UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CAMILA DE SOUSA DANTAS

UMA ESTRATÉGIA PARA SELEÇÃO DE ATRIBUTOS EM DADOS NÃO PARAMÉTRICOS COM APLICAÇÕES EM APRENDIZADO DE MÁQUINA

Recife
2025

CAMILA DE SOUSA DANTAS

UMA ESTRATÉGIA PARA SELEÇÃO DE ATRIBUTOS EM DADOS NÃO PARAMÉTRICOS COM APLICAÇÕES EM APRENDIZADO DE MÁQUINA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Redes de Computadores e Sistemas Distríbuidos

Orientador: Prof. Dr. Jamilson Ramalho Dantas

Coorientador: Prof. Dr. João Ferreira da Silva Júnior

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Dantas, Camila de Sousa.
Uma estratégia para seleção de atributos em dados não paramétricos com aplicações em aprendizado de máquina / Camila de Sousa Dantas. - Recife, 2025.
75f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós Graduação em Ciências da Computação, 2025.
Orientação: Jamilson Ramalho Dantas.
Inclui Referências.

1. Seleção de atributos; 2. Técnicas não paramétricas; 3. Redução de dimensionalidade. I. Dantas, Jamilson Ramalho. II. Título.

UFPE-Biblioteca Central

Camila de Sousa Dantas

“Uma estratégia para Seleção de Atributos em Dados Não Paramétricos com Aplicações em Aprendizado de Máquina”

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Redes de Computadores e Sistemas Distribuídos.

Aprovado em: 28/08/2025.

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
EDUARDO ANTONIO GUIMARAES TAVARES
Data: 09/12/2025 08:51:05-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Eduardo Antonio Guimarães Tavares
Centro de Informática / UFPE

gov.br Documento assinado digitalmente
ERMESON CARNEIRO DE ANDRADE
Data: 09/12/2025 13:55:47-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Ermeson Carneiro de Andrade
Departamento de Computação / UFRPE

gov.br Documento assinado digitalmente
JAMILSON RAMALHO DANTAS
Data: 10/12/2025 07:45:20-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Jamilson Ramalho Dantas
Centro de Informática / UFPE
(**orientador**)

À minha criança, e àquela que de mim nasceu.

AGRADECIMENTOS

A elaboração desta dissertação foi marcada por inúmeros desafios, acadêmicos e sobretudo pessoais, que exigiram resiliência, paciência e dedicação constante. Em meio às dificuldades, encontrei apoio fundamental em pessoas que me acompanharam nesta trajetória e a quem devo profunda gratidão.

Primeiramente, agradeço a Deus e à espiritualidade, meus guias e toda a força invisível que me sustentou nos momentos de maior dificuldade e me deu coragem para seguir adiante nesta caminhada.

Ao meu orientador, Prof. Dr. Jamilson Dantas, pela orientação firme e pela confiança depositada em meu trabalho, sempre oferecendo direcionamentos precisos e estimulando o pensamento crítico, pela compreensão e apoio nas horas difíceis. Ao meu coorientador, Prof. Dr. João Ferreira, pela condução científica e apoio espiritual em toda a caminhada, pela disponibilidade e pelo olhar atento aos detalhes que tanto enriqueceram esta pesquisa.

Expresso também minha gratidão aos colegas de pesquisa, que foram mais que colaboradores: tornaram-se amigos, companheiros de jornada e fonte de inspiração em momentos de incerteza desde as disciplinas iniciais. As discussões, as trocas de experiências e, sobretudo, a amizade cultivada ao longo deste percurso tornaram o caminho mais leve e enriquecedor.

Ao grupo de pesquisa MoDCS, agradeço pelo espaço de aprendizado coletivo, pelas oportunidades de crescimento intelectual e pelas contribuições que fortaleceram este trabalho.

À minha família, razão de toda força encontrada para todos os momentos.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização desta dissertação. Sem o apoio, a paciência e a colaboração de cada um, esta etapa não teria sido possível.

"É preciso ter perseverança e, acima de tudo, confiança em si mesmo. Devemos acreditar que somos dotados de algo e que esse algo deve ser alcançado." - Marie Curie.

RESUMO

A análise de dados não paramétricos, desbalanceados e de alta dimensionalidade é um desafio recorrente em diversas aplicações de Aprendizado de Máquina (AM), onde métodos tradicionais de Seleção de Características (FS) frequentemente falham devido a suposições restritivas (como normalidade dos dados) ou alto custo computacional. Este trabalho propõe uma estratégia abrangente de FS para sistemas baseados em AM por meio de uma abordagem não paramétrica, robusta e escalável. O modelo é estruturado em três estágios: filtragem, clusterização e ranqueamento, utilizando métricas adaptadas como entropia de Shannon, correlação de Spearman, distância de Bhattacharyya modificada e Informação Mútua Ajustada (AMI), que dispensam premissas rígidas sobre a distribuição dos dados. Implementado em Python, o algoritmo foi validado experimentalmente em múltiplos cenários, incluindo estudos de caso em cibersegurança com bases de dados reais de tráfego de rede e ataques cibernéticos, empregando classificadores como Floresta Aleatória (RF), validação cruzada e testes estatísticos não paramétricos (Friedman e Nemenyi). Os resultados demonstraram redução de 81,5% no número total de características, considerando a média da redução nos três *datasets* utilizados, sem comprometer a exatidão, com superioridade estatística (p-valor $< 0{,}05$) em métricas como exatidão(ou acurácia), Pontuação F1 (média harmônica de precisão e revocação) (F1) e Área sob a Curva ROC (Característica de Operação do Receptor) (AUC-ROC) em comparação a métodos tradicionais, além de reduzir o tempo de processamento em até 3,8 vezes em comparação com a classificação sobre os conjuntos de dados completos. A estratégia proposta não apenas melhora a eficiência computacional e a performance preditiva em problemas complexos, mas também amplia a explicabilidade e adaptabilidade a domínios com dados heterogêneos, oferecendo uma alternativa para a seleção de atributos em cenários onde dados não paramétricos são predominantes.

Palavras-chave: Seleção de Atributos. Técnicas Não Paramétricas. Desempenho de Sistemas. Redução de Dimensionalidade.

ABSTRACT

The analysis of non-parametric, imbalanced, and high-dimensional data remains a recurring challenge in numerous Machine Learning (ML) applications, where traditional feature selection (FS) methods often fail due to restrictive assumptions (e.g., data normality) or high computational costs. This work proposes a comprehensive FS strategy for ML-based systems through a non-parametric, robust, and scalable approach. The model is structured in three stages: filtering, clustering, and ranking, employing adapted metrics such as Shannon entropy, Spearman correlation, modified Bhattacharyya distance, and adjusted mutual information (AMI), which eliminate rigid assumptions about data distribution. Implemented in Python, the algorithm was experimentally validated across multiple scenarios, including cybersecurity case studies with real-world network traffic and cyberattack datasets, using classifiers such as Random Forest, 10-fold cross-validation, and non-parametric statistical tests (Friedman and Nemenyi). Results showed an average dimensionality reduction of 81.5% without compromising accuracy, achieving statistical superiority (p-value < 0.05) in metrics such as accuracy, F1-score, and AUC-ROC compared to traditional methods, while reducing processing time by up to 3.8×. The stability of the selections exceeded 90% agreement, demonstrating the reliability of the model. The proposed strategy not only enhances computational efficiency and predictive performance in complex problems but also improves explainability and adaptability to domains with heterogeneous data, providing an effective alternative for feature selection in scenarios dominated by non-parametric data.

