	✓	×	×	✓	✓	×
FLNET (2023)(53)	✓	×	×	✓	✓	×
Koumar et al. (2023)(54)	✓	×	✓	✓	✓	✓
MAWIFlow (2026)(21)	✓	×	✓	✓	✓	✓
Wang et al. (2025)(40)	✓	×	✓	✓	✓	✓
Goldschmidt et al. (2025)(3)	✓	✓	✓	✓	✓	×
Este estudo	✓	✓	✓	✓*	✓	✓

recebem marcação plena segundo os parâmetros descritos anteriormente, aplicados de modo uniforme a todos os trabalhos comparados.

É nesse ponto que a proposta desta dissertação se diferencia. Em vez de tratar isoladamente coleta, rotulagem ou avaliação temporal, o estudo articula essas dimensões em um único *pipeline*, integrando coleta contínua de tráfego benigno real, execução controlada de ataques, extração e discretização de atributos, rotulagem assistida, validação temporal prospectiva e diagnóstico distribucional, com objetivo explícito de avaliar detectores de anomalia em nível de fluxo sob deriva temporal.

3 METODOLOGIA

3.1 METODOLOGIA PROPOSTA

Esta pesquisa propõe uma metodologia para construir e avaliar *datasets* de fluxos destinados à detecção de intrusão sob deriva temporal. A metodologia organiza a coleta de tráfego benigno real, a injeção controlada de ataques, a preparação dos atributos, a rotulagem rastreável, a separação temporal dos dados e o diagnóstico de mudanças distribucionais.

A metodologia não depende de um algoritmo específico de detecção. Cada uma de suas etapas pode ser empregada com detectores *one-class*, supervisionados ou não supervisionados, desde que o modelo receba os atributos preparados pelo *pipeline* e produza uma decisão que possa ser avaliada nos recortes temporais. O EFC é utilizado neste trabalho como estudo de caso para demonstrar a aplicação da metodologia e examinar suas propriedades em um detector baseado em energia. A Figura 3.1 apresenta o fluxo geral da proposta.

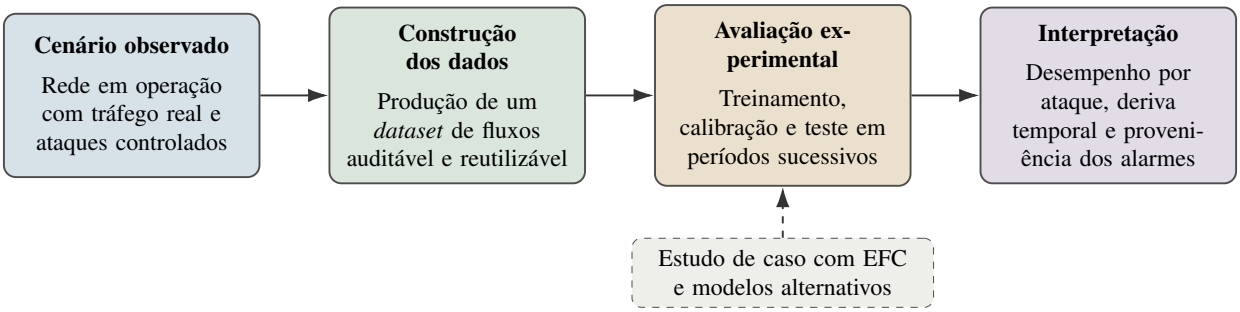


Figura 3.1: Funcionamento geral do trabalho desde a observação da rede até a interpretação dos resultados.

A Figura 3.1 posiciona a metodologia no fluxo completo da pesquisa. O trabalho parte de um cenário operacional, produz um *dataset* auditável, avalia detectores em períodos sucessivos e reúne evidências para interpretar o desempenho e a deriva temporal. As figuras seguintes detalham as transformações realizadas em cada etapa.

3.1.1 Construção e preparação do dataset

A construção e a preparação do *dataset* compreendem três estágios. O primeiro realiza a captura do tráfego e a execução controlada dos ataques. O segundo transforma os fluxos brutos em atributos adequados à modelagem. O terceiro consolida os rótulos, a proveniência e os recortes temporais. Cada estágio é representado em sua respectiva subseção.

3.1.1.1 Estágio 1. Captura de tráfego e simulação de ataques

Nesta fase, o tráfego benigno é coletado em uma rede real ou em um ambiente de teste representativo, enquanto ataques controlados são executados de forma planejada. O objetivo é produzir um corpus de

fluxos composto predominantemente por comunicações legítimas e, de modo deliberado, por instâncias conhecidas de atividade maliciosa. Essa combinação permite aproximar o cenário experimental de condições operacionais reais, preservando, ao mesmo tempo, o controle sobre os períodos, os alvos e os tipos de ataques injetados.

A Figura 3.2 representa o primeiro estágio e destaca a relação entre o plano experimental, a rede observada, a exportação dos fluxos e as evidências coletadas.

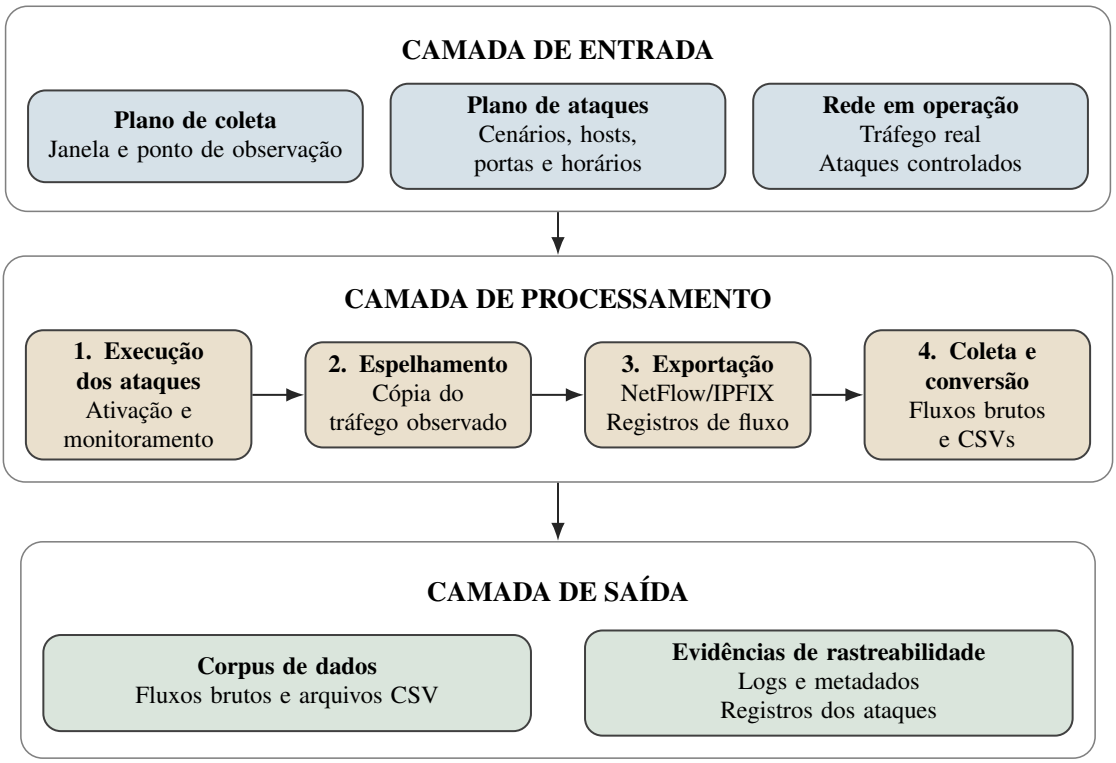


Figura 3.2: Arquitetura operacional da captura do tráfego e da execução controlada dos ataques.

A metodologia foi projetada para ser aplicada a redes em operação normal, sem exigir ambientes de teste isolados ou tráfego sintético. O estudo de caso descrito no Capítulo 4 operacionaliza essa premissa em uma rede governamental real, com tráfego predominantemente benigno e ataques controlados injetados em janelas planejadas.

O caráter realista do *dataset* decorre da coleta do tráfego benigno durante a operação normal da rede. Os fluxos registrados preservam padrões efetivos de comunicação, uso de aplicações, serviços internos, horários de atividade e interações entre hosts que não foram sintetizados para o experimento. Os ataques são introduzidos de forma controlada para que sua origem, duração, alvo e evidências permaneçam verificáveis. Essa combinação preserva a variabilidade do tráfego legítimo e permite estabelecer *ground truth* rastreável para a atividade maliciosa (6, 19).

O tráfego observado é espelhado em um ponto central da rede e exportado em formato NetFlow/IPFIX, conforme ilustrado na Figura 3.2. Essa abordagem permite registrar informações resumidas sobre as comunicações que atravessam o dispositivo exportador, como endereços IP de origem e destino, portas, protocolo, interface, volume de bytes, quantidade de pacotes e duração da comunicação. Diferentemente da captura integral de pacotes, a análise baseada em fluxos não depende da inspeção completa do conteúdo

transportado, o que reduz o custo de armazenamento e favorece sua aplicação em ambientes com grande volume de tráfego.

No processo de coleta, os pacotes com características comuns são agregados em registros de fluxo, mantidos temporariamente em cache pelo dispositivo exportador e, posteriormente, enviados ao coletor NetFlow/IPFIX. No ambiente experimental, a coleta é realizada pelo serviço `netflow-collector`, que armazena os fluxos brutos e os converte para arquivos CSV, utilizados nas etapas subsequentes de preparação do *dataset*. Dessa forma, torna-se possível obter visibilidade sobre quem se comunica com quem, por qual porta, usando qual protocolo, por quanto tempo e com qual volume de tráfego, sem depender da análise do *payload*.

A execução dos ataques controlados ocorre em paralelo à janela de coleta, a partir do serviço `attacker`, conforme o plano experimental previamente definido. Ataques como *port scan*, *brute force*, *web exploits*, DoS/DDoS e atividades associadas a *botnets* podem ser inseridos de maneira rastreável, permitindo associar os eventos maliciosos aos respectivos intervalos temporais, endereços de origem e destino, protocolos e portas envolvidos. Essas informações servem de base para a etapa posterior de rotulagem assistida.

Nessa etapa, o monitoramento não se limita à captura dos fluxos, mas também envolve o registro de evidências auxiliares sobre o período observado, como horários de execução dos ataques, disponibilidade dos serviços, eventos de rede e alertas eventualmente produzidos por ferramentas externas. Esses registros apoiam a interpretação posterior dos fluxos exportados e reduzem a dependência exclusiva dos metadados NetFlow/IPFIX no processo de rotulagem.

Por fim, embora o tráfego não associado aos ataques planejados componha majoritariamente a porção benigna do corpus, ele não deve ser automaticamente presumido como benigno. Por essa razão, recomenda-se o uso de mecanismos auxiliares de validação e monitoramento, como IDSs, registros de sistema e ferramentas de análise de fluxo, de modo a apoiar a identificação de anomalias não planejadas e aumentar a confiabilidade do processo de rotulagem.

3.1.1.2 Estágio 2. Extração e discretização de atributos

Com os registros de fluxo exportados, inicia-se a preparação dos atributos. Primeiro, são realizadas etapas de limpeza, padronização de tipos, tratamento de valores ausentes e atenuação de valores extremos. Em seguida, são selecionados atributos relevantes para a modelagem, como endereços IP, portas, protocolo, bytes, pacotes, duração, direção e sub-redes.

Para o EFC, a discretização constitui etapa central, uma vez que o modelo opera sobre atributos discretos. Nesta pesquisa, adotou-se uma estratégia híbrida, na qual o valor zero é isolado em um *bin* específico e os valores positivos são distribuídos por quantis.

No estudo de caso com o EFC, utilizaram-se $n\_bins = 30$, *port-topN*=50 para o agrupamento de portas de serviço e winsorização da duração em 1%. Atributos categóricos como direção (*lan2wan*, *wan2lan*, *intra*, *wan2wan*) e sub-redes /24 foram incorporados como contexto.

A Figura 3.3 representa as transformações aplicadas aos registros brutos e os artefatos produzidos para a modelagem.

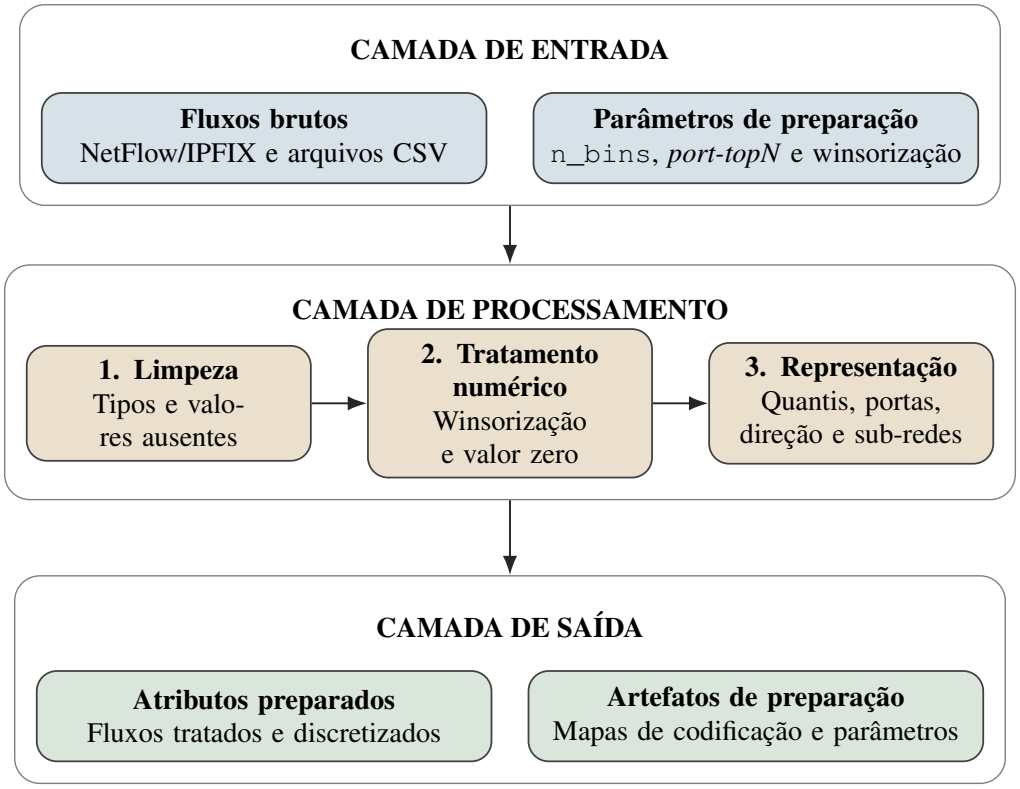


Figura 3.3: Preparação dos atributos e produção da representação utilizada pelos detectores.

3.1.1.3 Estágio 3. Rotulagem assistida e preparação de dados

A rotulagem assistida constitui etapa essencial para assegurar a auditabilidade do *dataset*. Nesta pesquisa, os rótulos maliciosos foram atribuídos pela correlação entre o plano de ataques controlados, as janelas temporais de execução, os endereços de origem e destino, os protocolos, as portas envolvidas e os metadados registrados durante a coleta. Cada fluxo candidato a malicioso foi vinculado a uma evidência de rotulagem, preservando a rastreabilidade entre o registro de fluxo e o cenário experimental que motivou sua anotação.

A regra primária de rotulagem considerou a sobreposição temporal entre o fluxo observado e a janela planejada do ataque, em conjunto com os critérios de origem, destino, porta e protocolo definidos no plano experimental. Nos ataques de varredura, admitiu-se correspondência por faixa de endereçamento ou CIDR, uma vez que esse tipo de ataque envolve múltiplos destinos. Nos ataques direcionados a serviços específicos, priorizou-se a correspondência exata entre origem, destino e porta. Os fluxos que não atenderam aos critérios mínimos de correspondência permaneceram rotulados como benignos, salvo quando havia evidência auxiliar suficiente para justificar sua reclassificação. A captura foi realizada de forma bidirecional no ponto central de espelhamento, registrando tanto os fluxos de ida quanto os de retorno.

O fluxo de ida do `syn_flood` satisfaz a correspondência entre o host atacante, o alvo e a porta 8080. O fluxo de retorno parte do alvo, tem como destino o host atacante e utiliza uma porta efêmera no

lado de destino. Como essa direção não satisfaz a correspondência exata definida para a ida do ataque, ela permaneceu sob rótulo benigno no pacote congelado. Esse retorno não representa uma comunicação legítima independente do ataque. Ele é gerado em resposta às tentativas TCP SYN enviadas ao alvo, ocorre entre os mesmos hosts envolvidos na campanha e concentra-se nas janelas do `syn_flood`. Seu impacto sobre as métricas é analisado na Seção 4.7.1.

Como mecanismo de controle de qualidade, os rótulos foram organizados em três níveis. O primeiro nível corresponde ao rótulo binário, que distingue tráfego benigno e malicioso. O segundo corresponde ao tipo de ataque, com classes como `syn_flood` e `aggressive_scan`. O terceiro corresponde à campanha específica, com execuções individuais como A01, A02, A03 e A04. Essa organização é relevante porque inconsistências no nível da campanha podem não comprometer a avaliação binária do detector, embora limitem inferências mais detalhadas sobre a contribuição de cada execução de ataque.