Keywords: Feature Selection. Non-parametric Techniques. System Performance. Dimensionality Reduction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM Aprendizado de Máquina.

AMI Informação Mútua Ajustada.

ANOVA Análise de Variância.

AUC-ROC Área sob a Curva ROC (Característica de Operação do Receptor).

BPSO Otimização por Enxame de Partículas Binária.

CD Diferença Crítica.

CICDDoS-2019 Conjunto de dados criado e usado em (Sharafaldin, Lashkari e Ghorbani, 2019).

CNN Rede Neural Convolucional.

CPU Unidade Central de Processamento.

DBSCAN Agrupamento Baseado em Densidade com Ruído.

DDoS Ataques de Negação de Serviço Distribuída.

DMS Medida de similaridade baseada em distância.

DT Árvore de Decisão.

F1 Pontuação F1 (média harmônica de precisão e revocação).

FS Seleção de Características.

HPC Contadores de Desempenho de Hardware.

HPC-Lab Conjunto de dados criado e usado em (Nascimento et al., 2021b).

IA Inteligência Artificial.

IDS Sistema de Detecção de Intrusões.

IG Ganho de Informação.

IoT Internet das Coisas.

IQR Intervalo interquartílico.

K-means Algoritmo de agrupamento k-médias.

KDE Estimativa de Densidade por Kernel.

KNN k-Vizinhos Mais Próximos.

LLM Modelo de linguagem de grande porte.

LSTM Memória de Longo e Curto Prazo.

MI Informação Mútua.

ML Aprendizado de Máquina.

MLP Perceptron Multicamadas.

mRMR Mínima Redundância e Máxima Relevância.

OPTICS Ordenação de Pontos para Identificar a Estrutura de Agrupamento.

PCA Análise de Componentes Principais.

RF Floresta Aleatória.

RFE Eliminação Recursiva de Atributos.

SDN Redes Definidas por Software.

SVM Máquina de Vetores de Suporte.

UNSW-NB15 Conjunto de dados criado e usado em (Moustafa e Slay, 2015).

XGBoost Impulsionamento de Gradiente Extremo.

LISTA DE SÍMBOLOS

Π	letra grega pi (maiúscula)
α	letra grega alfa
$\cup$	união de conjuntos
δ	letra grega delta
γ	letra grega gama
$\geq$	maior ou igual
$\in$	pertence
∞	infinito
$\leq$	menor ou igual
$\mathbb{N}$	conjunto dos números naturais
μ	letra grega mi
$\rightarrow$	seta para a direita
σ	letra grega sigma
$\surd$	raiz quadrada
$\subseteq$	subconjunto de
$\sum$	somatório
θ	letra grega teta
$\times$	multiplicação

SUMÁRIO

1 INTRODUÇÃO

A crescente sofisticação e frequência dos Ataques de Negação de Serviço Distribuída (DDoS) representam uma ameaça significativa à estabilidade e à disponibilidade de sistemas computacionais conectados em rede (Abd-Allah et al., 2025), (Berríos et al., 2025), (Kamalov et al., 2020), (Patel, 2025).

Em um cenário de tráfego intensivo, dinâmico e heterogêneo, técnicas baseadas em aprendizado de máquina vêm sendo amplamente empregadas para detectar padrões maliciosos e responder a tais ataques de forma automatizada e eficiente. No entanto, a eficácia desses sistemas está diretamente relacionada à qualidade das representações utilizadas nos dados de entrada, tornando a etapa de seleção de características um componente crítico do processo de modelagem.

A seleção de características tem como principal finalidade identificar, entre todas as variáveis observadas, aquelas que são mais informativas para a tarefa de predição, reduzindo a dimensionalidade dos dados e mitigando problemas como sobreajuste, redundância e multicolinearidade. Além disso, ao eliminar atributos irrelevantes ou ruidosos, essa etapa pode proporcionar ganhos substanciais em desempenho computacional, interpretabilidade e robustez dos classificadores (Pascoal et al., 2012). Na literatura (Heigl et al., 2021), (Emirmahmutoğlu e Atay, 2025), observa-se uma predominância de métodos paramétricos ou sensíveis à presença de *outliers* e a distribuições assimétricas, o que limita sua aplicação em cenários reais de tráfego de rede, onde essas condições são frequentemente observadas.

Neste contexto, esta dissertação propõe uma abordagem não-paramétrica de seleção de características, baseada na integração de métricas estatísticas robustas, como: entropia de Shannon (Pascoal et al., 2012), correlação de Spearman (Palamidessi e Romanelli, 2020), distância de Bhattacharyya modificada (Vergara e Estévez, 2015), distância de Mahalanobis robusta e informação mútua ajustada (Berbiche e Alami, 2024), organizadas em um fluxo de três estágios que combina filtragem, clusterização e ranqueamento. Ao ponderar simultaneamente aspectos de relevância, redundância e separabilidade, o modelo proposto busca selecionar subconjuntos de atributos capazes de maximizar o desempenho de classificadores, sem impor pressupostos de normalidade ou linearidade sobre os dados.

A proposta foi implementada em Python, utilizando bibliotecas consolidadas de aprendizado de máquina, e validada experimentalmente por meio de estudos de caso com três bases públicas representativas: Conjunto de dados criado e usado em (Sharafaldin, Lashkari e Ghorbani, 2019) (CICDDoS-2019), Conjunto de dados criado e usado em (Moustafa e Slay, 2015) (UNSW-NB15) e Conjunto de dados criado e usado em (Nascimento et al., 2021b) (HPC-Lab) . Utilizou-se o classificador *Random Forest* como referência, com validação cruzada estratificada e análise estatística dos resultados baseada em testes não-paramétricos de Friedman e pós-teste de Nemenyi. Os experimentos demonstraram reduções significativas na quantidade de atributos selecionados, acompanhadas por melhorias estatisticamente relevantes (com p-

value $< 0{,}05$) em métricas como acurácia, F1 e AUC-ROC, além de ganhos em tempo de processamento.

Com esta contribuição, espera-se avançar na construção de sistemas de detecção mais eficientes, interpretáveis e resilientes, reforçando o papel da estatística robusta e da análise não-paramétrica como fundamentos para a segurança computacional baseada em dados.

1.1 MOTIVAÇÃO

A crescente sofisticação dos ataques cibernéticos, em particular os DDoS, tem desafiado de forma contundente os mecanismos tradicionais de segurança em redes modernas (Ali et al., 2021). Neste contexto, os IDS baseados em AM têm se destacado como soluções promissoras (Javaid et al., 2020). Contudo, o desempenho desses sistemas depende fortemente da etapa de seleção de características, a qual visa reduzir a dimensionalidade dos dados, eliminar atributos redundantes ou irrelevantes e, consequentemente, melhorar a eficiência e a acurácia dos modelos preditivos (Nguyen et al., 2023).

Um dos principais desafios enfrentados na construção de IDS eficazes para Redes Definidas por Software (SDN) reside na natureza dos dados coletados. Esses dados frequentemente apresentam características não-paramétricas e distribuições desbalanceadas, o que compromete a aplicabilidade de técnicas tradicionais de seleção de características, muitas das quais assumem distribuições gaussianas (Singh, Singh e Roy; Zainudin et al.; Alhakami et al.; Liu et al., 2021, 2023, 2019, 2021). Tais pressupostos, embora convenientes para fins analíticos, raramente refletem a complexidade e a variabilidade intrínsecas ao tráfego de rede em ambientes reais, altamente dinâmicos e heterogêneos.

Nesse cenário, métodos não-paramétricos de seleção de características emergem como uma abordagem robusta e adaptável. Ao não dependerem de suposições estritas sobre a distribuição dos dados, essas técnicas são mais resilientes à presença de outliers, à assimetria entre classes e à variabilidade estrutural observada em cenários de big data (Sayed et al., 2022). Além disso, tais métodos se mostram particularmente adequados para lidar com conjuntos de dados de alta dimensionalidade, como os utilizados em tarefas de detecção de intrusões.

Estudos recentes (Das et al.; Upadhyay et al., 2021, 2021) têm demonstrado o potencial das abordagens não-paramétricas na identificação de atributos relevantes para a detecção de ataques em redes SDN. Esses métodos têm contribuído significativamente para o aumento da acurácia e para a redução das taxas de falsos positivos, aspectos fundamentais para a viabilidade prática de Sistema de Detecção de Intrusões (IDS) em ambientes críticos (Sayed et al.; Upadhyay et al., 2022, 2021). Tais avanços motivam a investigação de modelos que explorem o poder discriminativo das técnicas não-paramétricas, com vistas ao desenvolvimento de soluções mais eficazes, escaláveis e sensíveis ao contexto de dados reais.

1.2 OBJETIVOS

Este trabalho apresenta uma abordagem não paramétrica de seleção de características para sistemas de detecção de DDoS. O método baseia-se em métricas estatísticas robustas a *outliers* e a distribuições assimétricas — entropia de Shannon, correlação de Spearman, distância de Bhattacharyya modificada, distância de Mahalanobis robusta e informação mútua ajustada — organizadas em um fluxo de três estágios capaz de reduzir a dimensionalidade de conjuntos de dados de tráfego de rede sem comprometer (e, em muitos casos, elevando) o desempenho dos classificadores. Ao ponderar simultaneamente relevância, redundância e separabilidade, o modelo contribui para ganhos de eficiência computacional e para a melhoria da taxa de detecção em cenários de segurança de redes.

A proposta foi validada nas bases CICDDoS-2019, UNSW-NB15 e HPC-Lab, utilizando o classificador RF como referência, com validação cruzada estratificada e testes estatísticos não paramétricos de Friedman, seguidos pelo pós-teste de Nemenyi. Os resultados experimentais demonstram reduções substanciais no número de atributos e melhorias significativas nas métricas de acurácia, F1 e AUC-ROC (p-*value* $< 0{,}05$), além de redução no tempo de processamento.

Especificamente, os objetivos desta dissertação são:

- Propor uma estratégia para seleção de características totalmente não-paramétrica, mais aderente aos reais problemas da engenharia de *features*;
- Demonstrar que a elaboração dessa estratégia foi feita utilizando critérios explicáveis para a escolha e adaptação de métricas e métodos existentes;
- Avaliar quantitativamente o impacto da redução de dimensionalidade sobre desempenho, tempo de processamento e estabilidade das seleções;
- Apresentar os resultados experimentais ao comparar o método proposto a técnicas consagradas de seleção de características em múltiplas bases públicas e cenários de detecção de DDoS, empregando métricas como acurácia, precisão, F1, AUC-ROC e custo computacional;

1.3 ESTRUTURA DO DOCUMENTO

Os próximos capítulos do documento estão descritos brevemente a seguir.

- **Capítulo 2 – Referencial Teórico:** Introduz os conceitos necessários, baseados em uma revisão bibliográfica para compreensão das ideias apresentadas. Explora conceitos relacionados a seleção estatística de features, detecção de ameaças com modelos de Machine Learning, técnicas de redução de dimensionalidade e métodos de seleção de características;

- **Capítulo 3 – Trabalhos Relacionados:** Analisa pesquisas anteriores pertinentes ao tema, ressaltando suas contribuições e distinções em relação à proposta deste trabalho;
- **Capítulo 4 – Método Proposto:** Descreve detalhadamente a abordagem metodológica adotada, incluindo as etapas de análise, desenvolvimento do modelo e configuração das ferramentas utilizadas;
- **Capítulo 5 – Modelo proposto**: Apresenta a arquitetura do modelo desenvolvido, suas componentes e critérios de funcionamento;
- **Capítulo 6 – Resultados Experimentais:** Apresenta e discute os resultados obtidos nos experimentos, com base nos estudos de caso realizados;

  Capítulo 7 – Conclusão: Resume as principais contribuições da pesquisa e propõe direções para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo estabelece o arcabouço teórico fundamental que sustenta o desenvolvimento e a análise da presente dissertação. Discutem-se os sistemas de detecção de ameaças baseados em aprendizagem de máquina, os desafios impostos pela natureza não paramétrica dos dados em cibersegurança, e as principais abordagens e métricas utilizadas na FS.

2.1 APRENDIZADO DE MÁQUINA NA DETECÇÃO DE *MALWARE*

Malware refere-se a qualquer *software* projetado para causar danos ou explorar sistemas computacionais, incluindo vírus, *worms*, *trojans*, *ransomware*, entre outros. Dada sua prevalência em ambientes interconectados, representa ameaça crítica à segurança de indivíduos, organizações e nações (Bensaoud, Kalita e Bensaoud, 2024). A detecção eficaz de *malware* é, portanto, uma necessidade fundamental.

O AM tem se destacado como abordagem promissora na detecção de *malware*, permitindo análises automatizadas de grandes volumes de dados para identificar padrões maliciosos (Nawshin et al., 2024). Com sua capacidade de adaptação a novas ameaças, os algoritmos de ML têm sido amplamente empregados para aumentar a acurácia e reduzir falsos positivos em sistemas de detecção (Kim et al., 2023).