A Figura 3.4 apresenta a combinação das evidências usadas na rotulagem e os produtos necessários à avaliação temporal.

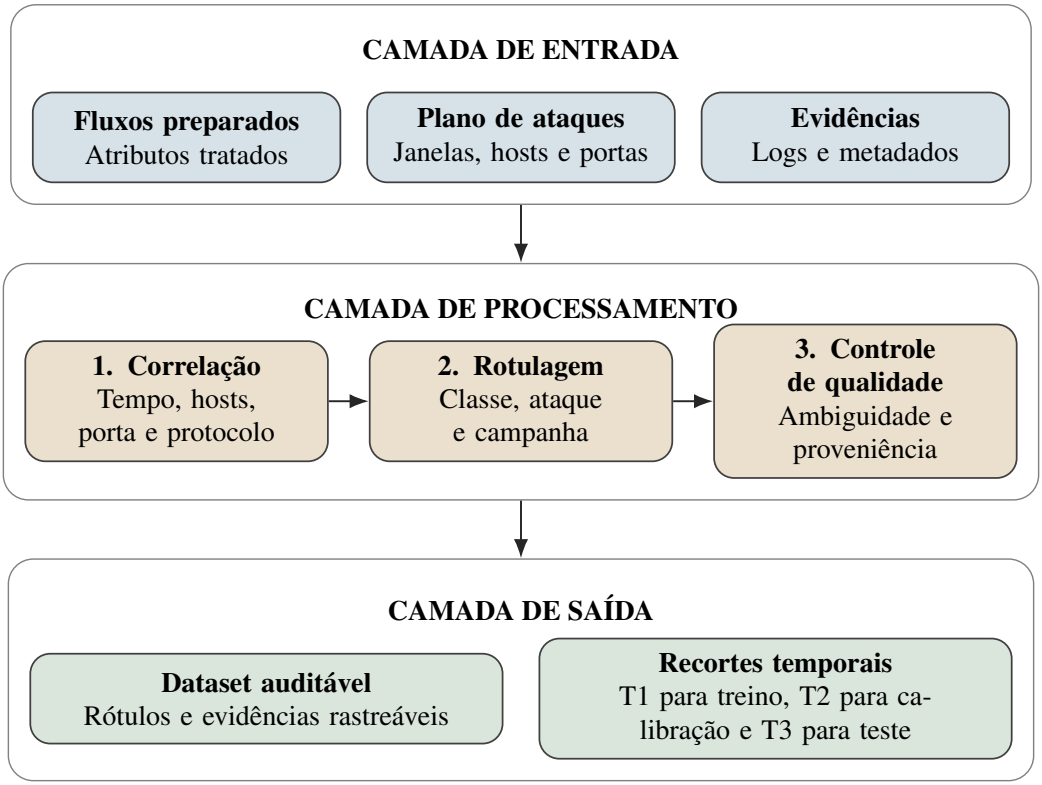


Figura 3.4: Rotulagem assistida, controle de qualidade e organização temporal dos dados.

Casos ambíguos, como fluxos próximos aos limites temporais das campanhas ou fluxos apenas parcialmente compatíveis com os critérios de origem e destino, foram tratados de forma conservadora. Na ausência de evidência suficiente, recomenda-se manter o rótulo benigno ou registrar o caso como ambíguo em artefato auxiliar, evitando a ampliação artificial da classe maliciosa. Essa decisão preserva a rastreabilidade do processo e permite que revisões futuras do *dataset* reavaliem os rótulos com base em novas evidências, como alertas de IDS, logs de firewall, registros de sistema ou inspeção dirigida.

Após a consolidação dos rótulos, o *pipeline* produz um conjunto de fluxos preparado para avaliação experimental. Suas entradas compreendem registros NetFlow/IPFIX, planos de coleta e ataque, metadados e evidências de execução. Suas saídas compreendem atributos tratados e discretizados, mapas de codificação, rótulos com evidências de proveniência e recortes temporais para treinamento, calibração e teste. Esses artefatos são produzidos antes da escolha do detector e podem ser reutilizados por modelos distintos.

3.1.2 Avaliação prospectiva e diagnóstico distribucional

Concluída a construção e a preparação do *dataset*, passa-se ao protocolo de avaliação prospectiva e ao diagnóstico distribucional entre janelas temporais. Esta fase examina o comportamento do detector em períodos cronologicamente posteriores ao treinamento e avalia sua sensibilidade a deslocamentos no tráfego benigno.

A avaliação temporal é necessária porque os fluxos de rede apresentam dependência temporal, ciclos de uso e alterações de perfil que enfraquecem a suposição de independência entre observações. Uma partição aleatória pode distribuir padrões muito semelhantes entre treinamento e teste e produzir uma estimativa otimista do desempenho. A separação cronológica preserva a ordem em que as informações se tornam disponíveis e aproxima a avaliação das condições de uso do detector em operação (56, 57, 58).

O recorte T1 contém somente tráfego benigno e define o perfil de normalidade usado no treinamento. Essa restrição evita que a fronteira de decisão de um detector *one-class* incorpore comportamentos associados aos ataques controlados. O recorte T2 contém tráfego benigno e ataques rotulados. Ele permite selecionar o ponto operacional com informações disponíveis antes da janela final. O recorte T3 permanece separado durante a calibração e representa o período futuro usado para teste.

O *cutoff* aplicado a T3 deve ser definido com informações de T1 e T2. Essa escolha reproduz uma condição operacional na qual os rótulos de eventos futuros não estão disponíveis para ajustar o detector. A busca do melhor *cutoff* com os rótulos de T3 pode ser mantida como análise retrospectiva de sensibilidade, mas não deve ser interpretada como resultado prospectivo.

Essa organização permite avaliar simultaneamente a capacidade de detecção do modelo e sua estabilidade diante da evolução temporal do tráfego. A comparação entre T2 e T3 indica se o detector mantém desempenho em uma janela posterior. O diagnóstico distribucional oferece evidência complementar sobre mudanças nas distribuições dos atributos utilizados pelo modelo.

A Figura 3.5 sintetiza a progressão cronológica do protocolo experimental. O diagnóstico distribucional é representado como etapa complementar à avaliação de desempenho, sendo calculado entre janelas temporais a partir dos fluxos benignos, de modo a caracterizar deslocamentos no perfil do tráfego legítimo sem confundi-los com o efeito volumétrico dos ataques controlados.

A separação em três janelas temporais permite observar o desempenho do detector em recortes cronologicamente distintos. T1 é utilizado para treinamento, T2 para calibração e avaliação intermediária e T3 para teste prospectivo final. Quando houver ajuste de *cutoff* diretamente em T3, esse resultado deve ser interpretado apenas como análise retrospectiva de sensibilidade.

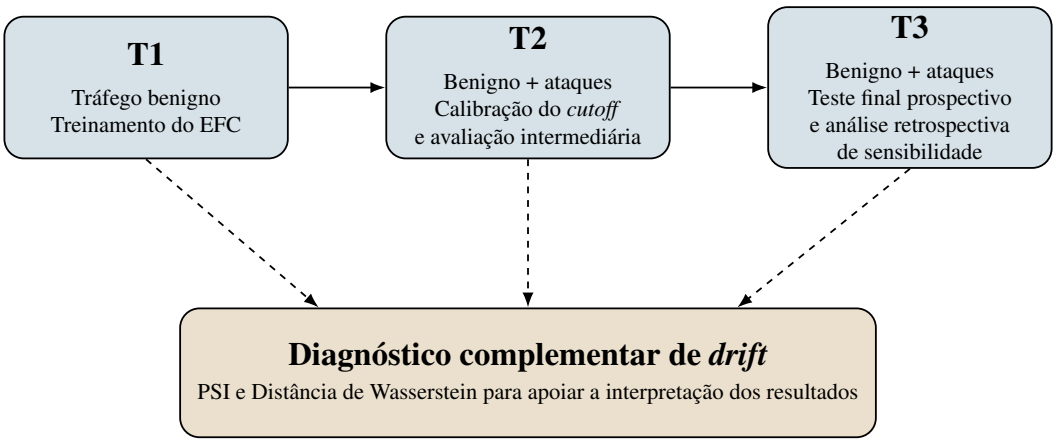


Figura 3.5: Protocolo de validação temporal. T1 define a normalidade, T2 seleciona o ponto operacional com informação historicamente disponível e T3 permanece fora da calibração para representar dados futuros.

Além das métricas de desempenho, a metodologia também incorpora uma etapa de diagnóstico distribucional, destinada a examinar se os atributos dos fluxos apresentam deslocamentos entre as janelas temporais analisadas. Para esse fim, empregam-se o *Population Stability Index* (PSI) e a Distância de Wasserstein, descritos a seguir.

3.1.2.1 Population Stability Index e Distância de Wasserstein

Como medida de estabilidade distribucional, emprega-se o *Population Stability Index* (PSI) para quantificar mudanças nas distribuições das principais *features* entre janelas temporais. O PSI compara a proporção de observações da população de referência e da população de comparação em um conjunto comum de intervalos, funcionando como indicador da presença e da intensidade de deslocamentos distribucionais. Por esse motivo, mostra-se particularmente compatível com o pipeline desta pesquisa, no qual atributos numéricos relevantes para o EFC são previamente discretizados.

O PSI foi escolhido porque compara as proporções observadas em intervalos comuns e se ajusta diretamente aos atributos discretizados utilizados no *pipeline*. Ele permite identificar quais distribuições alteraram sua composição entre as janelas temporais. A Distância de Wasserstein foi incluída porque quantifica o deslocamento de massa entre valores ordenados e permite distinguir mudanças entre valores próximos e mudanças entre valores distantes. O uso conjunto reduz as limitações de uma leitura baseada em uma única medida, especialmente quando o PSI é ampliado pela ausência de categorias ou intervalos em uma das janelas (59, 11).

Na literatura de monitoramento de modelos, o PSI tem sido utilizado como ferramenta para detecção de *distribution shift*, especialmente em cenários pós-implantação, nos quais se deseja comparar os dados observados em operação com aqueles utilizados no desenvolvimento do modelo. Embora o estudo de (59) tenha sido conduzido em outro domínio de aplicação, os autores indicam que o PSI apresenta potencial para detectar mudanças distribucionais, ao mesmo tempo em que ressaltam limitações importantes, como a dependência do número de *bins* e a dificuldade de comparação direta entre aplicações distintas.

No contexto desta pesquisa, o PSI não é empregado como métrica de desempenho do classificador, mas como instrumento auxiliar para evidenciar deslocamentos distribucionais entre janelas temporais. Assim, seu papel é identificar mudanças nas proporções dos atributos discretizados, como *bytes_bin*, *packets_bin* e *duration_bin*, em comparações como T1–T2, T1–T3 e T2–T3. Essa análise permite verificar se o tráfego benigno observado em períodos posteriores mantém distribuição semelhante àquela usada como referência no treinamento.

Como complemento ao PSI, emprega-se a Distância de Wasserstein para quantificar a magnitude do afastamento entre distribuições de atributos observadas em diferentes janelas temporais. Diferentemente de medidas baseadas apenas em divergência de proporções, a Distância de Wasserstein considera a geometria dos valores observados e pode ser interpretada como o esforço necessário para transformar uma distribuição em outra. Essa propriedade é especialmente útil para atributos contínuos ou discretizados com sentido ordinal, como volume de bytes, quantidade de pacotes e duração do fluxo.

Em estudo recente sobre *datasets* de NIDS, Layeghy et al. utilizaram a primeira Distância de Wasserstein para quantificar diferenças entre distribuições de atributos de tráfego benigno em *datasets* sintéticos e tráfego real de redes de produção, após análise visual por Função de Distribuição Acumulada e *box plots* (11). Essa aplicação é aderente ao objetivo desta pesquisa, pois permite avaliar se janelas temporais distintas apresentam deslocamentos mensuráveis nas distribuições dos atributos usados pelo detector.

Desse modo, PSI e Distância de Wasserstein são utilizados de forma complementar. O PSI oferece uma medida sintética da mudança de proporções entre intervalos, enquanto a Distância de Wasserstein fornece uma estimativa da magnitude do deslocamento de massa entre distribuições. As duas medidas são calculadas sobre os fluxos rotulados como benignos. Dessa forma, o diagnóstico procura caracterizar mudanças no tráfego legítimo sem atribuir ao comportamento benigno os efeitos introduzidos pelos ataques controlados. Sua interpretação deve ser feita em conjunto com as métricas de desempenho, a sensibilidade do *cutoff* e o perfil dos falsos positivos, evitando-se tratar o diagnóstico distribucional como prova isolada de *concept drift*.

3.1.2.2 Delimitação em relação à validação cruzada

Cabe ainda esclarecer que a validação cruzada não constitui o eixo principal de avaliação do *dataset*. Técnicas como *K-Fold* aleatório pressupõem, em maior ou menor grau, independência e distribuição aproximadamente idêntica entre as amostras, hipótese frágil no domínio de tráfego de rede e em cenários sujeitos a *dataset shift* (56, 60). Além disso, particionamentos aleatórios tendem a desconsiderar a ordem temporal dos dados, podendo permitir que padrões de janelas futuras influenciem indiretamente a avaliação do modelo.

Por essa razão, quando empregada, a validação cruzada deve ser interpretada apenas como análise complementar de robustez ou de apoio à calibração do ponto operacional, preferencialmente em formato temporal bloqueado ou *rolling-origin* (57, 58, 61). Ainda assim, ela não substitui a avaliação temporal final em T3, que permanece como o teste mais representativo do comportamento do detector em uso prospectivo. Nesta pesquisa, portanto, a validade experimental é sustentada principalmente pela separação cronológica T1–T2–T3 e pelo diagnóstico distribucional complementar baseado em PSI e Distância de Wasserstein.

3.2 ARQUITETURA

A Seção 3 definiu a metodologia em nível geral. Esta seção descreve sua implementação no estudo de caso conduzido. A arquitetura apresentada reúne os serviços, máquinas virtuais, configurações e scripts utilizados para executar a coleta, a preparação dos dados, a rotulagem e a avaliação experimental. Esses componentes operacionalizam a metodologia proposta, mas não a restringem ao EFC.

A implementação foi organizada em três serviços principais. O serviço `netflow-collector` é responsável pela recepção dos registros NetFlow/IPFIX e pela conversão dos fluxos coletados para arquivos CSV. O serviço `attacker` executa os ataques controlados conforme o plano experimental definido em `configs/plan_attacks.yaml`. No estudo de caso, o serviço `efc-runner` executa o processamento dos dados e a instanciação do EFC. A coleta, a discretização, a rotulagem assistida e a divisão temporal permanecem conceitualmente separadas do modelo e podem fornecer dados para outros detectores. A Tabela 3.1 resume os principais aspectos técnicos desses componentes.

Tabela 3.1: Detalhamento dos serviços conteinerizados.

Serviço	Imagem base	Componentes e configuração	Função, recursos e dependências
`netflow-collector`	`debian:stable-slim`	Utiliza `nfdump`, `tcpdump`, `coreutils`, `tzdata`, `zip`. Expõe a porta `9995/udp`. Volumes: `data/flows`, `data/logs`, `data/csv` e `configs:ro`.	Responsável pela recepção e rotação dos fluxos NetFlow/IPFIX.
`attacker`	`kali-rolling`	Utiliza `python3-pip`, `python3-dateutil`, `nmap`, `hping3`, `git`, `pyyaml` e `Slowloris`. Executa `python3 orchestrator.py` com base em `plan_attacks.yaml`. Volumes: `data/logs` e `configs`.	Responsável pela execução programada dos ataques controlados. Atua de forma coordenada com a janela de coleta e não expõe porta de serviço.
`efc-runner`	`python:3.11-slim`	Utiliza `gcc`, `g++`, `git`, `curl`, `tzdata`, além de `pandas`, `numpy`, `scikit-learn`, `pyyaml`, `rich`, `matplotlib`, `scipy` e `seaborn`. Clona o pacote EFC. Volumes: `pipeline`, `configs`, `data` e `meta`.	Executa o processamento comum do *pipeline* e, no estudo de caso, o treino, a predição e a avaliação do EFC. Limite de `0.70 CPU` e `16G RAM`; reserva de `0.25 CPU` e `4G RAM`. Depende de `netflow-collector`.

A arquitetura recebe registros NetFlow/IPFIX, planos versionados de coleta e ataque, parâmetros de preparação e evidências geradas durante a execução. O processamento produz arquivos de fluxos, atributos preparados, mapas de discretização, rótulos rastreáveis, recortes temporais e relatórios de avaliação. As etapas de coleta, preparação, rotulagem e organização temporal não dependem do algoritmo empregado. O modelo de detecção utiliza os artefatos produzidos por essas etapas e acrescenta suas próprias saídas de treino, predição e avaliação.