Dentre as principais aplicações de ML na detecção de *malware*, destacam-se:

- *Detecção baseada em assinaturas*: automatiza a geração e o reconhecimento de padrões de *malware* conhecidos (Stevens et al., 2024).
- *Detecção baseada em comportamento*: identifica atividades maliciosas com base em anomalias comportamentais, úteis na detecção de ameaças desconhecidas ou *zero-day* (Galli et al., 2024).
- *Classificação de malware*: algoritmos supervisionados (e.g., árvores de decisão, redes neurais) distinguem entre amostras benignas e maliciosas com base em atributos extraídos (Bensaoud e Kalita, 2024).
- *Classificação por família*: categoriza *malware* com base em estrutura de código, comportamento ou trechos recorrentes, auxiliando em análises forenses e estratégias de resposta (Zhang, Liu e Liu, 2024).
- *Detecção de variantes*: permite identificar versões modificadas de ameaças conhecidas, promovendo resposta mais ágil a novas cepas (Madamidola, Ngobigha e Ez-zizi, 2024).
- *Seleção e extração de características*: algoritmos de ML auxiliam na identificação automática dos atributos mais relevantes, reduzindo a dimensionalidade e custo computacional (Maribana, Chindipha e Brown, 2023).

- *Abordagens de comitê* (*ensemble*): combinam múltiplos modelos para aumentar a robustez, minimizar erros e tratar desbalanceamento de dados (Muthusamy e Charles, 2025).
- *Robustez a ataques adversariais*: técnicas como *adversarial training* visam aumentar a resiliência dos modelos a manipulações intencionais de entrada (Li et al., 2024).

2.2 OTIMIZAÇÃO NA DETECÇÃO DE *MALWARE*

A otimização na detecção de *malware* envolve a aplicação de métodos computacionais que visam maximizar a eficácia e eficiência dos sistemas de identificação de ameaças. Tais métodos buscam soluções ótimas para problemas como seleção de atributos, configuração de modelos e alocação de recursos computacionais (Shar et al., 2024).

No contexto de sistemas inteligentes de segurança, a otimização contribui de forma significativa para o equilíbrio entre acurácia, robustez e custo computacional. As principais estratégias incluem:

- **Seleção de características**: algoritmos de otimização identificam subconjuntos de atributos que maximizam o desempenho preditivo ou minimizam erros de classificação, como falsos positivos e negativos (Hasan et al., 2025).
- **Alocação de recursos**: a distribuição eficiente de recursos — Unidade Central de Processamento (CPU), memória, largura de banda — é essencial para garantir desempenho em tempo real, especialmente em cenários de grande volume de dados ou arquiteturas restritas, como dispositivos Internet das Coisas (IoT) (Li e Zhao, 2024).

2.3 DESAFIOS E FUNDAMENTOS DA MODELAGEM ESTATÍSTICA NÃO PARAMÉTRICA EM CIBERSEGURANÇA

Aplicações em cibersegurança, como análise de tráfego de rede e biometria, frequentemente lidam com dados assimétricos, com *outliers* e relações não lineares, inviabilizando pressupostos de testes paramétricos clássicos, como o teste t deStudent ou Análise de Variância (ANOVA) (Alasmar et al., 2021). Tais testes pressupõem normalidade e homocedasticidade; quando essas condições são violadas — cenário comum em tráfego de rede de cauda pesada — a validade inferencial e a capacidade de generalização dos modelos ficam comprometidas (Arp et al., 2022).

Nesse contexto, a modelagem estatística **não paramétrica** apresenta-se como alternativa robusta. Métodos baseados em ranques (e.g., Wilcoxon–Mann–Whitney, Kruskal–Wallis, Friedman), técnicas de *bootstrapping* e estimadores de densidade via *kernel* dispensam supo-

sições rígidas sobre a distribuição dos dados e, por isso, mostram-se adequados à variabilidade e heterogeneidade de dados reais(Chu, Ling e Yuan, 2024).

Essas abordagens também desempenham papel central na FS em contextos de alta dimensionalidade: métricas como correlação de Spearman, distância de Bhattacharyya e Informação Mútua Ajustada (AMI), em versões não paramétricas, permitem avaliar relevância e redundância de atributos com maior acurácia (Li e Fard, 2022). Com isso, é possível identificar subconjuntos informativos que aprimoram não apenas o desempenho preditivo de modelos de aprendizado de máquina, mas também sua interpretabilidade e eficiência computacional — fatores decisivos na construção de sistemas de Inteligência Artificial(IA) robustos e transparentes para detecção proativa de ameaças (Arreche, Guntur e Abdallah, 2024).

2.3.1 A Necessidade de Abordagens Não Paramétricas na Detecção de Ameaças

Sistemas de detecção baseados em Aprendizado de Máquina (AM) têm se consolidado como ferramentas essenciais no combate a ataques de negação de serviço distribuídos (DDoS). Ao aprender padrões de tráfego legítimo, tais sistemas conseguem identificar desvios em tempo real, mesmo em redes de alta velocidade (Nguyen e Armitage, 2022).

Entretanto, a alta **dimensionalidade** dos dados — composta por métricas derivadas de HPC, estatísticas de pacotes, fluxos e Qualidade de Serviço — introduz desafios substanciais: maior custo computacional, dificuldade de interpretação dos modelos e limitação de escalabilidade(Taherkordi, Mohammadi e Franke, 2020).

Embora estudos reportem exatidão superior a 90% com diversas técnicas de seleção ou extração de atributos (Kuruvila; Li, 2021, 2022), no que cabe aos eventos de microarquitetura, a coleta de múltiplos eventos microarquiteturais exige sondas adicionais e, em dispositivos embarcados, a leitura simultânea de HPC é bastante restrita (geralmente entre 2 e6 eventos)(Das et al., 2019). Estratégias como o *multiplexing* de contadores têm sido utilizadas para contornar essa limitação, mas adicionam complexidade e latência. (Zhang, Liu e Liu, 2024). Já em dados de tráfego de redes, as vantagens em tempo de detecção, por exemplo, ganham enorme vantagem com uma estratégia bem definida para seleção de atributos.

Neste cenário, abordagens **não paramétricas** para FS tornam-se particularmente promissoras. Por não dependerem de pressupostos sobre a distribuição dos dados, essas técnicas oferecem maior robustez frente a assimetrias, outliers e variabilidade estrutural — aspectos comuns em dados de segurança de rede.

Esta dissertação propõe um fluxo completo de análise que combina:

1. reduzir drasticamente o número de atributos mantendo (ou aumentando) a capacidade de detecção;
2. minimizar tempo e memória de execução para viabilizar implantação em tempo real;
3. garantir interpretabilidade por meio de métricas estatísticas transparentes.

2.3.2 Métricas Estatísticas Não Paramétricas Utilizadas

Conjuntos de tráfego de rede raramente obedecem a distribuições normais; logo, adotamos métricas estatísticas não parametrizadas que dispensam suposições de normalidade e são menos sensíveis a *outliers*.

2.3.2.1 Distância de Medida de Similaridade

A Medida de similaridade baseada em distância (DMS) é uma métrica estatística não paramétrica empregada para quantificar a dissimilaridade de uma instância em relação ao conjunto de dados, após a exclusão de uma determinada característica. Seu objetivo é avaliar a relevância de um atributo com base no impacto estrutural de sua remoção sobre os dados(Mitra, Murthy e Pal, 2002).

A DMS é definida por:

$$\mathrm{DMS}(x) = (x_{-i} - \mathrm{med}_{-i})^T R_{-i}^{-1} (x_{-i} - \mathrm{med}_{-i}),$$

em que:

- x_{-i} representa a instância sem o atributo f_i,
- med_{-i} é o vetor de medianas das demais variáveis,
- R_{-i}^{-1} é a inversa da matriz de correlação de Spearman dos atributos restantes.

Ao empregar medianas e correlação de Spearman em vez de média e covariância, a DMS torna-se robusta a *outliers*, assimetrias e relações não lineares—características frequentes em dados de tráfego de rede(Feng, Lu e Zhang, 2024).

Valores elevados de DMS indicam que a remoção do atributo f_i causa uma distorção significativa na coerência da instância com relação ao conjunto, sugerindo sua importância estrutural.

Entre suas aplicações destacam-se:

- Seleção não supervisionada de características;
- Detecção de anomalias;
- Análise de redundância e relevância de variáveis;
- Pré-processamento em pipelines de aprendizado de máquina em contextos de segurança e saúde digital.

2.3.2.2 Entropia.

A **entropia**, conforme definida por Shannon, quantifica a incerteza associada a uma variável aleatória (Shannon, 1948). Em aprendizado de máquina, é amplamente utilizada para avaliar a variabilidade de atributos e sua capacidade discriminativa em tarefas supervisionadas.

Seja X uma variável discreta com distribuição $P(x)$. Sua entropia é dada por:

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x), \tag{2.1}$$

onde $\mathcal{X}$ é o conjunto de valores possíveis de X. O logaritmo geralmente é na base 2 (bits).

A entropia atinge valor máximo quando X é uniformemente distribuída, e é mínima (zero) quando X é determinística. No contexto de FS, a entropia permite avaliar o quão informativo é um atributo sobre a variável-alvo.

O IG formaliza essa contribuição:

$$IG(T, X) = H(T) - H(T|X), \tag{2.2}$$

sendo T a variável-alvo e $H(T|X)$ a entropia condicional após observar X.

Entre suas propriedades, destacam-se:

- não negatividade;
- valor máximo sob distribuição uniforme;
- invariância a permutações;
- aplicabilidade a variáveis categóricas ou discretizadas.

2.3.2.3 Correlação de Spearman.

A **correlação de Spearman** é uma medida não paramétrica que avalia o grau de associação *monotônica* entre duas variáveis, baseada na ordenação de seus valores (Spearman, 1904). Diferentemente da correlação de Pearson, não exige normalidade nem linearidade, sendo robusta a *outliers* e assimetrias - características comuns em dados de tráfego de rede(Feng, Lu e Zhang, 2024).

Dada duas variáveis X e Y, a correlação de Spearman é definida por:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X_r, Y_r)}{\sigma_{X_r} \sigma_{Y_r}}, \tag{2.3}$$

onde X_r e Y_r representam os postos de X e Y, respectivamente. O coeficiente varia de -1 (correlação negativa perfeita) a $+1$ (positiva perfeita), sendo zero na ausência de associação monotônica.

Em FS, essa métrica é útil para:

- Detectar **redundância entre atributos**, auxiliando na eliminação de variáveis altamente correlacionadas;
- Capturar **dependências não lineares** de natureza monotônica;
- Servir de base para métricas robustas, como a DMS.

2.3.2.4 Ganho de Informação.

O Ganho de Informação (IG) é uma métrica derivada da teoria da informação que quantifica a redução de incerteza de uma variável-alvo T ao conhecer o valor de uma variável X (Quinlan, 1986). É amplamente utilizado em FS e construção de DT.

Formalmente, é definido como:

$$IG(T, X) = H(T) - H(T|X), \tag{2.4}$$

em que $H(T)$ é a entropia de T e $H(T|X)$ representa a entropia condicional após observar X. Valores mais altos de IG indicam que X fornece maior informação sobre T.

Na seleção de atributos, o IG é utilizado para:

- Classificar variáveis por relevância preditiva;
- Eliminar atributos irrelevantes;
- Guiar algoritmos como ID3, C4.5 e RF.

Entre suas vantagens:

- Capta relações não lineares;
- Não exige normalidade nem escalas padronizadas;
- Funciona bem com variáveis categóricas ou discretizadas.

Por outro lado, o IG tende a favorecer atributos com alta cardinalidade. Para corrigir esse viés, métricas como o IG normalizado e a AMI têm sido empregadas (Estévez et al., 2009).

O IG permanece uma ferramenta central em pipelines de aprendizado supervisionado, especialmente em domínios como cibersegurança, onde a identificação de atributos altamente informativos é crítica para a detecção de ameaças.

2.3.2.5 Distância de Bhattacharyya.

A **Distância de Bhattacharyya** quantifica a similaridade entre duas distribuições de probabilidade e é amplamente utilizada para medir a *separabilidade entre classes* em problemas de classificação e FS (Fukunaga, 1990).

Dadas duas distribuições $f(x)$ e $g(x)$, a distância é definida como:

$$D_B(f,g) = -\ln\left(\int \sqrt{f(x)\,g(x)}\,dx\right), \tag{2.5}$$

sendo o termo dentro do logaritmo conhecido como *coeficiente de Bhattacharyya*, que mede a sobreposição entre distribuições. Quanto maior o D_B, maior a dissimilaridade entre as classes.

No caso de distribuições normais multivariadas, a métrica pode ser aproximada por:

$$D_B \approx \frac{1}{8}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) + \frac{1}{2}\ln\left(\frac{|\Sigma|}{\sqrt{|\Sigma_1||\Sigma_2|}}\right), \tag{2.6}$$

com $\Sigma = \frac{1}{2}(\Sigma_1+\Sigma_2)$ representando a covariância média.

Na seleção de atributos, D_B é usado para:

- Avaliar a capacidade discriminativa de um atributo;
- Priorizar variáveis que maximizam a separação entre classes;
- Complementar métricas como entropia e correlação.

As principais vantagens da métrica incluem:

- Alta sensibilidade à sobreposição entre distribuições;
- Aplicabilidade a espaços multivariados e dados empíricos;
- Adequação a contextos não paramétricos com *outliers* e assimetrias.