Para a execução do experimento, foram disponibilizadas duas máquinas virtuais em ambiente Proxmox Virtual Environment (PVE), cada uma com 4 vCPUs e 16 GB de RAM. Uma delas foi utilizada como host de coleta e processamento, executando os serviços `netflow-collector` e `efc-runner`; a outra foi destinada à execução do serviço `attacker`. No ponto de coleta da rede de produção, o exportador de fluxos `softflowd` foi configurado para encaminhar os registros NetFlow/IPFIX ao host responsável pelo serviço `netflow-collector`. Definiu-se explicitamente apenas o timeout geral em 300 segundos, mantendo-se os demais nos valores padrão do `softflowd`, a saber, timeout TCP de 3600 segundos, UDP de 300 segundos, TCP-FIN de 300 segundos, TCP-RST de 120 segundos e *maxlife* de uma semana. Como o timeout geral governa apenas os fluxos que não são TCP nem UDP, os fluxos TCP de longa duração permanecem ativos segundo o timeout específico de protocolo e o *maxlife*, o que explica a presença de fluxos com duração de várias horas no *dataset*. Esses parâmetros determinam a granularidade dos registros exportados e afetam diretamente os atributos `duration_bin`, `bytes_bin` e `packets_bin`

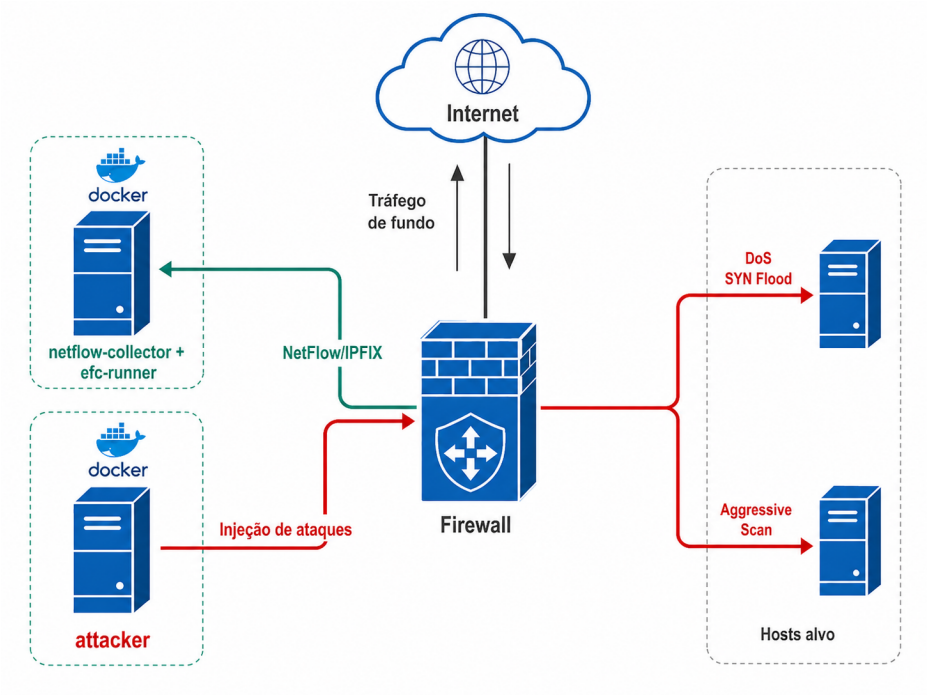


Figura 3.6: Topologia do experimento proposto

utilizados pelo EFC.

O serviço `netflow-collector` recebe registros NetFlow pelo daemon `nfcapd` na porta UDP 9995 e converte os arquivos de fluxo coletados em arquivos CSV, constituindo a base para as etapas posteriores de preparação do *dataset*. De acordo com a documentação do projeto, os fluxos brutos são armazenados em `data/flows`, os registros de log em `data/logs` e os arquivos CSV exportados em `data/csv`. O comportamento da coleta é controlado por `configs/plan_collect.yaml`, que define a janela temporal do experimento; o coletor aguarda o instante de início, encerra automaticamente ao final da janela e, caso não haja horário de término definido, permanece em execução contínua até interrupção manual.

A Figura 3.6 apresenta a topologia lógica do ambiente experimental. O cenário é composto por um host de coleta e processamento, no qual são executados os serviços `netflow-collector` e `efc-runner`; por um host atacante, responsável pela execução controlada dos cenários maliciosos; por um dispositivo central de rede; e pelos hosts alvo. Essa organização permite observar, de forma integrada, o tráfego legítimo da rede e o tráfego decorrente da injeção de ataques, preservando os metadados necessários para a construção, rotulagem e avaliação do *dataset*.

A execução dos ataques é coordenada pelo serviço `attacker`. O plano de agendamento dos cenários de ataque é realizado pelo arquivo `configs/plan_attacks.yaml`. Esse arquivo reúne informações como o identificador do ataque, os horários de início e término, os endereços de origem e destino, o protocolo, a porta de destino, o tipo de ataque, a ferramenta empregada e os respectivos parâmetros de execução. O disparo dos cenários é conduzido pelo script `orchestrator.py`, permitindo sincronizar

a geração de eventos maliciosos com a janela de coleta e, posteriormente, reutilizar esses metadados na etapa de rotulagem assistida.

Na etapa de preparação, o script `build\_features.py` transforma os CSVs brutos em uma representação discreta de fluxos e produz mapas de discretização e codificação categórica. Em seguida, o script `label\_flows.py` utiliza o plano de ataques e as evidências disponíveis para realizar a rotulagem assistida e organizar os subconjuntos T1, T2 e T3. Esses artefatos constituem a saída reutilizável do *pipeline* de construção do *dataset*.

No estudo de caso, o EFC é treinado com os fluxos benignos de T1. O ponto operacional é calibrado com T2 e aplicado a T3 no teste prospectivo. A execução produz predições, métricas, histogramas de energia e matrizes de confusão. Outro detector pode substituir o EFC nessa etapa desde que utilize os subconjuntos e a regra temporal definidos pela metodologia.

4 RESULTADOS EXPERIMENTAIS

Este capítulo apresenta os experimentos conduzidos para avaliar a metodologia proposta de construção e validação temporal de *datasets* de fluxo para NIDS. A análise considera a caracterização do cenário experimental, a composição do *dataset* gerado, a configuração do EFC, a avaliação de desempenho em janelas temporais distintas, a sensibilidade do *cutoff*, a estabilidade distribucional do tráfego benigno, o perfil dos falsos positivos, a incerteza estatística por *bootstrap* e a comparação com modelos *one-class* alternativos. O objetivo é verificar não apenas a capacidade de detecção do classificador, mas também sua robustez diante de deslocamentos temporais no tráfego benigno.

4.1 CENÁRIO EXPERIMENTAL

O experimento foi executado sobre a arquitetura descrita no capítulo anterior, baseada em serviços conteinerizados para coleta de fluxos, execução controlada de ataques e processamento analítico. O serviço `netflow-collector` recebeu os registros NetFlow/IPFIX exportados pela infraestrutura monitorada; o serviço `attacker` executou os ataques controlados conforme o plano experimental; e o serviço `efc-runner` concentrou as etapas de processamento, discretização, treinamento, predição e avaliação do EFC.

O tráfego foi coletado em uma rede governamental em operação normal, com entre 100 e 300 hosts ativos durante o período experimental. O perfil benigno observado foi composto predominantemente por navegação web, acesso a serviços internos e comunicações administrativas, refletindo padrões típicos de um ambiente de produção. Essa caracterização é relevante porque os modelos foram treinados e avaliados sobre tráfego benigno genuíno, oriundo de uma rede operacional real, o que constitui um dos requisitos centrais para a validade empírica da metodologia proposta.

A janela de coleta foi definida no arquivo `configs/plan_collect.yaml`, cobrindo o período de 23/09/2025 às 09:45 até 26/09/2025 às 21:00, no fuso horário `America/Sao_Paulo`. Ao todo, foram observadas aproximadamente 83 horas e 15 minutos de tráfego em ambiente de produção, com preservação dos arquivos `nfcapd`.

Embora essa janela de coleta seja suficiente para demonstrar o funcionamento do protocolo de separação temporal e permitir a análise de deslocamentos entre T1, T2 e T3, ela não captura ciclos operacionais mais longos, como variações semanais, finais de semana, mudanças sazonais, períodos de manutenção ou eventos excepcionais da rede. Assim, os deslocamentos observados devem ser interpretados como evidência de instabilidade temporal de curto prazo no ambiente estudado, e não como caracterização exaustiva da deriva em redes de produção ao longo de períodos prolongados.

O volume analisado compreendeu 657.550 fluxos rotulados em `meta/labels/labels.csv`. Desses, 640.638 fluxos foram utilizados nos recortes temporais T1, T2 e T3, enquanto 16.912 ficaram fora das

janelas de avaliação. Em termos de tráfego agregado registrado nos arquivos CSV, foram observados aproximadamente 537,5 GB e 712,2 milhões de pacotes. Esses valores referem-se ao volume de tráfego da rede monitorada, e não ao volume de registros NetFlow recebidos pelo coletor. Cada registro NetFlow ocupa tipicamente entre 50 e 100 bytes, de modo que o *throughput* efetivo do serviço `netflow-collector` permaneceu na ordem de poucos Mbps, bem dentro da capacidade do contêiner com os recursos alocados.

Tabela 4.1: Recortes temporais utilizados no experimento.

Split	Janela temporal	Uso no experimento	Amostras
T1	23/09/2025 09:45 a 24/09/2025 09:45	Treinamento *one-class* e definição do limiar de referência	442.713
T2	25/09/2025 22:00 a 26/09/2025 08:00	Calibração do ponto operacional e avaliação intermediária	82.066
T3	26/09/2025 10:00 a 26/09/2025 20:00	Teste prospectivo com o ponto operacional definido em T2	115.859

A integridade das coletas foi verificada antes da divisão temporal por meios complementares. A inspeção dos arquivos `nfcapd` confirmou continuidade temporal, com ausência de lacunas nos *timestamps* dos arquivos binários durante as 83h15min de observação. A ausência de ataques em T1 foi confirmada por inspeção do `configs/plan_attacks.yaml`. Todos os quatro cenários controlados (A01–A04) foram agendados para 26/09/2025, inteiramente fora da janela de T1, o que garante a pureza da partição de treinamento *one-class*. Os rótulos produzidos por `label_flows.py` foram verificados por rastreabilidade. O campo `evidence` em `meta/labels/labels.csv` registra, para cada fluxo malicioso, o identificador do ataque que motivou o rótulo, permitindo auditoria individual de cada classificação à sua origem no plano de ataques.

4.2 CARACTERIZAÇÃO DO *DATASET* GERADO

O *dataset* preserva um desequilíbrio de classes compatível com cenários reais de detecção de intrusão, nos quais eventos maliciosos representam uma fração minoritária do tráfego total. No conjunto rotulado completo, 632.442 fluxos foram classificados como benignos e 25.108 como maliciosos, correspondendo a 3,82% de amostras maliciosas. Nos períodos T2 e T3, a proporção de tráfego malicioso foi maior em razão da injeção controlada de ataques durante as respectivas janelas de coleta.

Tabela 4.2: Distribuição de classes por recorte temporal.

Split	Benignos	Maliciosos	Percentual malicioso
T1	442.713	0	0,00%
T2	71.225	10.841	13,21%
T3	101.592	14.267	12,31%
Total rotulado	632.442	25.108	3,82%

A diversidade de protocolos foi representada por três categorias codificadas: TCP, UDP e protocolo indefinido. O tráfego foi dominado por TCP, mas preservou fluxos UDP e registros não classificados. A diversidade de serviços também foi mantida por meio do atributo `svc_port_bucket`, com 101 categorias de portas ou serviços. Entre as portas mais frequentes observaram-se `443`, `80`, `8080`, `7680`, `53`, `3128`, `389` e `445`. A presença de tráfego criptografado foi inferida de forma aproximada a partir de portas usualmente associadas a TLS ou serviços tunelados, com destaque para `443`; essa inferência, entretanto,

deve ser interpretada como indicador indireto operacional baseado em metadados de fluxo, e não como confirmação criptográfica do conteúdo da sessão.

Os ataques controlados foram definidos no arquivo `configs/plan_attacks.yaml`. A Tabela 4.3 resume os cenários injetados, suas respectivas janelas, ferramentas e alvos.

Tabela 4.3: Plano de ataques controlados executados no experimento.

Ataque	Janela	Tipo	Ferramenta	Alvo
A01	26/09/2025 02:00–02:30	`aggressive_scan`	`nmap -sV -O -A -p-`	`192.168.110.0/24`
A02	26/09/2025 04:30–05:00	`syn_flood`	`hping3 -S -p 8080 -i u1000`	`192.168.106.20:8080`
A03	26/09/2025 15:00–16:00	`aggressive_scan`	`nmap -sV -O -A -p-`	`192.168.110.0/24`
A04	26/09/2025 17:00–18:00	`syn_flood`	`hping3 -S -p 8080 -i u1000`	`192.168.106.20:8080`

A escolha dos ataques `syn_flood` e `aggressive_scan` seguiu um critério de contraste metodológico. Os dois cenários produzem padrões distintos nos atributos disponíveis em registros de fluxo e permitem avaliar se o protocolo proposto revela diferenças de desempenho entre atividades maliciosas de naturezas diferentes.

O `syn_flood`, executado com `hping3`, representa um ataque de impacto baseado no envio intenso de pacotes TCP SYN ao mesmo serviço. Esse padrão tende a alterar de forma acentuada o número de pacotes, o volume de bytes, a duração e a direção das comunicações. Portanto, ele permite examinar o comportamento do detector diante de uma anomalia volumétrica com expressão evidente no nível de cada fluxo.

O `aggressive_scan`, executado com `nmap`, representa uma atividade de reconhecimento que distribui tentativas de conexão entre múltiplos hosts, portas e serviços. Cada fluxo individual pode apresentar baixo volume e curta duração, mesmo quando o conjunto das sondagens caracteriza uma atividade maliciosa relevante. Esse cenário permite verificar se o detector identifica padrões que permanecem próximos do tráfego benigno quando observados isoladamente.

A seleção não pretende cobrir exaustivamente o espaço de ataques contra redes. Seu objetivo é avaliar a metodologia diante de dois perfis contrastantes. O primeiro apresenta alta expressão volumétrica por fluxo. O segundo apresenta baixa expressão individual e maior dependência da correlação entre fluxos. Essa distinção permite interpretar os resultados por tipo de ataque e evita que métricas agregadas ocultem limitações do detector.

A seleção também pode ser interpretada no vocabulário do MITRE ATT&CK. O `aggressive_scan` corresponde à técnica T1046, *Network Service Discovery*, inserida na tática de reconhecimento. O `syn_flood` corresponde à subtécnica T1498.001, *Direct Network Flood*, inserida na tática de impacto. Esse enquadramento relaciona os cenários do experimento a comportamentos reconhecidos na literatura de segurança e favorece a rastreabilidade entre os ataques executados, os rótulos do *dataset* e as categorias empregadas em NIDS (20, 62, 63).

Na rotulagem final, os ataques foram agrupados nas classes `syn_flood` e `aggressive_scan`. Em T2, foram identificados 10.162 fluxos `syn_flood` e 679 fluxos `aggressive_scan`. Em T3, esses

quantitativos foram de 13.609 e 658 fluxos, respectivamente. As evidências dos *scans* utilizaram correspondência por janela temporal e CIDR. Para `syn_flood`, entretanto, a evidência salva aparece como `exact=1;attack:A02` tanto em T2 quanto em T3. Esse registro preserva a identificação do tipo de ataque, mas não distingue com o mesmo grau de granularidade as campanhas A02 e A04 no campo de evidência. Esse aspecto não altera a avaliação binária entre tráfego benigno e malicioso, nem compromete a análise agregada por tipo de ataque quando `syn_flood` é tratado como uma única classe. Entretanto, recomenda-se cautela em inferências sobre diferenças entre as campanhas A02 e A04 consideradas isoladamente. Por essa razão, os resultados principais desta dissertação concentram-se nos níveis binário e por tipo de ataque.

A quantidade de fluxos rotulados como `aggressive_scan` (679 em T2 e 658 em T3) é compatível com a agregação característica do NetFlow/IPFIX, em que as numerosas tentativas de conexão geradas pela varredura `nmap` sobre a faixa-alvo são condensadas em um conjunto reduzido de registros de fluxo de baixa intensidade, em vez de um fluxo por sonda. Essa contagem foi verificada de forma independente nos artefatos do pacote reprodutível e mostrou-se estável; o *recall* praticamente nulo observado para essa classe não decorre, portanto, de sub-rotulagem, mas da representação desses fluxos no espaço de atributos, conforme detalhado na análise energética da Seção 4.4.1.

4.3 CONFIGURAÇÃO DO MODELO E PROTOCOLO TEMPORAL DE AVALIAÇÃO

O EFC foi utilizado em configuração *one-class*, treinado exclusivamente com fluxos benignos de T1. O *wrapper* implementado em `pipeline/src/efc_model.py` cria rótulos `y=0` para todas as amostras de treinamento, define `base_class=0` e instancia o `EnergyBasedFlowClassifier` com `n_bins=30` e `cutoff_quantile=0.95`. A decisão do modelo é baseada na comparação entre a energia atribuída a cada fluxo e um limiar de decisão, denominado *cutoff*; energias iguais ou superiores ao limiar são classificadas como anômalas.