2.3.2.6 Teste de Friedman.

O **Teste de Friedman** é um teste estatístico não paramétrico para comparação de múltiplos algoritmos avaliados sobre os mesmos blocos (e.g.,*datasets* ou *folds* de validação cruzada), sendo uma alternativa robusta à ANOVA de medidas repetidas (Friedman, 1937). É indicado quando as suposições de normalidade ou homocedasticidade são violadas, especialmente em amostras pequenas ou com *outliers*(Demšar, 2006a).

Dado k algoritmos e n blocos, cada algoritmo é ranqueado dentro de cada bloco. A estatística do teste é:

$$Q = \frac{12}{n\,k(k+1)}\left(\sum_{j=1}^{k} R_j^2\right) - 3n(k+1), \tag{2.7}$$

onde R_j é a soma dos postos do j-ésimo algoritmo. Sob a hipótese nula H_0 (sem diferenças entre os algoritmos), Q segue uma distribuição qui-quadrado com $k-1$ graus de liberdade:

$$Q \sim \chi^2_{k-1}. \tag{2.8}$$

Na prática, o teste é amplamente utilizado para:

- Comparar o desempenho de modelos preditivos em múltiplos conjuntos;
- Validar diferenças entre métodos de FS;
- Suportar análises estatísticas em experimentos repetidos ou pareados.

Entre suas vantagens estão:

- Ausência de pressupostos paramétricos;
- Aplicabilidade a dados assimétricos ou com ruído;
- Existência de variantes, como o teste de Iman-Davenport, que melhora sua aproximação.

Nesta dissertação, o Teste de Friedman é adotado para validar estatisticamente as diferenças entre subconjuntos de atributos selecionados, assegurando que os ganhos observados sejam significativos e não resultem de flutuações aleatórias.

2.4 SELEÇÃO DE CARACTERÍSTICAS

A FS é uma etapa essencial no desenvolvimento de modelos de aprendizado de máquina, com o objetivo de identificar o subconjunto mais relevante de atributos para representar os dados sem perda informacional significativa. Ao remover características irrelevantes, redundantes ou ruidosas, a FS contribui para a **redução da dimensionalidade**, maior **eficiência computacional**, aumento da **exatidão ou acurácia preditiva** e melhora na **interpretabilidade** dos modelos (Li et al.; Nguyen et al., 2022, 2023).

Em domínios de alta dimensionalidade — como tráfego de rede, bioinformática e segurança cibernética — atributos desnecessários podem comprometer a generalização dos modelos, aumentar o custo computacional e intensificar o risco de *overfitting*. A FS atua mitigando esses efeitos ao selecionar apenas as variáveis que contribuem efetivamente para a discriminação entre classes (Ayad, Fahmy e Abdelrahman, 2024).

Objetivos Principais da Seleção de Características

A FS busca:

- **Reduzir a dimensionalidade**, mantendo ou melhorando o desempenho preditivo;
- **Eliminar atributos irrelevantes e redundantes**, focando nas variáveis mais informativas;
- **Aumentar a robustez e generalização**, reduzindo o risco de *overfitting*;
- **Diminuir o tempo de treinamento e inferência**, otimizando recursos;
- **Melhorar a interpretabilidade**, facilitando a compreensão dos modelos.

As principais abordagens para FS incluem os métodos *Filter*, *Wrapper* e *Embedded*. Além disso, métodos *Híbridos* combinam vantagens dessas estratégias para alcançar melhor equilíbrio entre desempenho e custo computacional.

2.4.1 Classificação dos Métodos de Seleção de Características

2.4.2 Métodos *Filter (Filter Methods)*

Métodos de filtro avaliam atributos independentemente do modelo preditivo, utilizando métricas estatísticas como entropia, correlação, variância e informação mútua (Guyon e Elisseeff, 2003). São rápidos, escaláveis e adequados para grandes volumes de dados, mas não consideram interações entre variáveis.

2.4.3 Métodos *Wrapper (Wrapper Methods)*

Métodos *wrapper* utilizam o modelo de aprendizado como caixa-preta para avaliar diferentes subconjuntos de atributos com base em seu desempenho preditivo (Kohavi e John, 1997). Apesar de geralmente mais precisos que os filtros, apresentam maior custo computacional, especialmente em cenários de alta dimensionalidade.

2.4.4 Métodos Embutidos (Embedded Methods)

Métodos embutidos realizam a seleção de atributos durante o treinamento do modelo, aproveitando mecanismos internos para induzir esparsidade ou calcular importâncias (Tibshirani, 1996). São geralmente mais eficientes que os *wrappers* e capturam interações entre atributos, embora possam estar restritos às hipóteses do algoritmo subjacente.

2.4.5 Métodos Híbridos (Hybrid Methods)

Métodos híbridos combinam abordagens *filter* e *wrapper* em múltiplas etapas, buscando equilibrar desempenho preditivo e custo computacional (Bolón-Canedo, Sánchez-Maroño e Alonso-Betanzos, 2015). Tipicamente, realizam uma pré-filtragem inicial para reduzir o espaço de busca e, em seguida, aplicam métodos mais custosos apenas sobre subconjuntos promissores.

2.4.6 Critérios de Avaliação em FS

A eficácia de métodos de FS pode ser avaliada com base em múltiplos critérios:

- **Desempenho preditivo**, medido por métricas como exatidão, F1-Score e AUC;
- **Estabilidade da seleção**, frente a variações nos dados ou amostragens;
- **Redução da dimensionalidade**, refletida na queda no número de atributos e no tempo de processamento;
- **Significância estatística**, obtida por testes não paramétricos como o de Friedman (Demšar, 2006a).

2.4.7 Importância na Cibersegurança

No contexto de segurança cibernética, FS é crucial para identificar atributos discriminativos em tráfego de rede, eventos do sistema ou contadores de *hardware*. Técnicas **não paramétricas e robustas** são preferidas por dispensarem hipóteses sobre a distribuição dos dados e tolerarem melhor ruídos e valores extremos(Ayad, Fahmy e Abdelrahman; Suhaimi et al., 2024, 2022).

Assim, FS vai além do pré-processamento: constitui um componente estratégico para garantir **eficiência, interpretabilidade e desempenho** em soluções inteligentes, especialmente em sistemas de detecção de intrusões e ambientes de missão crítica.

3 TRABALHOS RELACIONADOS

A literatura recente sobre FS em detecção de ameaças evoluiu de filtros simples, voltados a um único *dataset*, para pipelines híbridos capazes de operar em tempo real, em múltiplas bases e até mesmo fundindo sinais heterogêneos (tráfego + HPC). Esta seção apresenta estudos que se dedicaram à etapa de seleção de características para de modelos de aprendizagem de máquina no contexto da detecção de ameaças.

3.1 FS PARA DETECÇÃO DE INTRUSÃO EM REDES

Em (Yin et al., 2023) os autores combinam Ganho de Informação (IG) com Eliminação Recursiva de Atributos (RFE) sobre o UNSW-NB15. O método selecionou 18 de 49 variáveis, elevando a F1-score de um Perceptron Multicamadas (MLP) de 94,3% para 97,6% e reduzindo 47% do tempo de inferência. A validação 10-*fold* demonstrou estabilidade mesmo sob forte desbalanceamento de classes, mas o custo do wrapper Eliminação Recursiva de Atributos (RFE) ainda inviabiliza atualização em fluxo contínuo. Em (Tripathi e Sharma, 2024) os autores utilizaram a importância do RF como pré ranqueamento e, depois, o filtro Mínima Redundância e Máxima Relevância (mRMR), o conjunto de atributos no caiu 55%. A exatidão alcançou 96,2% com queda de 55ms na latência média de decisão. Testes de Friedman confirmam que o subconjunto mRMR + RF supera IG e χ^2 com significância $<0{,}01$, mas o artigo não verifica robustez adversarial.

3.2 FS PARA DETECÇÃO DE ATAQUES DDoS

Em (Yu, Chen e Li, 2024), por meio de Otimização por Enxame de Partículas Binária (BPSO) em 84 features do , selecionaram 24 atributos que, em Impulsionamento de Gradiente Extremo (XGBoost), renderam precisão 99,1% e duplicaram o*throughput* de classificação. Os autores mostram convergência em 36 iterações, porém admitem sensibilidade a parâmetros de inércia do enxame.

Em (Chanu e Sarma, 2023), os autores propuseram um *voting hybrid* (correlação, χ^2 e *ReliefF*) que reteve 27features do mesmo *dataset*, atingindo F1 98,4% e reduzindo 42% da latência em controladores SDN. Apesar do ganho, não há análise estatística formal nem teste cruzado em bases adicionais.

O estudo (Han, Zhang e Liu, 2024), através do algoritmo *Marginal-Gain FS with RF* (MFS–RF) priorizaram variáveis cujo acréscimo marginal de AUC ultrapassa um limiar adaptativo. Em tráfego OpenFlow sintético, bastaram 15 das 76 métricas para manter AUC 98,7% e cortar 60% do consumo de memória do controlador, mas o estudo se restringe a cenários laboratoriais.

Em (Ogaili et al., 2022), os autores exploraram três meta-heurísticas (Salp Swarm, Gray Wolf e PSO) sobre o , o trabalho identifica subconjuntos com até 90% de redução de dimensionalidade e obtém exatidão de 99,9% em SVM e KNN. Embora o resultado seja expressivo, falta comparação direta com heurísticas de menor custo e medidas de robustez a *concept-drift*.

O estudo (Zhang, Han e Liu, 2024), focado em SDN 5G, o estudo aplica um FS proprietário a cinco conjuntos e demonstra que a filtragem proposta acelera a detecção sem perda de F1 ($\geq 98\%$). O artigo detalha impacto em latência de $20\,\mu$s por fluxo, mas não libera o *codebase*, dificultando a reprodutibilidade.

3.3 FS PARA BOTNETS E IOT

Em (Al-Sarem et al., 2022), os autores propuseram um pipeline duplo EO + BRO reduz 46→11 atributos. Com *LightGBM*, AUC chega a 0,993 e FPR a 0,8%. Os autores analisam *concept-drift* em 12 meses de tráfego real, mas não discutem consumo de energia em *gateways* Internet das Coisas (IoT).

No estudo (Pereira e Silva, 2025), com χ^2 seguido de *Sequential Forward Selection*, foi treinado um SVM nos conjuntos IoT-23/BoT-IoT mantidos em ARM Cortex-A53. O *Recall* permaneceu $\geq 97\%$ para tráfego de 100 kpps, consumindo apenas 31 MB de RAM; não há, porém, comparações com outros algoritmos de seleção para justificar a escolha por esse método.

Em (Chen et al., 2024), os autores introduziram o algoritmo GQBWSSA (versão aprimorada do Salp Swarm) que, no CICIoT2023, manteve exatidão de 99,7% em binária e 99,4% em multi-classe, reduzindo 80% dos atributos. O artigo discute tempo de treino 6× menor que GA tradicional, embora não considere métricas de consumo energético em dispositivos de borda.

Em (Ma et al., 2025), os autores propuseram um modelo para selecionar 20% das features em cinco bases IoT, usando as *features* para fine-tunar Modelo de linguagem de grande porte (LLM), gerando amostras sintéticas que corrigem desbalanceamento. A abordagem eleva macro-F1 em 4,2p.p. sobre LightGBM puro e reduz redundância >80%; entretanto, a dependência de LLM amplia custo computacional em gateways.

3.4 FS BASEADA EM HPC E ABORDAGENS MULTIMODAIS

Em (Nascimento, Lima e Pereira, 2023), os autores utilizaram correlação para selecionar 8 HPC entre 49, o modelo RF detectou HTTP Flood no HPC-Lab com precisão 97,8% e overhead 3× menor de coleta, mostrando viabilidade de monitoramento micro-arquitetural.

No estudo (Alduailij et al., 2022) foi aplicada MI para decidir componentes de uma

Análise de Componentes Principais (PCA) guiado, fusionando HPCs com tráfego do conjunto CICDDoS-2019. O

3.5 SÍNTESE CRÍTICA

A abordagem não-paramétrica proposta nesta dissertação endereça lacunas ao *(i)* fundir sinais multimodais, *(ii)* validar via teste de Friedman em bases distintas e *(iii)*, provendo uma visão sistêmica ainda pouco explorada na literatura.

A Tabela 1 sintetiza uma análise comparativa entre os principais trabalhos recentes sobre FS aplicados à detecção de ataques DDoS e a proposta desta dissertação. Os critérios de comparação abrangem sete dimensões consideradas fundamentais para avaliar a abrangência, robustez e reprodutibilidade dos métodos: suporte a dados **multimodais**, uso de métricas **não paramétricas**, avaliação em **múltiplos conjuntos de dados** (*multi- dataset*), **validação estatística** dos resultados e consideração de **latência ou overhead computacional** (Lat./OH).

Observa-se que a maioria dos trabalhos analisados apresenta limitações como ausência de validação estatística formal, uso restrito a um único *dataset* e falta de medidas robustas contra perturbações adversariais. Poucos estudos exploram métricas não paramétricas ou disponibilizam seus repositórios de forma completa. Em contraste, este trabalho se destaca por cumprir integralmente todos os critérios avaliados. Tais aspectos reforçam a originalidade e a contribuição técnica desta dissertação no contexto de sistemas inteligentes para segurança de redes.