O protocolo atribui funções não intercambiáveis aos três recortes temporais. T1 contém somente tráfego benigno e é empregado para aprender o perfil de normalidade e obter o limiar de referência. Como não contém ataques rotulados, T1 não fornece informação suficiente para escolher um ponto operacional orientado por F1-score. T2 constitui o primeiro período posterior com exemplos benignos e maliciosos conhecidos e permite calibrar o compromisso entre detecção e falsos positivos. T3 permanece fora dessa escolha e recebe o ponto operacional definido em T2.

Essa organização preserva a ordem em que dados e rótulos estariam disponíveis durante uma implantação. O desempenho em T3 mede a capacidade de transferir para dados futuros uma decisão tomada com informações anteriores. Os rótulos de T3 são utilizados somente depois da predição para calcular as métricas. A busca do melhor F1-score em T3 permanece como análise retrospectiva de sensibilidade e não integra o resultado prospectivo principal.

Esse controle temporal constitui um diferencial metodológico porque evita que informações da janela final influenciem a configuração avaliada. Uma partição aleatória ou uma calibração realizada diretamente

em T3 poderia produzir uma estimativa otimista e ocultar a instabilidade do detector diante da evolução do tráfego.

As variáveis utilizadas pelo modelo são apresentadas na Tabela 4.4. O mapa de discretização registrou 29 intervalos para `bytes`, 29 para `packets` e 30 para `duration`. O mapa categórico registrou 3 categorias de protocolo, 101 buckets de serviço/porta, 4 direções de tráfego, 3.459 sub-redes /24 de origem e 3.648 sub-redes /24 de destino.

Tabela 4.4: Atributos utilizados pelo EFC no experimento.

Ordem	Atributo	Tipo de representação
1	`bytes_bin`	Volume de bytes discretizado
2	`packets_bin`	Quantidade de pacotes discretizada
3	`duration_bin`	Duração do fluxo discretizada
4	`proto_enc`	Protocolo codificado
5	`svc_port_bucket_enc`	Bucket de porta ou serviço codificado
6	`direction_enc`	Direção do tráfego codificada
7	`src_subnet24_enc`	Sub-rede /24 de origem codificada
8	`dst_subnet24_enc`	Sub-rede /24 de destino codificada

Para a análise desta dissertação, distinguem-se dois estados experimentais. O primeiro corresponde aos artefatos utilizados na publicação original, em que T3 estava consistente, mas T2 apresentava divergência entre *tuning*, predição e avaliação. O segundo corresponde ao pacote congelado `runs/efc_repro_20260427_053541`, regenerado ponta a ponta com os mesmos planos, a mesma estratégia de rotulagem e a mesma configuração base do modelo, mas com congelamento de versões, *hashes*, modelo, *cutoffs* e relatórios. Por essa razão, os resultados do pacote congelado são tratados como referência principal da dissertação.

No pacote reprodutível congelado, o limiar base associado a `q=0.95` em T1 foi 180,630. Também foram registrados os limiares correspondentes ao mesmo quantil nas distribuições de energia de T2 e T3, respectivamente 209,423 e 213,680, como referência descritiva para comparação entre janelas. Em seguida, o ajuste fino do limiar foi realizado por busca em grade no intervalo $q \in [0.90, 0.999]$, com passo de 0,001 e objetivo F1-score. O melhor ponto por F1 foi `q=0.960`, com *cutoff* 209,913, para T2; e `q=0.916`, com *cutoff* 209,385, para T3.

Tabela 4.5: Limiar de referência e pontos de melhor F1 no pacote reprodutível.

Split	Cutoff em q=0,95	Melhor q por F1	Cutoff por melhor F1
T1	180,630	–	–
T2	209,423	0,960	209,913
T3	213,680	0,916	209,385

Os pontos de melhor F1 em T2 e T3 possuem estatutos metodológicos distintos. O ponto de T2 integra o protocolo prospectivo porque é selecionado antes do teste final. O ponto de T3 utiliza os rótulos da própria janela e serve apenas para quantificar a sensibilidade do detector ao limiar. As comparações prospectivas aplicam em T3 o ponto operacional definido em T2.

4.4 RESULTADOS DE DETECÇÃO DO EFC

Os resultados desta seção seguem a distinção de regimes definida na Seção 4.3. A Tabela 4.6 apresenta a avaliação prospectiva estrita do EFC, com o ponto de operação calibrado em T2 e aplicado a T3 sem uso dos rótulos dessa janela para seleção do limiar. Em seguida, a Tabela 4.7 reúne os pontos de melhor F1 por janela, que em T3 constituem análise retrospectiva de sensibilidade do ponto de operação. Em T2, por ser a janela de calibração, os dois regimes coincidem.

Tabela 4.6: Desempenho prospectivo estrito do EFC.

Split (regime)	Precisão	Recall	F1-score	FPR	TN	FP	FN	TP
T2 (calibração)	0,781	0,937	0,852	4,00%	68.375	2.850	679	10.162
T3 (prospectivo)	0,616	0,954	0,749	8,34%	93.117	8.475	658	13.609

Na avaliação prospectiva, o EFC manteve *recall* elevado nos dois períodos, mas apresentou redução de precisão e aumento do FPR bruto em T3. Essa variação não deve ser interpretada isoladamente como perda de especificidade sobre o tráfego legítimo. A análise de proveniência da Seção 4.7.1 mostra que o valor agregado de T3 inclui tráfego de retorno induzido pelo `syn_flood` e mantido sob rótulo benigno.

A Tabela 4.7 apresenta os pontos de melhor F1 em cada janela. Em T3, esse resultado utiliza os rótulos da própria janela e tem finalidade retrospectiva. O resultado prospectivo da dissertação é o obtido com o *cutoff* calibrado em T2 e aplicado a T3 sem uso prévio de seus rótulos.

Tabela 4.7: Desempenho global do EFC nos pontos de melhor F1 por janela.

Split	Precisão	Recall	F1-score	FPR	TN	FP	FN	TP
T2	0,781	0,937	0,852	4,00%	68.375	2.850	679	10.162
T3	0,615	0,954	0,748	8,40%	93.058	8.534	658	13.609

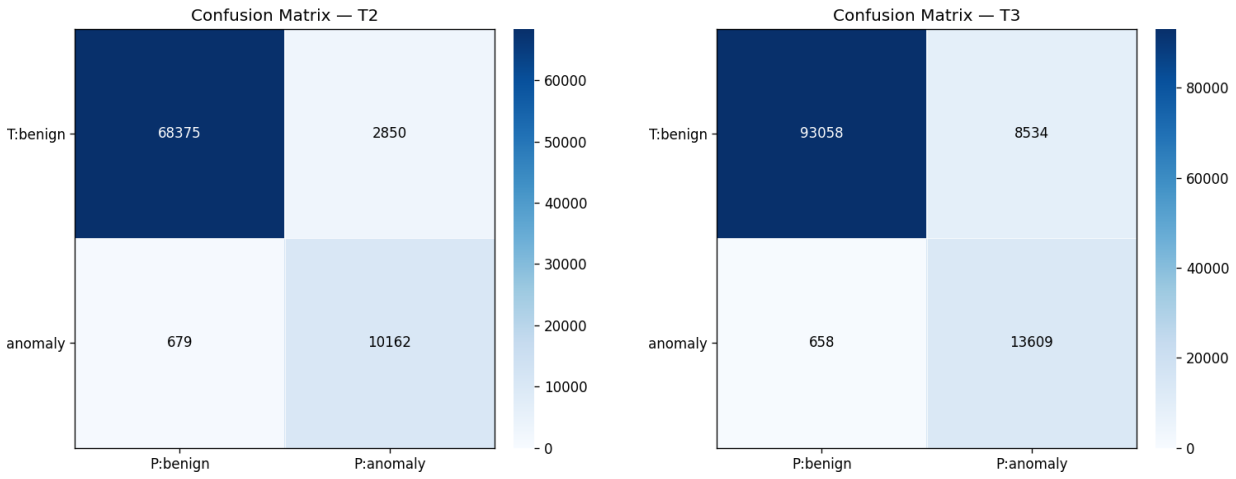
As matrizes de confusão da Figura 4.1 complementam as métricas ao mostrar a distribuição dos acertos e erros. Os falsos negativos permaneceram baixos em termos globais, enquanto a maior quantidade de falsos positivos em T3 explica a redução de precisão e o aumento do FPR bruto.

A análise por tipo de ataque, apresentada na Tabela 4.8, demonstra que o resultado agregado é fortemente influenciado pelo predomínio quantitativo do ataque `syn_flood`. Esse ataque foi detectado integralmente em T2 e T3, enquanto o `aggressive_scan` praticamente não foi detectado. Assim, o alto *recall* global não deve ser interpretado como desempenho homogêneo para todos os padrões de ataque.

Tabela 4.8: Desempenho do EFC por tipo de ataque.

Split	Ataque	Suporte	Detectados	Não detectados	Recall
T2	`syn_flood`	10.162	10.162	0	1,000
T2	`aggressive_scan`	679	0	679	0,000
T3	`syn_flood`	13.609	13.609	0	1,000
T3	`aggressive_scan`	658	0	658	0,000

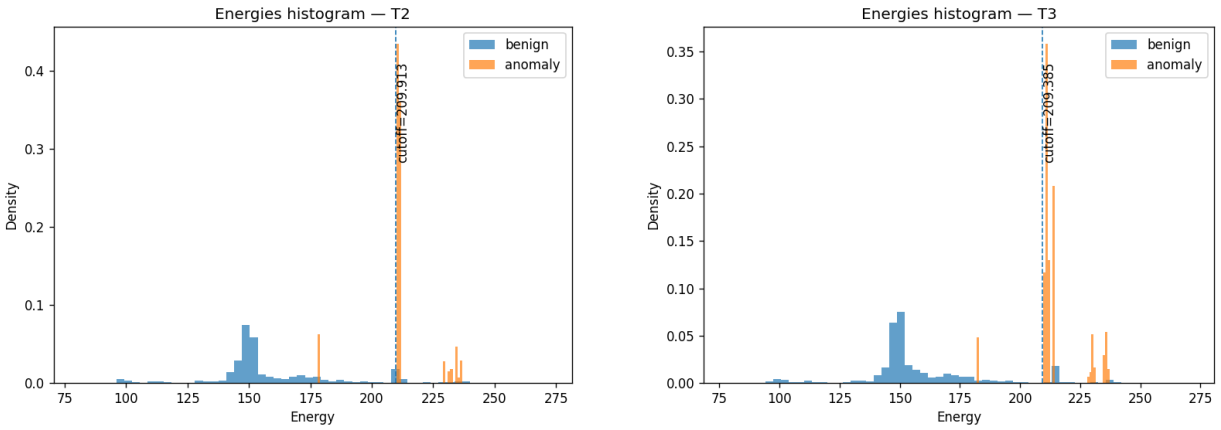
Os histogramas de energia permitem observar como fluxos benignos e maliciosos se distribuem em relação ao limiar de decisão. A Figura 4.2 mostra a separação parcial entre as classes, mas também evidencia sobreposição nas caudas de energia. Essa sobreposição explica simultaneamente a ocorrência de



(a) T2 (b) T3

Figura 4.1: Matrizes de confusão do EFC nos recortes temporais T2 e T3, no ponto de melhor F1 de cada janela.

falsos positivos, quando fluxos benignos cruzam o limiar de anomalia, e de falsos negativos, quando fluxos maliciosos permanecem em regiões de energia compatíveis com a normalidade aprendida.



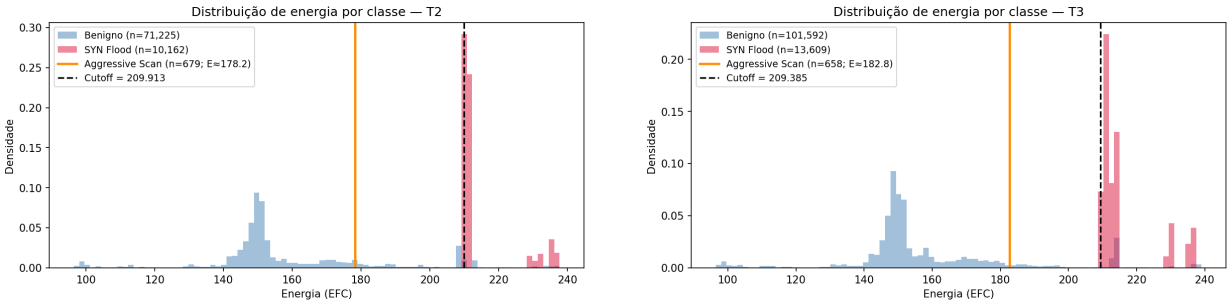
(a) T2 (b) T3

Figura 4.2: Distribuição das energias atribuídas pelo EFC aos fluxos benignos e maliciosos em T2 e T3.

4.4.1 Análise energética por tipo de ataque

Para investigar a explicação mais provável para o *recall* praticamente nulo observado para o `aggressive_scan`, analisou-se a distribuição da energia atribuída pelo EFC às três classes de tráfego presentes em T2 e T3. A Figura 4.3 sobrepõe as distribuições e a Tabela 4.9 resume as estatísticas por classe, incluindo a fração de fluxos abaixo do *cutoff*.

Os resultados indicam que a diferença de desempenho decorre da separabilidade no espaço de energia. O `syn_flood` permaneceu acima do *cutoff* nos dois recortes, enquanto o `aggressive_scan` se concentrou dentro da distribuição benigna e teve todos os seus fluxos classificados como benignos. O padrão se repetiu em T2 e T3.



(a) T2 (b) T3

Figura 4.3: Distribuição de energia por classe (benigno, `syn_flood` e `aggressive_scan`) em T2 e T3, com o limiar de decisão (*cutoff*).

Tabela 4.9: Estatísticas de energia por classe e fração de fluxos abaixo do *cutoff*.

Split	Classe	n	Mediana	P95	% < cutoff
T2	Benigno	71.225	150,9	209,4	96,0%
T2	`syn_flood`	10.162	211,4	234,8	0,0%
T2	`aggressive_scan`	679	178,2	178,2	100,0%
T3	Benigno	101.592	150,9	213,7	91,6%
T3	`syn_flood`	13.609	212,3	236,1	0,0%
T3	`aggressive_scan`	658	182,8	182,8	100,0%

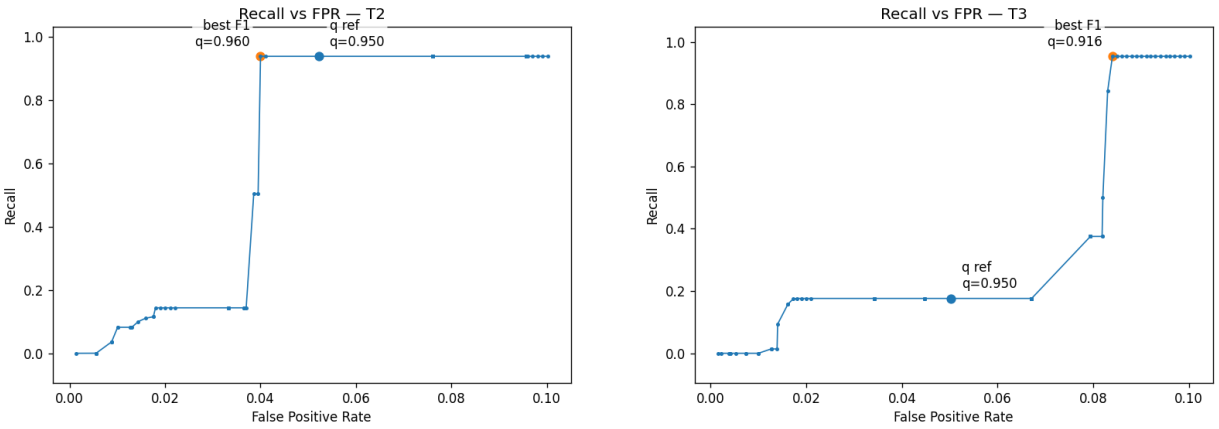
Esse comportamento é compatível com uma limitação da representação discretizada. Após o mapeamento em *bins*, os fluxos de varredura convergem para poucas configurações e recebem energias próximas às do tráfego benigno. Um *cutoff* capaz de capturá-los também sinalizaria grande quantidade de fluxos legítimos. Como não foi realizada uma ablação sistemática do número de *bins*, essa interpretação permanece como hipótese explicativa sustentada pela análise energética. O contraste com o PCA e a sensibilidade ao ponto operacional são discutidos na Seção 5.3.

4.5 SENSIBILIDADE DO *CUTOFF*

A análise de sensibilidade do *cutoff* foi realizada por busca em grade sobre quantis no intervalo de 0,900 a 0,999, com passo de 0,001. Para cada ponto, foram recalculados o limiar numérico, a precisão, o *recall*, o F1-score e a taxa de falsos positivos. Essa análise permite avaliar como diferentes pontos operacionais alteram o compromisso entre sensibilidade e especificidade.

A sensibilidade do limiar é sintetizada pela relação entre *recall* e FPR, por ser a visualização mais diretamente associada ao compromisso operacional de um NIDS. As curvas Precision-Recall e ROC foram preservadas como artefatos complementares, mas não são reproduzidas nesta seção para evitar redundância visual.

A Figura 4.4 mostra a relação entre *recall* e taxa de falsos positivos para T2 e T3. Em T2, o detector mantém *recall* elevado em uma região mais ampla de limiares. Em T3, a redução do FPR por meio de quantis mais altos provoca queda rápida do *recall*, indicando que a busca por menor taxa de alarmes falsos pode comprometer substancialmente a capacidade de detecção.



(a) T2 (b) T3

Figura 4.4: Relação entre *recall* e taxa de falsos positivos na análise de sensibilidade do *cutoff*.

A região de quantis mais altos reduziu significativamente o FPR, mas ao custo de perda severa de *recall*. Quando `q` se aproxima de 0,985, o FPR cai para cerca de 1,62% em ambos os splits, mas o *recall* fica abaixo de 0,16. No extremo `q=0.999`, o detector praticamente deixa de sinalizar anomalias. Portanto, a região de menor FPR não representa, por si só, uma condição operacional adequada para NIDS, pois sacrifica excessivamente a capacidade de detecção.

4.6 ANÁLISE DE ESTABILIDADE DISTRIBUCIONAL

A análise de estabilidade distribucional foi conduzida por meio do *Population Stability Index* (PSI) e da Distância de Wasserstein, aplicados às *features* discretizadas do pipeline. O PSI foi utilizado para sintetizar a intensidade do deslocamento distribucional por atributo, enquanto a Distância de Wasserstein foi empregada como medida complementar da magnitude do deslocamento de massa entre bins. Em conjunto, essas medidas permitem evidenciar desalinhamento distribucional entre janelas temporais, sem tratá-lo como métrica de desempenho do classificador ou como prova isolada de *concept drift*. Cabe destacar que o diagnóstico distribucional foi computado exclusivamente sobre os fluxos rotulados como benignos de cada janela, isolando a deriva do tráfego legítimo da contaminação volumétrica introduzida pelos ataques. Assim, os valores elevados de PSI em `packets_bin` e `bytes_bin` refletem mudanças no perfil benigno entre janelas, e não o efeito direto do `syn_flood` sobre os atributos de volume.

A Tabela 4.10 apresenta os principais atributos que contribuíram para a deriva distribucional segundo o PSI.

Tabela 4.10: Principais atributos com maior deriva distribucional segundo o PSI.

Comparação	Maior deriva	PSI	Segunda maior	PSI	Terceira maior	PSI
T1 x T2	`packets_bin`	8,964	`bytes_bin`	3,535	`dst_subnet24_enc`	0,168
T1 x T3	`packets_bin`	0,881	`bytes_bin`	0,600	`dst_subnet24_enc`	0,116
T2 x T3	`packets_bin`	16,227	`bytes_bin`	6,991	`dst_subnet24_enc`	0,098

Os resultados indicam que a deriva se concentrou principalmente nas variáveis de volume, em especial `packets_bin` e `bytes_bin`. A comparação T2 x T3 apresentou a maior ruptura distribucional, tanto pelo PSI quanto pela Distância de Wasserstein, sugerindo mudança relevante na composição do tráfego benigno entre os dois períodos de avaliação.

Essa leitura exige cautela por duas razões. Primeiro, como mostra a auditoria de proveniência da Seção 4.7.1, a maior parte do aumento agregado do FPR em T3 decorre de tráfego de retorno do ataque, e não de deslocamento do benigno, de modo que o diagnóstico distribucional aqui apresentado, calculado apenas sobre fluxos benignos, é a base mais adequada para avaliar a deriva. Segundo, mesmo entre os fluxos benignos, a análise não sustenta a interpretação simplificada de que T3 seria apenas mais distante de T1 do que T2 em todas as dimensões. Em termos marginais, T1 x T2 exibiu deriva mais forte que T1 x T3 para as principais variáveis de volume, enquanto T2 x T3 revelou a maior ruptura entre períodos. Isso indica que o deslocamento do tráfego benigno em T3 não decorre apenas de uma maior distância global em relação ao treino de T1, mas de uma reorganização mais complexa da distribuição benigna entre os períodos de avaliação.

A Distância de Wasserstein confirmou deslocamentos relevantes de massa entre bins, especialmente em `packets_bin`, com valores de 4,94 em T1 x T2, 3,70 em T1 x T3 e 8,55 em T2 x T3. Assim, há evidência quantitativa de deriva temporal no tráfego benigno, mas sua manifestação é heterogênea entre pares de janelas e grupos de atributos.

4.7 PERFIL DOS FALSOS POSITIVOS

Para compreender o aumento agregado do FPR em T3, foi realizada uma análise dos falsos positivos por perfil, agregando protocolo, direção, bucket de serviço, porta de destino, sub-redes de origem e destino, além da faixa horária. O objetivo é verificar se os alarmes falsos se distribuem de forma difusa ou se estão concentrados em segmentos específicos do tráfego.

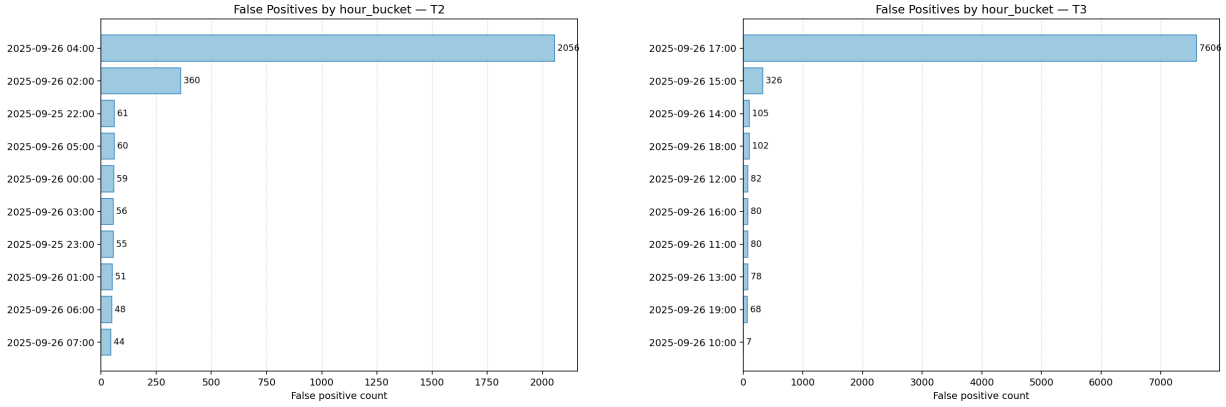
No corpo principal, a análise é apresentada de forma sintética por meio da Tabela 4.11, que consolida as dimensões mais informativas. Os gráficos completos por protocolo, direção, porta de destino, bucket de serviço e sub-rede foram preservados como artefatos complementares.

Tabela 4.11: Principais concentrações dos falsos positivos por dimensão de tráfego.

Split	Dimensão	Categoria dominante	FP	% FP
T2	Direção	`intra`	2.816	98,8%
	Origem	`192.168.106.0/24`	2.008	70,5%
	Destino	`192.168.102.0/24`	2.646	92,8%
	Faixa horária	`2025-09-26 04:00`	2.056	72,1%
	Bucket de serviço	`other`	2.008	70,5%
T3	Direção	`intra`	8.508	99,7%
	Origem	`192.168.106.0/24`	7.519	88,1%
	Destino	`192.168.102.0/24`	8.110	95,0%
	Faixa horária	`2025-09-26 17:00`	7.606	89,1%
	Bucket de serviço	`other`	7.514	88,0%

Os falsos positivos se concentraram em tráfego interno entre as sub-redes `192.168.106.0/24` e `192.168.102.0/24`. O padrão mostra que os alarmes não se distribuíram uniformemente pelo tráfego benigno e permaneceram associados a um contexto operacional específico.

A Figura 4.5 apresenta a distribuição dos falsos positivos por faixa horária. Diferentemente das demais dimensões, essa visualização foi mantida no corpo principal porque evidencia a coincidência temporal entre os falsos positivos e as janelas de ataque controlado. Em T2, a maior concentração ocorre às 04:00, próxima ao ataque A02; em T3, a maior concentração ocorre às 17:00, correspondente à janela do ataque A04.



(a) T2 (b) T3

Figura 4.5: Distribuição dos falsos positivos por faixa horária em T2 e T3.

O aumento do FPR decorreu de uma concentração de alarmes em tráfego associado ao contexto operacional do `syn_flood`. A auditoria da Seção 4.7.1 examina a origem desses fluxos.

4.7.1 Proveniência dos falsos positivos e tráfego de retorno do ataque

A concentração dos falsos positivos em tráfego interno e em janelas próximas aos ataques motivou uma auditoria de proveniência, conduzida sobre os artefatos congelados sem alteração de rótulos. Para cada fluxo benigno classificado como anomalia, verificou-se o envolvimento com os hosts do plano de ataque, a saber, o host atacante (`192.168.102.62`, origem de todos os fluxos maliciosos), o alvo do `syn_flood` (`192.168.106.20:8080`) e a faixa-alvo do `aggressive_scan` (`192.168.110.0/24`), além da sobreposição com as janelas temporais dos cenários A01–A04.

A Tabela 4.12 mostra que a maior parte dos falsos positivos corresponde ao retorno do alvo do `syn_flood` para o host atacante. Esses fluxos utilizam portas de destino efêmeras compatíveis com respostas *SYN-ACK* e *RST* e ocorrem dentro das janelas das campanhas A02 e A04. O tráfego benigno sem envolvimento dos hosts do contexto de ataque representa uma parcela minoritária dos alarmes.

A origem desse padrão está na regra de rotulagem adotada para ataques direcionados a serviço. O fluxo de ida do `syn_flood` satisfaz a correspondência entre o host atacante, o alvo e a porta 8080. O fluxo de retorno parte do alvo, tem como destino o host atacante e utiliza uma porta efêmera no lado de destino.

Tabela 4.12: Proveniência dos falsos positivos do EFC no ponto de melhor F1 por janela.

Split	FP total	Retorno alvo→atacante	Contexto de ataque	Benigno genuíno
T2	2.850	1.994 (70,0%)	2.606 (91,4%)	244 (8,6%)
T3	8.534	7.511 (88,0%)	8.223 (96,4%)	311 (3,6%)

Como essa direção não satisfaz a correspondência exata definida para a ida do ataque, ela permaneceu sob rótulo benigno no pacote congelado.

Esse retorno não representa uma comunicação legítima independente do ataque. Ele é gerado em resposta às tentativas TCP SYN enviadas ao alvo, ocorre entre os mesmos hosts envolvidos na campanha e concentra-se integralmente nas janelas do `syn_flood`. A diferença é relevante para a interpretação das métricas. Segundo o rótulo binário congelado, esses fluxos são contabilizados como falsos positivos. Segundo sua proveniência operacional, eles pertencem ao contexto do ataque e não ao tráfego benigno genuíno usado para avaliar a especificidade do detector.

A taxa agregada não constitui uma medida limpa de deriva do tráfego benigno porque é dominada pelo retorno do `syn_flood`. A Tabela 4.13 apresenta o FPR após a exclusão apenas do retorno e após a exclusão de todo o tráfego que envolve hosts do contexto de ataque. No ponto de melhor F1, o FPR benigno permaneceu praticamente estável entre T2 e T3.

Tabela 4.13: Taxa de falsos positivos bruta e após remoção do tráfego do contexto de ataque.

Split	Limiar	FPR bruto	Sem retorno	Sem contexto
T2	melhor F1 (209,913)	4,00%	1,30%	0,42%
T2	referência T1 (180,630)	14,91%	7,35%	3,69%
T3	melhor F1 (209,385)	8,40%	1,09%	0,36%
T3	T2 prospectivo (209,913)	8,34%	1,02%	0,36%
T3	referência T1 (180,630)	13,08%	6,14%	3,38%

A diferença entre os FPRs brutos de T2 e T3 não mede exclusivamente uma mudança no tráfego benigno. A evidência de deriva permanece sustentada pelo PSI, pela Distância de Wasserstein e pelo FPR depurado sob o limiar de referência de T1. Esse conjunto indica uma mudança moderada entre períodos comerciais equivalentes e separa esse efeito daquele introduzido pela rotulagem do tráfego de retorno.

4.8 INTERVALOS DE CONFIANÇA POR BOOTSTRAP

Para avaliar a estabilidade das métricas globais, foi aplicada reamostragem por *bootstrap*. Esse procedimento consiste em gerar múltiplas amostras artificiais a partir do próprio conjunto avaliado, por meio de sorteios com reposição. Nesta pesquisa, foram geradas 2.000 reamostragens para cada split avaliado. Em cada uma delas, foram recalculadas a matriz de confusão e as métricas precisão, *recall*, F1-score e FPR, mantendo-se, em cada janela, o limiar de melhor F1 reportado na Tabela 4.7, de modo que os intervalos de T3 referem-se ao regime retrospectivo de sensibilidade. A partir da distribuição empírica obtida para cada métrica, foram estimados intervalos de confiança de 95%, permitindo avaliar se os valores observados

permanecem estáveis diante de variações amostrais. A semente aleatória *seed*=42 foi fixada apenas para garantir a reprodutibilidade dos sorteios realizados no procedimento.

Tabela 4.14: Métricas globais do EFC com intervalos de confiança de 95% por *bootstrap* (regime retrospectivo em T3).

Split	Métrica	Estimativa pontual	IC95% inferior	IC95% superior
T2	Precisão	0,781	0,774	0,788
T2	Recall	0,937	0,933	0,942
T2	F1-score	0,852	0,848	0,857
T2	FPR	4,00%	3,86%	4,14%
T3	Precisão	0,615	0,608	0,621
T3	Recall	0,954	0,951	0,957
T3	F1-score	0,748	0,743	0,753
T3	FPR	8,40%	8,23%	8,57%

Os intervalos indicam que o *recall* permaneceu alto e estável nos dois períodos. Por outro lado, os intervalos de FPR bruto entre T2 e T3 ficaram separados, o que indica que a diferença agregada não é apenas flutuação amostral. Como mostra a auditoria de proveniência (Seção 4.7.1), essa diferença reflete sobretudo o tráfego de retorno do `syn_flood` contabilizado como falso positivo, e não um deslocamento do tráfego benigno genuíno, cujo FPR permaneceu estável entre as janelas.

4.9 COMPARAÇÃO COM MODELOS *ONE-CLASS* ALTERNATIVOS

A comparação reuniu mecanismos distintos de decisão. O *Isolation Forest* utiliza particionamento aleatório, o PCA utiliza erro de reconstrução e o EFC utiliza energia de Potts. *Isolation Forest* e PCA também apresentam uso recorrente em avaliações de NIDS e viabilidade computacional para o volume analisado (4, 3). Todos os modelos foram treinados com fluxos benignos de T1, calibrados em T2 e avaliados nos mesmos recortes. O One-Class SVM foi incluído como tentativa controlada e interrompido após o piloto em T2.

4.9.1 *Isolation Forest*

O *Isolation Forest* foi treinado sobre T1 benigno e avaliado em T2 e T3 com ajuste de *cutoff* por quantil. Os melhores pontos por F1 foram `q=0.933` em T2 e `q=0.940` em T3. A Tabela 4.15 apresenta a comparação com o EFC, usando o pacote reprodutível congelado como referência principal para os resultados do EFC.

Tabela 4.15: Comparação entre EFC e *Isolation Forest* no ponto de melhor F1 por janela (regime retrospectivo em T3).

Modelo	Split	Precisão	Recall	F1-score	FPR
EFC	T2	0,781	0,937	0,852	4,00%
Isolation Forest	T2	0,680	0,937	0,788	6,70%
EFC	T3	0,615	0,954	0,748	8,40%
Isolation Forest	T3	0,691	0,954	0,801	6,00%

Em T2, o EFC apresentou melhor equilíbrio entre sensibilidade e especificidade que o *Isolation Forest*. No regime retrospectivo de T3, o baseline obteve melhor desempenho agregado. A inversão reforça a necessidade de considerar o período avaliado, o regime de calibração e o desempenho por tipo de ataque.

A comparação por tipo de ataque qualifica essa leitura. Em ambos os recortes, EFC e *Isolation Forest* detectaram integralmente o `syn_flood` (*recall* 1,000), e nenhum dos dois detectou o `aggressive_scan` (*recall* 0,000 em T2 e em T3). Assim, o baseline não supera a principal limitação qualitativa do experimento, pois a fragilidade estrutural diante de ataques de baixa expressão volumétrica é compartilhada pelos dois detectores.

Os intervalos de confiança por *bootstrap* confirmam essa leitura. Em T2, os intervalos de precisão e FPR favorecem o EFC. Em T3, eles favorecem o *Isolation Forest* no regime retrospectivo.

4.9.2 Principal Component Analysis (PCA)

Também foi implementado um baseline *one-class* baseado em PCA, utilizando o erro médio quadrático de reconstrução no espaço padronizado como *score* de anomalia. O fluxo seguiu a mesma estratégia de isolamento adotada nos demais baselines, com scripts separados, artefatos próprios e ausência de impacto sobre o pipeline consolidado do EFC.

Inicialmente, o PCA foi treinado com `n_components=0.95`, mas essa configuração mostrou desempenho inadequado em T2. Em seguida, foi conduzida calibração controlada em T2, testando `n_components` em {2, 4, 6, 0.80, 0.90, 0.95, 0.99} e `svd_solver` em {`auto`, `full`, `randomized`}, quando compatível. O melhor resultado foi obtido com `n_components=0.80`, `svd_solver=auto` e `whiten=False`. Com essa configuração, os melhores pontos por F1 foram `q=0.919` em T2 e `q=0.979` em T3.

Tabela 4.16: Comparação entre EFC, *Isolation Forest* e PCA no ponto de melhor F1 por janela (regime retrospectivo em T3).

Modelo	Split	Precisão	Recall	F1-score	FPR
EFC	T2	0,781	0,937	0,852	4,00%
Isolation Forest	T2	0,680	0,937	0,788	6,70%
PCA	T2	0,644	0,961	0,771	8,10%
EFC	T3	0,615	0,954	0,748	8,40%
Isolation Forest	T3	0,691	0,954	0,801	6,00%
PCA	T3	0,854	0,872	0,862	2,10%

Em T2, o PCA obteve maior *recall*, mas apresentou o maior FPR e F1-score inferior ao EFC. No regime retrospectivo de T3, o PCA superou os demais modelos nas métricas globais. Os intervalos de confiança por *bootstrap* indicam estabilidade amostral desse ponto.

A análise por tipo de ataque qualifica esse resultado. O PCA detectou as duas classes em T2, mas deixou de detectar o `aggressive_scan` no ponto retrospectivo de T3. A vantagem agregada precisa, portanto, ser interpretada em conjunto com a cobertura das classes.

4.9.3 Tentativa controlada com One-Class SVM

Também foi realizada uma tentativa controlada de incluir One-Class SVM como baseline *one-class* adicional. Para reduzir o custo computacional, o treinamento foi realizado sobre uma subamostra reprodutível de 25.000 fluxos benignos de T1, com `StandardScaler` embutido no pipeline do modelo, `kernel=rbf`, `gamma` em {`scale`, `auto`} e busca controlada sobre `nu` em {0,001, 0,005, 0,01, 0,02, 0,05}.

O piloto inicial em T2 indicou desempenho muito baixo e motivou uma busca controlada com dez combinações de `nu` e `gamma`. O melhor resultado utilizou `nu=0.05`, `gamma=scale` e `q=0.927`, com precisão de 0,146, *recall* de 0,082, F1-score de 0,105 e FPR de 7,30%.

Mesmo após o ajuste, o One-Class SVM permaneceu abaixo dos demais modelos e apresentou maior custo operacional. Como o treinamento utilizou uma subamostra de 25.000 fluxos benignos de T1, o resultado não demonstra limitação intrínseca do método. Ele registra apenas desempenho insuficiente no espaço de atributos, no orçamento computacional e no protocolo adotados. A tentativa foi encerrada em T2.

4.10 SÍNTESE COMPARATIVA DOS MODELOS

A comparação confirma que o problema não é trivial para detectores *one-class*. Em T2, o EFC apresentou o melhor equilíbrio agregado, enquanto o PCA obteve maior *recall* com mais falsos positivos. O One-Class SVM permaneceu pouco competitivo sob a subamostragem e o orçamento adotados.

Em T3, a ordenação depende do regime de calibração. *Isolation Forest* e PCA superam o EFC no regime retrospectivo apresentado na Tabela 4.16. Quando o ponto calibrado em T2 é aplicado prospectivamente a T3, o EFC apresenta o maior F1-score agregado, o *Isolation Forest* perde sensibilidade e o PCA amplia o FPR. A Tabela 4.17 reúne esses resultados.

Tabela 4.17: Comparação dos modelos em avaliação prospectiva estrita em T3 (*cutoff* de T2 aplicado a T3).

Modelo	Precisão	Recall	F1-score	FPR	Recall `syn_flood`	Recall `agr_scan`
EFC	0,616	0,954	**0,749**	8,34%	1,000	0,000
Isolation Forest	0,313	0,176	0,225	5,43%	0,185	0,000
PCA	0,544	0,927	0,686	10,90%	0,924	**0,998**

A leitura por tipo de ataque impede uma conclusão de superioridade simples. No regime prospectivo, o PCA é o único detector que recupera o `aggressive_scan`, enquanto o EFC permanece sem detectar essa classe. A diferença entre os regimes confirma que a interpretação depende simultaneamente do protocolo de calibração e da cobertura dos ataques, conforme discutido na Seção 5.3.

Em conjunto, os comparativos mostram que métricas globais, cobertura por ataque e regime de calibração podem produzir ordenações distintas. O protocolo temporal e a análise estratificada são necessários para distinguir vantagem agregada, estabilidade prospectiva e capacidade de detectar classes minoritárias.

Tabela 4.18: Recall por tipo de ataque e F1-score global por modelo (regime retrospectivo em T3).

Split	Modelo	Recall `syn_flood`	Recall `agr_scan`	F1 global
T2	EFC	1,000	0,000	0,852
	Isolation Forest	1,000	0,000	0,788
	PCA	0,958	**1,000**	0,771
T3	EFC	1,000	0,000	0,748
	Isolation Forest	1,000	0,000	0,801
	PCA	0,914	0,000	**0,862**

Optou-se por não aplicar testes de significância pareados, como o de McNemar, uma vez que a comparação central deste trabalho é qualitativa e estratificada por tipo de ataque, e não uma afirmação de superioridade global de um detector sobre outro. A variabilidade amostral das métricas agregadas já é quantificada pelos intervalos de confiança por *bootstrap* reportados para cada modelo.

4.11 ARTEFATOS CONSOLIDADOS

Os principais artefatos produzidos pelo experimento foram organizados em arquivos de dados, mapas de transformação, predições, relatórios estatísticos, gráficos e manifesto de reprodutibilidade. Entre os artefatos centrais estão os splits discretizados `T1.csv`, `T2.csv` e `T3.csv`; os arquivos de predição `predictions_T2.csv` e `predictions_T3.csv`; os mapas `discretization.json` e `categoricals.json`; os relatórios de *tuning*; os histogramas de energia; os relatórios por tipo de ataque; os arquivos de análise de deriva; os perfis de falsos positivos; os relatórios de *bootstrap*; e os artefatos dos baselines *Isolation Forest*, PCA e One-Class SVM.

Além disso, o pacote congelado em `runs/efc_repro_20260427_053541` contém o manifesto completo da execução, o resumo de métricas finais, os limiares utilizados e os hashes SHA-256 dos insumos, códigos, modelos e relatórios. Esse congelamento permite rastrear a origem dos resultados apresentados e contribui para a auditabilidade do experimento.

O pacote congelado também permitiu reconciliar os resultados da dissertação com os valores publicados em (22). A principal diferença ocorreu em T2 e decorreu da reexecução consistente do mesmo ciclo de *tuning*, predição e avaliação. T3 apresentou apenas variações marginais. Por essa razão, todas as tabelas e análises desta dissertação adotam o pacote congelado como referência, enquanto os valores da publicação permanecem registrados como resultado da execução anterior.

A comparação prospectiva estrita em T3 (Tabela 4.17) é um resultado derivado dos artefatos congelados, e não uma nova execução. Os valores foram recomputados a partir dos *scores* congelados das predições de T3, aplicando-se a cada modelo o *cutoff* calibrado em T2 por melhor F1, registrado nos relatórios de *tuning*. Esse procedimento não usa os rótulos de T3 para seleção do limiar. Esses rótulos são utilizados apenas para calcular ex post as métricas reportadas. A derivação foi registrada como artefato complementar de reprodutibilidade, por meio do script de comparação prospectiva e do relatório derivado correspondente, ambos armazenados fora do pacote congelado para preservar os *hashes* originais.

As matrizes de confusão, curvas ROC, curvas Precision-Recall, gráficos completos de deriva e gráficos completos de perfil dos falsos positivos foram mantidos como artefatos complementares. Essa decisão evita redundância no corpo principal, uma vez que as informações essenciais desses gráficos já são sintetizadas nas tabelas de métricas, de deriva, de perfil dos falsos positivos e de comparação entre modelos.

Os artefatos foram organizados de modo a preservar a rastreabilidade do experimento, incluindo scripts, configurações, manifestos, hashes, relatórios agregados e resultados derivados, ressalvadas as questões de privacidade, segurança institucional e exposição de metadados sensíveis da rede monitorada.

4.12 SÍNTESE DOS RESULTADOS

A avaliação prospectiva mostrou que o EFC preservou alto *recall* global em T2 e T3. A redução da precisão e o aumento do FPR bruto em T3 exigem interpretação contextual. A auditoria de proveniência mostrou que a maior parte desse aumento decorre do tráfego de retorno induzido pelo `syn_flood` e mantido sob rótulo benigno. O diagnóstico de deriva deve, portanto, apoiar-se nas distribuições dos fluxos benignos e no FPR depurado sob o limiar de referência de T1.

O desempenho agregado foi influenciado pela detecção do `syn_flood`, enquanto o `aggressive_scan` permaneceu pouco detectado pelo EFC. Nos modelos alternativos, a ordenação variou com o regime de calibração. A avaliação deve, portanto, combinar protocolo temporal, métricas por ataque, auditoria dos falsos positivos e diagnóstico distribucional.

5 DISCUSSÃO

Este capítulo discute os resultados do Capítulo 4 à luz do problema de pesquisa, das escolhas metodológicas e das limitações inerentes à construção de *datasets* realistas para NIDS baseados em fluxo. A análise considera a validação temporal, o ponto operacional, a estabilidade distribucional, o desempenho por tipo de ataque, a proveniência dos falsos positivos e a comparação com modelos *one-class* alternativos.

A metodologia produziu um *dataset* com tráfego benigno real, ataques controlados, separação temporal e artefatos auditáveis de rotulagem, treino, predição e avaliação. O EFC manteve alto *recall* global, mas apresentou comportamentos distintos entre ataques e entre regimes de calibração. A auditoria mostrou ainda que o FPR bruto de T3 foi dominado pelo retorno do `syn_flood`, enquanto o diagnóstico distribucional identificou deslocamento moderado no tráfego benigno.

A contribuição discutida não consiste em demonstrar superioridade universal do EFC. O resultado central está na capacidade do protocolo de revelar instabilidades que uma avaliação estática poderia ocultar. A separação cronológica, a distinção entre calibração prospectiva e análise retrospectiva, o diagnóstico distribucional e a avaliação estratificada permitem identificar quando e por que um detector se mostra estável ou frágil.

5.1 PROTOCOLO TEMPORAL COMO EIXO DE VALIDADE EXPERIMENTAL

O protocolo temporal constitui o principal diferencial da avaliação porque reproduz a ordem de disponibilidade das informações em operação. T1 define a normalidade, T2 fornece os rótulos usados para selecionar o ponto operacional e T3 verifica a transferência dessa decisão para dados posteriores. Essa separação impede que os rótulos da janela final influenciem a configuração do detector.

A preservação da ordem cronológica oferece uma condição mais exigente do que uma partição aleatória. A mistura de fluxos provenientes de períodos distintos poderia distribuir padrões semelhantes entre treinamento, calibração e teste. Esse procedimento reduziria artificialmente as diferenças entre janelas e produziria uma estimativa excessivamente favorável da estabilidade do modelo.

A interpretação das janelas deve considerar o período do dia. T2 cobre a madrugada e o início da manhã, enquanto T1 e T3 abrangem predominantemente horários comerciais. A comparação entre T2 e T3 combina evolução temporal e variação diurna. A comparação entre T1 e T3 oferece evidência mais controlada de deslocamento entre períodos operacionais equivalentes.

O aumento do FPR bruto em T3 não constitui evidência direta de deriva benigna. A auditoria mostrou que esse resultado é dominado pelo tráfego de retorno do `syn_flood`. A evidência de deslocamento temporal deve ser sustentada pelo diagnóstico distribucional calculado sobre fluxos benignos e pelo FPR depurado sob o limiar de referência de T1.

Os valores extremos de PSI em `packets_bin` nas comparações que envolvem T2 exigem cautela. T2 popula apenas 15 dos 29 valores observados em T1 e T3. Os 14 intervalos ausentes em um dos lados recebem proporções próximas de zero e ampliam o somatório do PSI. A comparação entre T1 e T3 não apresenta intervalos ausentes e resulta em PSI de 0,881, compatível com deslocamento moderado. A Distância de Wasserstein complementa essa leitura porque mede o deslocamento de massa sem a mesma amplificação causada por intervalos esparsos.

O valor do protocolo independe do detector empregado. A inversão da ordenação dos modelos entre os regimes prospectivo e retrospectivo demonstra que a escolha do procedimento de calibração pode alterar a conclusão experimental. A leitura conjunta das métricas, do diagnóstico distribucional e da proveniência permite discutir deslocamento dos dados sem extrapolar para uma afirmação isolada de *concept drift*.

5.2 SENSIBILIDADE DO *CUTOFF* E IMPLICAÇÕES OPERACIONAIS

A análise de sensibilidade mostrou que o ponto operacional determina o compromisso entre *recall* e FPR. Quantis mais altos reduzem alarmes falsos, mas podem eliminar grande parte da detecção. A configuração operacional deve considerar o custo relativo de falsos positivos e falsos negativos no ambiente monitorado.

A calibração prospectiva e a análise retrospectiva possuem finalidades distintas. O ponto selecionado em T2 pode ser aplicado a T3 porque utiliza apenas informações anteriores ao teste. Um ponto escolhido com os próprios rótulos de T3 representa o melhor cenário retrospectivo naquela janela e não uma condição disponível durante a implantação.

Detectores baseados em energia precisam de monitoramento contínuo do ponto operacional. A análise de tendência das energias benignas, a definição de faixas de alerta e a validação de casos limítrofes podem apoiar uma recalibração futura. Qualquer atualização deve utilizar dados anteriores à janela que será avaliada para preservar a validade prospectiva.

5.3 DESEMPENHO POR TIPO DE ATAQUE

A análise por tipo de ataque evitou uma interpretação excessivamente favorável das métricas globais. O alto *recall* do EFC foi sustentado pela detecção integral do `syn_flood`, que domina a classe maliciosa. O `aggressive_scan` permaneceu sem detecção relevante e revelou uma fragilidade que o resultado agregado ocultaria.

A análise energética indica que o `syn_flood` ocupa uma região acima do *cutoff*, enquanto a varredura permanece dentro da distribuição benigna. A discretização faz esses fluxos convergirem para poucas configurações com energias próximas às do tráfego legítimo. Como não houve ablação sistemática do número de *bins*, essa explicação deve ser tratada como hipótese sustentada pelos resultados e não como causa isolada definitivamente demonstrada.

O contraste com o PCA ajuda a delimitar essa limitação. Com o ponto calibrado em T2 e aplicado prospectivamente a T3, o PCA detecta o `aggressive_scan`, enquanto o EFC permanece sem detectar a classe. A diferença indica uma limitação da representação discretizada do EFC sob o protocolo avaliado e não uma insuficiência necessária dos atributos de fluxo disponíveis. O resultado retrospectivo do PCA em T3 utiliza um limiar mais conservador e oculta essa capacidade.

Métricas por tipo de ataque devem acompanhar os resultados agregados. O aprimoramento do EFC pode considerar atributos de frequência, contadores por origem e destino, diversidade de portas e agregações temporais. Essas extensões permitiriam representar atividades cuja evidência emerge da correlação entre fluxos e não do volume de cada registro isolado.

5.4 PERFIL DOS FALSOS POSITIVOS E INTERPRETAÇÃO OPERACIONAL

A auditoria apresentada na Seção 4.7.1 mostra que os falsos positivos se concentram em um padrão específico. A maior parte envolve a resposta do alvo do `syn_flood` ao host atacante durante as janelas das campanhas. Esse padrão não caracteriza uma degradação uniforme do detector sobre as comunicações legítimas da rede.

Os fluxos de retorno foram contabilizados como falsos positivos porque a regra de rotulagem preservada no pacote congelado identifica somente a direção de ida do ataque. Contudo, eles são induzidos pela campanha e não constituem tráfego benigno independente. Essa distinção permite interpretar o FPR agregado com cautela e mostra a importância de preservar evidências de proveniência na construção do *dataset*.

Do ponto de vista operacional, esse resultado é importante porque falsos positivos concentrados são mais tratáveis do que falsos positivos difusos. Quando os alarmes se agrupam por sub-rede, serviço, horário ou direção, torna-se possível investigar causas locais, criar regras de supressão contextual, ajustar políticas de monitoramento ou revisar a discretização de determinados atributos. Essa possibilidade exige cautela, pois exceções excessivamente amplas podem mascarar ataques reais.

A análise de falsos positivos deve, portanto, ser vista como parte integrante da avaliação de NIDS, e não apenas como diagnóstico secundário. Ela permite transformar uma métrica agregada, como FPR, em informação operacional interpretável, indicando onde e quando o detector tende a errar.

5.5 COMPARAÇÃO COM MODELOS *ONE-CLASS* ALTERNATIVOS

Os baselines mostram que o problema não é trivial para detectores *one-class*. O EFC apresentou o melhor equilíbrio agregado em T2. No regime retrospectivo de T3, *Isolation Forest* e PCA obtiveram resultados globais superiores. No regime prospectivo, a ordenação se inverteu e o EFC apresentou o maior F1-score agregado.

A cobertura por ataque modifica essa leitura. O PCA foi o único modelo que detectou o `aggressive_scan` em T3 sob o protocolo prospectivo, enquanto o EFC e o *Isolation Forest* permaneceram sem detectar a classe. O One-Class SVM apresentou baixa competitividade sob a subamostragem e o orçamento avaliados, sem permitir uma conclusão sobre limitação intrínseca do método.

Não há um modelo superior em todas as dimensões observadas. A escolha deve considerar estabilidade temporal, custo dos falsos positivos, cobertura das classes e interpretabilidade dos erros. A dependência dos resultados em relação ao regime de calibração reforça o papel do protocolo temporal na comparação controlada entre modelos.

5.6 DESIGN DA COLETA E IMPLICAÇÕES PARA *DATASETS* REALISTAS

As decisões anteriores à coleta determinam o alcance da avaliação. Os ataques foram programados fora de T1, o que preservou a pureza do treinamento *one-class*. O plano versionado permite verificar essa condição e transforma a ausência de contaminação em uma propriedade auditável do experimento.

Os planos de coleta e de ataques foram mantidos em artefatos distintos. Essa separação preserva a rastreabilidade entre o ambiente observado e as intervenções realizadas. Ela também permite reprocessar os dados brutos e revisar critérios de rotulagem sem repetir a aquisição do tráfego.

O tráfego benigno foi obtido em uma rede operacional e preserva variações que não foram determinadas pelo pesquisador. Os ataques foram executados com ferramentas, alvos e janelas conhecidos. A combinação aproxima o *dataset* do comportamento legítimo de uma rede real e mantém controle sobre a atividade maliciosa (6, 19). A cobertura dos ataques permanece limitada aos cenários planejados.

A janela de aproximadamente 83 horas foi suficiente para demonstrar o protocolo e observar instabilidade de curto prazo. Ela não caracteriza ciclos semanais, sazonalidade ou mudanças prolongadas da infraestrutura. A coleta longitudinal em outros ambientes constitui uma extensão necessária para avaliar a generalização temporal em horizontes maiores (21).

Um *dataset* realista precisa combinar separação temporal, rotulagem rastreável e preservação dos artefatos experimentais. O congelamento de configurações, modelos, predições, relatórios e *hashes* reduz o risco de divergência entre calibração e avaliação. A contribuição metodológica reside nesse processo auditável e não apenas no desempenho obtido por um detector específico.

5.7 LIMITAÇÕES E CONSIDERAÇÕES OPERACIONAIS

A interpretação deve considerar limitações de rotulagem, escopo temporal, representação dos dados e robustez estatística. Os rótulos são rastreáveis ao plano de ataques, mas a evidência do `syn_flood` não distingue as campanhas A02 e A04 com o mesmo grau de confiança. Essa limitação não compromete a

avaliação binária nem a análise por tipo de ataque. Ela impede conclusões mais detalhadas por campanha. Também permanece a possibilidade residual de anomalias não planejadas no tráfego tratado como benigno.

Os achados estão limitados a uma rede e a uma coleta de aproximadamente 83 horas. O período permite avaliar janelas cronologicamente distintas, mas não abrange ciclos semanais, sazonalidade, manutenções ou mudanças prolongadas da infraestrutura. Avaliações em períodos mais extensos e em redes com outros perfis são necessárias para examinar a generalização dos resultados.

A representação utiliza atributos discretizados de volume, protocolo, serviço, direção e sub-redes. Essa escolha é compatível com tráfego cifrado e com restrições de privacidade, mas limita a detecção de atividades cuja evidência depende da correlação entre fluxos. O desempenho do `aggressive_scan` mostra a necessidade de investigar atributos derivados e agregações temporais. Como somente dois perfis de ataque foram avaliados, a cobertura comportamental não representa o conjunto de ameaças relevantes para uma implantação.

Em uma implantação, o *cutoff* deve ser definido com dados anteriores à janela avaliada e monitorado continuamente. Regras de supressão contextual podem reduzir alarmes concentrados, mas exigem validação para não ocultar ataques reais. O protocolo demonstra como medir esses efeitos, mas não define uma política operacional universal de recalibração ou supressão.

Os intervalos por *bootstrap* pressupõem independência entre os fluxos. Essa suposição é limitada porque o tráfego apresenta correlação por sessão, aplicação, origem, destino e período de coleta. Os intervalos podem subestimar a variabilidade amostral. Uma avaliação futura pode empregar *block bootstrap* com blocos temporais ou arquivos de coleta.

Essas limitações delimitam o alcance dos resultados e orientam avaliações futuras com coletas mais longas, maior diversidade de ataques, atributos temporais e procedimentos estatísticos que preservem a dependência entre fluxos.

6 CONSIDERAÇÕES FINAIS

Esta dissertação teve como objetivo geral propor e avaliar empiricamente, por meio de um estudo de caso controlado, uma metodologia para construção de *datasets* realistas baseados em fluxos de rede, com foco na avaliação de detectores de anomalia sob deriva temporal. Esse objetivo foi atendido pela implementação de um *pipeline* experimental capaz de coletar tráfego benigno em ambiente real, inserir ataques controlados, extrair e preparar atributos de fluxo, realizar rotulagem assistida, organizar recortes temporais prospectivos e avaliar o detector em janelas cronologicamente distintas. As conclusões devem ser interpretadas à luz do ambiente avaliado, da duração da coleta e dos tipos de ataque contemplados, sem prejuízo da contribuição metodológica do estudo.

Os resultados indicam que a avaliação de NIDS baseados em aprendizado de máquina não deve se limitar a partições aleatórias nem a métricas agregadas. A validação temporal evidenciou aspectos operacionais relevantes, como sensibilidade ao *cutoff*, deriva T1–T3 entre períodos comerciais equivalentes evidenciada pelo diagnóstico distribucional e pelo FPR depurado, concentração dos erros em perfis específicos de tráfego e diferença de desempenho entre tipos de ataque. A auditoria de proveniência dos falsos positivos mostrou ainda que o aumento agregado do FPR em T3 decorreu majoritariamente de tráfego de retorno do ataque na fronteira de rotulagem, um achado que só foi possível pela rastreabilidade do processo. Esses resultados reforçam a importância de metodologias que incorporem temporalidade, rastreabilidade, reprodutibilidade e análise de estabilidade distribucional.

Os objetivos específicos da pesquisa foram contemplados da seguinte forma:

1. **Desenvolver um *pipeline* que integre captura contínua de tráfego benigno real com injeção controlada de ataques.** Esse objetivo foi contemplado pela definição de um processo experimental que combina coleta de fluxos NetFlow/IPFIX em ambiente real, execução planejada de ataques controlados, extração de atributos, preparação dos dados e organização dos artefatos. A arquitetura conteinerizada, composta pelos serviços `netflow-collector`, `attacker` e `efc-runner`, contribuiu para separar responsabilidades, favorecer a portabilidade e preservar a rastreabilidade das execuções.

2. **Implementar um mecanismo de rotulagem assistida e rastreável.** A rotulagem associou fluxos coletados aos planos de ataque, considerando janelas temporais, endereços, portas, protocolos e artefatos de coleta. Esse processo permitiu consolidar o *ground truth* e distinguir fluxos benignos e maliciosos de forma auditável. A aplicação da metodologia também evidenciou a importância de complementar a rotulagem com alertas de IDS, logs de sistemas, registros de firewall e monitores de fluxo em campanhas futuras.

3. **Estabelecer um protocolo de validação baseado em divisões temporais prospectivas.** Esse objetivo foi atendido pela separação dos dados em T1, T2 e T3, preservando a ordem cronológica da coleta. T1 foi utilizado para treinamento do EFC apenas com tráfego benigno; T2 funcionou como

janela intermediária de calibração e avaliação; e T3 foi reservado para avaliação temporal posterior. Essa estrutura permitiu quantificar o impacto do deslocamento temporal sobre precisão, *recall*, F1-score e FPR, além de viabilizar a análise distribucional por PSI e Distância de Wasserstein.

4. **Aplicar a metodologia proposta em um estudo de caso com o EFC.** A metodologia foi aplicada ao *Energy-based Flow Classifier* em modo *one-class*, treinado exclusivamente com fluxos benignos de T1. O EFC apresentou alto *recall* global em T2 e T3, sobretudo pela detecção consistente dos fluxos `syn_flood`. Por outro lado, a redução da precisão agregada em T3 e a baixa detecção da classe `aggressive_scan` evidenciaram diferenças relevantes entre tipos de ataque. A auditoria de proveniência mostrou ainda que o aumento agregado do FPR em T3 decorreu majoritariamente de tráfego de retorno do ataque rotulado como benigno, o que a própria rastreabilidade do processo permitiu identificar. Assim, o estudo de caso mostrou que a metodologia permite mensurar desempenho e revelar condições em que o modelo se mostra mais ou menos sensível.

5. **Analisar, em caráter exploratório, o impacto do ajuste do ponto operacional do detector em cenários prospectivos sujeitos a deslocamento temporal.** Esse objetivo foi explorado por meio da análise de sensibilidade do *cutoff*, da caracterização dos falsos positivos e da comparação com modelos *one-class* alternativos. A análise evidenciou o compromisso entre redução do FPR e preservação do *recall*; os falsos positivos concentraram-se em tráfego interno, sub-redes específicas e janelas próximas aos ataques; e os *baselines* com *Isolation Forest*, PCA e *One-Class SVM* permitiram comparar diferentes comportamentos sob o mesmo protocolo. No caso do One-Class SVM, a tentativa foi limitada por subamostragem e custo computacional, de modo que seus resultados contextualizam o protocolo adotado sem estabelecer uma limitação intrínseca do método. Os resultados sugerem que a mitigação da perda de desempenho depende da calibração do ponto operacional, do enriquecimento dos atributos, do monitoramento contínuo da deriva e da avaliação estratificada por tipo de ataque.

Além dos objetivos definidos, a pesquisa evidenciou duas contribuições complementares. A primeira consiste em demonstrar que métricas globais podem ocultar diferenças relevantes de desempenho em *datasets* desbalanceados, uma vez que o alto *recall* do EFC foi influenciado pela detecção dos fluxos `syn_flood`, enquanto a classe `aggressive_scan` apresentou baixa detecção. A segunda está na importância dos *baselines* para contextualizar o modelo principal. No ponto de melhor F1 de cada janela, que em T3 corresponde ao regime retrospectivo, o PCA obteve o melhor desempenho agregado, embora não tenha detectado a classe de varredura nesse ponto. Essa ausência reflete o limiar conservador imposto pela otimização de F1 sobre a classe majoritária, e não incapacidade do modelo. Sob avaliação prospectiva estrita, com o limiar calibrado em T2 aplicado a T3, a comparação passa a refletir a capacidade de transferência do ponto operacional entre janelas. Nesse regime, o PCA recupera a varredura e o EFC apresenta o maior F1-score agregado. Esse condicionamento reforça que conclusões baseadas em métrica global dependem do protocolo de calibração.

Assim, a principal contribuição desta dissertação está na proposição e demonstração empírica, em estudo de caso controlado, de um processo metodológico para construir, documentar e avaliar *datasets* realistas de NIDS baseados em fluxo. A metodologia permitiu observar fenômenos que poderiam ser mascarados por avaliações tradicionais, como deslocamento temporal, instabilidade do limiar de decisão,

concentração operacional de falsos positivos, variação de desempenho entre ataques e impacto de modelos alternativos sob um mesmo protocolo experimental. Esses elementos fornecem base mais sólida para interpretar detectores *one-class* em redes dinâmicas e orientar aprimoramentos futuros do *pipeline*, dos atributos e do modelo.

Os achados delimitam o alcance da metodologia ao contexto experimental avaliado, mas também indicam sua utilidade como base para estudos futuros. A ampliação da janela de coleta, a inclusão de novos ataques, o enriquecimento dos atributos de fluxo, a aplicação em diferentes ambientes de rede e a exploração de estratégias de recalibração contínua podem fortalecer a generalização da proposta. Dessa forma, a metodologia apresentada oferece uma base reprodutível e auditável, validada internamente no estudo de caso conduzido, para construção e avaliação temporal de *datasets* de NIDS baseados em fluxo, contribuindo para avaliações mais realistas de detectores de anomalia em redes dinâmicas.

6.1 TRABALHOS FUTUROS

Os trabalhos futuros decorrem dos aspectos metodológicos e operacionais observados nesta pesquisa. Uma primeira direção consiste em ampliar a avaliação da metodologia por meio de coletas em múltiplos ambientes de rede e por períodos mais extensos, permitindo verificar se os padrões de deslocamento temporal, concentração de falsos positivos e sensibilidade ao *cutoff* observados neste estudo se mantêm em redes com diferentes perfis operacionais.

Recomenda-se também aumentar a representatividade temporal e comportamental do *dataset*. Coletas mais longas, com ciclos semanais, mudanças de turno, variações sazonais, manutenções programadas e eventos operacionais atípicos, poderiam caracterizar melhor a estabilidade do tráfego benigno e a intensidade da deriva ao longo do tempo. Além disso, a inclusão de ataques mais diversos, como força bruta, exploração de aplicações web, movimentação lateral, exfiltração de dados, comunicação com servidores C2, botnets, ataques distribuídos e estratégias *low-and-slow*, permitiria avaliar a metodologia diante de padrões maliciosos menos dependentes de alterações volumétricas evidentes.

Outra linha de continuidade envolve o enriquecimento dos atributos utilizados pelos detectores. A baixa detecção da classe `aggressive_scan` sugere que atributos calculados apenas no nível do fluxo individual podem ser insuficientes para capturar ataques de reconhecimento, varreduras graduais e comportamentos distribuídos temporalmente. Assim, trabalhos futuros podem incorporar atributos agregados por janelas temporais, como taxa de novas conexões por origem, diversidade de portas, dispersão de destinos, entropia de serviços, contadores por sub-rede e agregações por pares origem-destino. Avaliar o efeito da granularidade da discretização sobre a separação entre varredura e tráfego benigno, testando sistematicamente diferentes números de *bins*, permitiria determinar se o colapso dos fluxos de `aggressive_scan` para configurações discretas compartilhadas com o tráfego benigno pode ser atenuado sem comprometer a estabilidade do modelo. A concentração dos fluxos de varredura em poucos valores de energia distintos sugere ainda que assinaturas de regularidade energética por janela temporal poderiam complementar a regra de limiar único sobre fluxos individuais, substituindo a decisão por fluxo pela detecção de padrões de concentração anômala ao longo do tempo. Também se mostra promissora a investigação de versões multiclasse

e *open-set* do EFC, capazes de distinguir tráfego benigno, ataques conhecidos e eventos desconhecidos em cenários nos quais a classe maliciosa permanece aberta ou incompletamente observada.

A rotulagem e a calibração operacional do detector também constituem frentes relevantes de aprimoramento. A integração da rotulagem assistida com alertas de IDS, logs de sistemas, registros de firewall, telemetria de aplicações e ferramentas de análise de fluxos pode reduzir ambiguidades e fortalecer a confiabilidade do *ground truth*. De forma complementar, os resultados indicam a necessidade de investigar estratégias adaptativas de calibração do *cutoff*, incluindo limiares dinâmicos, quantis móveis, controle da taxa de alertas e atualização baseada no monitoramento contínuo da distribuição de energia. Essas abordagens podem contribuir para mitigar perdas de desempenho em ambientes sujeitos a deslocamento temporal.

A calibração adaptativa do limiar de decisão constitui uma direção natural para investigação por meio de aprendizado por reforço. Nessa formulação, o agente utilizaria como estado os sinais de deriva distribucional já produzidos pelo *pipeline*, como PSI, Distância de Wasserstein e tendência da distribuição de energia benigna, aprendendo uma política de ajuste do *cutoff* que preserve o equilíbrio entre *recall* e especificidade ao longo do tempo sem exigir rótulos de ataque para operação. Essa direção estende os achados sobre a instabilidade do limiar entre T2 e T3 para um regime de calibração contínua, conectando os mecanismos de diagnóstico distribucional já empregados nesta pesquisa a um arcabouço formal de decisão sequencial sob incerteza.

Por fim, recomenda-se aproximar o *pipeline* de um cenário operacional contínuo e ampliar sua avaliação em ambientes heterogêneos, como redes corporativas, acadêmicas, industriais, IoT/IIoT e governamentais. Essa evolução permitiria comparar perfis de tráfego benigno, intensidades de deriva e padrões de ataque, além de avaliar a capacidade de generalização da metodologia. Recomenda-se ainda aprofundar a análise de incerteza estatística com *block bootstrap* aplicado a janelas temporais, arquivos de coleta ou sessões agregadas, preservando a dependência temporal entre os fluxos. A disponibilização controlada de scripts, configurações, mapas de atributos, procedimentos de anonimização, relatórios e manifestos de execução também pode contribuir para auditoria, replicação e adaptação por outros pesquisadores, observadas as restrições de privacidade, segurança institucional e exposição de metadados sensíveis da rede monitorada.

REFERÊNCIAS BIBLIOGRÁFICAS

1 CERT.br. *Incidentes Notificados ao CERT.br*. 2025. Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil. Disponível em: <https://stats.cert.br/incidentes/>.

2 PONTES, C.; SOUZA, M.; GONDIM, J.; MAROTTA, M. A.; BISHOP, M. A new method for flow-based network intrusion detection using the inverse potts model. *IEEE Transactions on Network and Service Management*, v. 18, n. 2, p. 1125–1136, 2021.

3 GOLDSCHMIDT, P.; CHUDá, D. Network intrusion datasets: A survey, limitations, and recommendations. *Computers & Security*, v. 156, p. 104510, 2025.

4 PINTO, D.; AMORIM, I.; MAIA, E.; PRAçA, I. A review on intrusion detection datasets: Tools, processes, and features. *Computer Networks*, v. 262, p. 111177, 2025.

5 ENISA. *ENISA Threat Landscape 2025 Booklet*. [S.l.], 2025. Accessed: 2026-04-18. Disponível em: <https://www.enisa.europa.eu/sites/default/files/2025-10/ENISA%20Threat%20Landscape%202025%20Booklet.pdf>.

6 FERRIYAN, A.; THAMRIN, A. H.; TAKEDA, K.; MURAI, J. Generating network intrusion detection dataset based on real and encrypted synthetic attack traffic. *Applied Sciences*, v. 11, p. 7868, 2021.

7 MONDRAGON, J. C.; BRANCO, P.; JOURDAN, G.-V.; GUTIERREZ-RODRIGUEZ, A. E.; BISWAL, R. R. Advanced ids: A comparative study of datasets and machine learning algorithms for network flow-based intrusion detection systems. *Applied Intelligence*, v. 55, p. 608, 2025.

8 RUFFO, V. G. da S.; LENT, D. M. B.; KOMARCHESQUI, M.; SCHIAVON, V. F.; ASSIS, M. V. O. de; CARVALHO, L. F.; PROENçA, M. L. Anomaly and intrusion detection using deep learning for software-defined networks: A survey. *Expert Systems with Applications*, v. 256, p. 124982, 2024.

9 SARHAN, M.; LAYEGHY, S.; MOUSTAFA, N.; PORTMANN, M. Netflow datasets for machine learning-based network intrusion detection systems. In: *Big Data Technologies and Applications*. Cham: Springer, 2021. p. 117–135.

10 SARHAN, M.; LAYEGHY, S.; PORTMANN, M. Towards a standard feature set for network intrusion detection system datasets. *Mobile Networks and Applications*, v. 27, n. 1, p. 357–370, 2022.

11 LAYEGHY, S.; GALLAGHER, M.; PORTMANN, M. Benchmarking the benchmark — comparing synthetic and real-world network IDS datasets. *Journal of Information Security and Applications*, v. 80, p. 103689, 2024.

12 BERTOLI, G. d. C.; JUNIOR, L. A. P.; SAOTOME, O.; SANTOS, A. L. dos. Generalizing intrusion detection for heterogeneous networks: A stacked-unsupervised federated learning approach. *arXiv preprint arXiv:2209.00721*, 2022. Disponível em: <https://arxiv.org/abs/2209.00721>.

13 MOUSTAFA, N.; SLAY, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: *Military Communications and Information Systems Conference (MilCIS)*. [S.l.: s.n.], 2015. p. 1–6.

14 SHARAFALDIN, I.; LASHKARI, A. H.; GHORBANI, A. A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: *Proceedings of the 4th International Conference on Information Systems Security and Privacy*. [S.l.: s.n.], 2018. p. 108–116.

15 MOUSTAFA, N. New generations of internet of things datasets for cybersecurity applications based machine learning: TON_IoT datasets. In: *Proceedings of the eResearch Australasia Conference*. [S.l.: s.n.], 2019. p. 21–25.

16 ALSAEDI, A.; MOUSTAFA, N.; TARI, Z.; MAHMOOD, A.; ANWAR, A. TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems. *IEEE Access*, v. 8, p. 165130–165150, 2020.

17 GHARIB, A.; SHARAFALDIN, I.; LASHKARI, A. H.; GHORBANI, A. A. An evaluation framework for intrusion detection dataset. In: *International Conference on Information Science and Security*. [S.l.: s.n.], 2017. p. 0–4.

18 GUERRA, J. L.; CATANIA, C.; VEAS, E. Datasets are not enough: Challenges in labeling network traffic. *Computers & Security*, v. 120, p. 102810, 2022.

19 CORDERO, C. G.; VASILOMANOLAKIS, E.; WAINAKH, A.; MüHLHäUSER, M.; NADJM-TEHRANI, S. On generating network traffic datasets with synthetic attacks for intrusion detection. *ACM Transactions on Privacy and Security*, v. 24, n. 2, 2021.

20 DANIEL, N.; KAISER, F. K.; GILADI, S.; SHARABI, S.; MOYAL, R.; SHPOLYANSKY, S.; MURILLO, A.; ELYASHAR, A.; PUZIS, R. Labeling network intrusion detection system (NIDS) rules with MITRE ATT&CK techniques: Machine learning vs. large language models. *Big Data and Cognitive Computing*, v. 9, n. 2, p. 23, 2025.

21 SCHRAVEN, J.; WINDMANN, A.; NIGGEMANN, O. Mawiflow benchmark: Realistic flow-based evaluation for network intrusion detection. In: INSTICC. *Proceedings of the 12th International Conference on Information Systems Security and Privacy (ICISSP)*. [S.l.]: SciTePress, 2026. v. 1, p. 549–556. ISBN 978-989-758-800-6. ISSN 2184-4356.

22 OLIVEIRA, E. R. d.; ALBUQUERQUE, R. d. O.; GONDIM, J. J. C. Methodology for building realistic network intrusion detection datasets: A case study with the energy-based flow classifier. *Proceedings of the 10th Workshop on Communication Networks and Power Systems (WCNPS)*, 2025. 6 p.

23 HINDY, H.; BROSSET, D.; BAYNE, E.; SEEAM, A. K.; TACHTATZIS, C.; ATKINSON, R.; BELLEKENS, X. A taxonomy of network threats and the effect of current datasets on intrusion detection systems. *IEEE Access*, v. 8, p. 104650–104675, 2020.

24 VIEGAS, E. K.; SANTIN, A. O.; OLIVEIRA, L. S. Toward a reliable anomaly-based intrusion detection in real-world environments. *Computer Networks*, v. 127, p. 200–216, 2017.

25 AYAD, A. G.; EL-GAYAR, M. M.; HIKAL, N. A.; SAKR, N. A. Efficient real-time anomaly detection in IoT networks using one-class autoencoder and deep neural network. *Electronics*, v. 14, n. 1, p. 104, 2025.

26 ELDHAI, A. M.; HAMDAN, M.; ABDELAZIZ, A.; HASHEM, I.; BABIKER, S. F.; MARSONO, M. N.; HAMZAH, M.; JHANJHI, N. Z. Improved feature selection and stream traffic classification based on machine learning in software-defined networks. *IEEE Access*, 2024. Early access / accepted version.

27 NUAIMI, T. A.; ZAABI, S. A.; ALYILIELI, M.; ALMASKARI, M.; ALBLOOSHI, S.; ALHABSI, F.; YUSOF, M. F. B.; BADAWI, A. A. A comparative evaluation of intrusion detection systems on the Edge-IIoT-2022 dataset. *Intelligent Systems with Applications*, v. 20, p. 200298, 2023.

28 SINGLA, A.; BERTINO, E.; VERMA, D. Preparing network intrusion detection deep learning models with minimal data using adversarial domain adaptation. *Proceedings of the ACM Asia Conference on Computer and Communications Security*, p. 127–140, 2020.

29 SHARMILA, B. S.; NAGAPADMA, R. Quantized autoencoder (qae) intrusion detection system for anomaly detection in resource-constrained iot devices using RT-IoT2022 dataset. *Cybersecurity*, v. 6, p. 41, 2023.

30 PONTES, C. F. T. *Um novo método baseado no modelo de Potts para detecção de intrusão em rede*. 2020. Trabalho de Conclusão de Curso, Universidade de Brasília.

31 LOPES, D. A. G.; MAROTTA, M. A.; LADEIRA, M.; GONDIM, J. J. C. Botnet detection based on network flow analysis using inverse statistics. In: *17th Iberian Conference on Information Systems and Technologies (CISTI)*. Madrid, Spain: [s.n.], 2022.

32 MUNHOZ, J. V. B. *Classificação de fluxos de rede IoT MQTT utilizando o EFC*. 2022. Trabalho de Conclusão de Curso, Universidade de Brasília.

33 SOUZA, M. M. C.; PONTES, C. T.; GONDIM, J. J. C.; GARCIA, L. P. F.; DASILVA, L.; CAVALCANTE, E. F. M.; MAROTTA, M. A. A novel open set energy-based flow classifier for network intrusion detection. *Computers & Security*, v. 157, p. 104569, 2025.

34 GAMA, J.; ŽLIOBAITė, I.; BIFET, A.; PECHENIZKIY, M.; BOUCHACHIA, A. A survey on concept drift adaptation. *ACM Computing Surveys*, v. 46, n. 4, p. 44:1–44:37, 2014.

35 LIPPMANN, R. P.; FRIED, D. J.; GRAF, I.; HAINES, J. W.; KENDALL, K. R.; MCCLUNG, D.; WEBER, D.; WEBSTER, S. E.; WYSCHOGROD, D.; CUNNINGHAM, R. K.; ZISSMAN, M. A. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In: *Proceedings DARPA Information Survivability Conference and Exposition (DISCEX)*. [S.l.: s.n.], 2000. v. 2, p. 12–26.

36 TAVALLAEE, M.; BAGHERI, E.; LU, W.; GHORBANI, A. A. A detailed analysis of the KDD CUP 99 data set. In: *IEEE Symposium on Computational Intelligence for Security and Defense Applications*. [S.l.: s.n.], 2009. p. 1–6.

37 MOUSTAFA, N.; SLAY, J. The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. *Information Security Journal: A Global Perspective*, v. 25, p. 18–31, 2016.

38 LIU, L.; ENGELEN, G.; LYNAR, T.; ESSAM, D.; JOOSEN, W. Error prevalence in NIDS datasets: A case study on CIC-IDS2017 and CSE-CIC-IDS-2018. In: *IEEE Conference on Communications and Network Security (CNS)*. [S.l.: s.n.], 2022. p. 254–262.

39 MOUSTAFA, N. A new distributed architecture for evaluating ai-based security systems at the edge: Network ton iot datasets. *Sustainable Cities and Society*, v. 72, p. 102994, 2021.

40 JIANG, W.; ZHANG, B.; ZHU, Q.; LIAO, C.; WANG, W. A real network environment dataset for traffic analysis. *Scientific Data*, v. 12, p. 756, 2025.

41 SHIRAVI, A.; SHIRAVI, H.; TAVALLAEE, M.; GHORBANI, A. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Comput. Secur.*, v. 31, p. 357–374, 2012.

42 BHUYAN, M. H.; BHATTACHARYYA, D. K.; KALITA, J. K. Towards generating real-life datasets for network intrusion detection. *International Journal of Network Security*, v. 17, p. 683–701, 2015.

43 HAIDER, W.; HU, J.; SLAY, J.; TURNBULL, B. P.; XIE, Y. Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling. *Journal of Network and Computer Applications*, v. 87, p. 185–192, 2017.

44 VELARDE-ALVARADO, P.; GONZALEZ, H.; MARTíNEZ-PELáEZ, R.; MENA, L. J.; OCHOA-BRUST, A.; MORENO-GARCíA, E.; FéLIX, V. G.; OSTOS, R. A novel framework for generating personalized network datasets for NIDS based on traffic aggregation. *Sensors*, v. 22, n. 5, p. 1847, 2022.

45 CORDERO, C. G.; VASILOMANOLAKIS, E.; MILANOV, N.; KOCH, C.; HAUSHEER, D.; MüHLHäUSER, M. ID2T: A DIY dataset creation toolkit for intrusion detection systems. In: *Proceedings of the Conference on Communications and Network Security (CNS'15)*. [S.l.]: IEEE, 2015. p. 739–740.

46 VASILOMANOLAKIS, E.; CORDERO, C. G.; MILANOV, N.; MüHLHäUSER, M. Towards the creation of synthetic, yet realistic, intrusion detection datasets. In: *Proceedings of the IEEE/IFIP Network Operations and Management Symposium (NOMS)*. [S.l.: s.n.], 2016. p. 1209–1214.

47 GARGIULO, F.; MAZZARIELLO, C.; SANSONE, C. Automatically building datasets of labeled IP traffic traces: A self-training approach. *Applied Soft Computing*, v. 12, p. 1640–1649, 2012.

48 BEAUGNON, A.; CHIFFLIER, P.; BACH, F. ilab: An interactive labelling strategy for intrusion detection. In: *Research in Attacks, Intrusions, and Defenses*. Cham: Springer, 2017. p. 120–140.

49 GUERRA, J. L.; VEAS, E.; CATANIA, C. A. A study on labeling network hostile behavior with intelligent interactive tools. In: *IEEE Symposium on Visualization for Cyber Security (VizSec)*. [S.l.: s.n.], 2019. p. 1–10.

50 TORRES, J. L.; CATANIA, C. A.; VEAS, E. Active learning approach to label network traffic datasets. *Journal of Information Security and Applications*, v. 49, p. 102388, 2019.

51 FERRAG, M. A.; FRIHA, O.; HAMOUDA, D.; MAGLARAS, L.; JANICKE, H. Edge-iiotset: A new comprehensive realistic cyber security dataset of iot and iiot applications for centralized and federated learning. *IEEE Access*, Institute of Electrical and Electronics Engineers Inc., v. 10, p. 40281–40306, 2022. ISSN 21693536.

52 S., B.; NAGAPADMA, R. *RT-IoT2022* . 2023. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5P338.

53 KUMAR, P.; LIU, J.; TAYEEN, A. S. M.; MISRA, S.; CAO, H.; HARIKUMAR, J.; PEREZ, O. Flnet2023: Realistic network intrusion detection dataset for federated learning. In: *MILCOM 2023 - 2023 IEEE Military Communications Conference (MILCOM)*. [S.l.]: IEEE, 2023. p. 345–350. ISBN 979-8-3503-2181-4.

54 KOUMAR, J.; HYNEK, K.; ČEJKA, T. Network traffic classification based on single flow time series analysis. *arXiv preprint arXiv:2307.13434*, 7 2023. Disponível em: <http://arxiv.org/abs/2307.13434>.

55 NUAIMI, T. A.; ZAABI, S. A.; ALYILIELI, M.; ALMASKARI, M.; ALBLOOSHI, S.; ALHABSI, F.; YUSOF, M. F. B.; BADAWI, A. A. A comparative evaluation of intrusion detection systems on the edge-iiot-2022 dataset. *Intelligent Systems with Applications*, Elsevier B.V., v. 20, 11 2023. ISSN 26673053.

56 BERGMEIR, C.; BENÍTEZ, J. M. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, v. 191, p. 192–213, 2012.

57 TASHMAN, L. J. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, v. 16, n. 4, p. 437–450, 2000.

58 ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. *Statistics Surveys*, v. 4, p. 40–79, 2010.

59 KHADEMI, A.; HOPKA, M.; UPADHYAY, D. Model monitoring and robustness of in-use machine learning models: Quantifying data distribution shifts using population stability index. *arXiv preprint arXiv:2302.00775*, 2023. Preprint, accessed 2026-04-23. Disponível em: <https://arxiv.org/abs/2302.00775>.

60 QUIÑONERO-CANDELA, J.; SUGIYAMA, M.; SCHWAIGHOFER, A.; LAWRENCE, N. D. (Ed.). *Dataset Shift in Machine Learning*. Cambridge, MA: The MIT Press, 2009. ISBN 9780262170055.

61 Scikit-Learn Developers. *TimeSeriesSplit — scikit-learn documentation*. 2026. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html>. Acesso em: 2026-04-23.

62 MITRE ATT&CK. *Network Service Discovery, Technique T1046*. 2026. Acesso em: 2026-08-15. Disponível em: <https://attack.mitre.org/techniques/T1046/>.

63 MITRE ATT&CK. *Network Denial of Service: Direct Network Flood, Sub-technique T1498.001*. 2026. Acesso em: 2026-08-15. Disponível em: <https://attack.mitre.org/techniques/T1498/001/>.