Tabela 1 – Comparação entre este trabalho e estudos de FS para detecção de DDoS

Trabalho	**Multimodal**	**Não-param.**	**Multi-dataset**	**Validação estat.**	**Lat./OH**
(Yu, Chen e Li, 2024)	Não	Não	Não	Não	Sim
(Chanu e Sarma, 2023)	Não	Parcial	Não	Não	Sim
(Han, Zhang e Liu, 2024)	Não	Parcial	Não	Não	Sim
(Ogaili et al., 2022)	Não	Parcial	Não	Não	Não
(Zhang, Han e Liu, 2024)	Não	Não	Sim	Não	Sim
(Nascimento, Lima e Pereira, 2023)	Não	Sim	Não	Não	Sim
(Alduailij et al., 2022)	Sim	Parcial	Não	Não	Não
Este trabalho(2025)	Sim	Sim	Sim	Sim	Sim

Legenda: **Sim** = presente; **Parcial** = implementado de forma limitada (ex.: apenas parte da métrica ou repositório incompleto); **Não** = ausente.

4 MÉTODO PROPOSTO

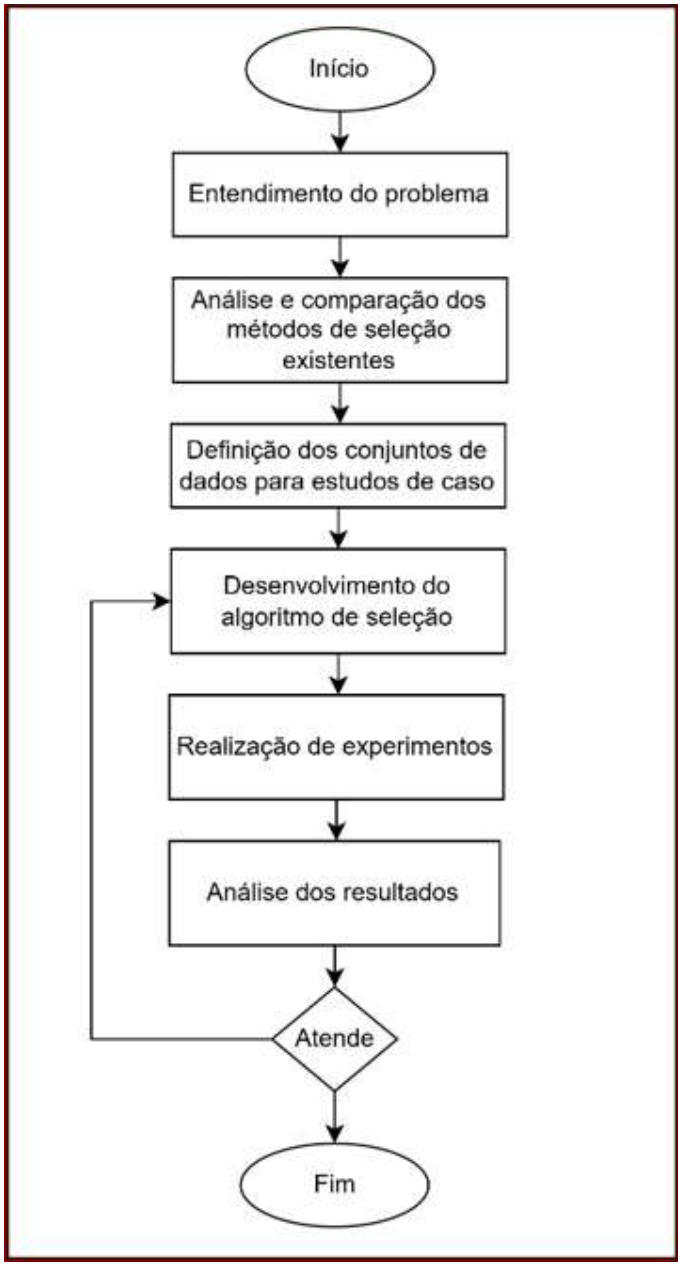
Este capítulo apresenta o método proposto para o desenvolvimento de um modelo de seleção de características com base em métricas estatísticas não paramétricas, aplicado a IDS. A abordagem visa reduzir a dimensionalidade de conjuntos de dados heterogêneos, típicos de ambientes de cibersegurança, preservando ou ampliando a capacidade de discriminação de padrões maliciosos.

Inicialmente, é apresentada uma visão geral das etapas que compõem o processo metodológico, destacando-se as principais atividades desenvolvidas para atingir os objetivos da pesquisa. Na sequência, cada etapa é descrita em detalhes.

A metodologia foi fundamentada em uma análise da literatura especializada, considerando os principais métodos de seleção de atributos utilizados em sistemas inteligentes de detecção de ameaças. Os critérios de comparação adotados — incluindo robustez estatística, eficiência computacional e viabilidade de implantação em tempo real — nortearam as decisões de projeto e culminaram na elaboração de uma arquitetura de seleção de características adequada a dados de tráfego com natureza não paramétrica.

A Figura 1 apresenta uma visão geral da metodologia proposta.

Figura 1 – Visão do método proposto



4.1 VISÃO GERAL

O método proposto adota uma abordagem não paramétrica para seleção de características. Essa abordagem fundamenta-se em técnicas estatísticas que dispensam pressupostos sobre a distribuição dos dados, tornando-se particularmente adequada para contextos reais em que as hipóteses clássicas de normalidade não são verificadas. As principais etapas são descritas a seguir:(Hollander, Wolfe e Chicken, 2015).

4.2 ENTENDIMENTO DO PROBLEMA

Esta seção descreve a investigação conduzida para compreender o problema, delimitar o escopo da solução e identificar os requisitos estatísticos, métricas e parâmetros relevantes para o desenvolvimento do modelo proposto. São detalhadas a revisão da literatura especializada, a implementação de experimentos preliminares com diferentes abordagens de seleção de características e a definição dos conjuntos de dados e critérios de avaliação utilizados.

Esta etapa buscou identificar as principais abordagens utilizadas na literatura para seleção de atributos em IDS, bem como os modelos de aprendizado de máquina mais recorrentes nesse contexto. A análise permitiu alinhar a solução proposta com as práticas consolidadas na área, considerando a presença ou ausência de seleção de atributos no pré-processamento e os critérios que orientam essa decisão.

Com base no estudo da literatura recente em seleção de características, elaborou-se um ranqueamento dos modelos de aprendizado mais empregados, o que subsidiou a escolha do algoritmo *Random Forest* para a fase de validação do modelo.

Tabela 2 – Frequência de uso de modelos de ML na literatura revisada

Modelo de ML	**Frequência em Estudos Revisados**
RF	26
XGBoost	21
SVM	18
LSTM	17
DT	14
CNN	13
MLP	12
Ensemble (Voting/Bagging)	11
Autoencoder	10
LightGBM	10
CatBoost	8
Naive Bayes	9
Isolation Forest	9
KNN	7
One-Class SVM	6
K-means	5
DBSCAN	4

A literatura especializada apresenta uma variedade de abordagens que combinam estatística, aprendizado de máquina e teoria da informação para selecionar subconjuntos relevantes de atributos. Esta subseção realiza uma revisão dos principais métodos utilizados na última década em detecção de ameaças. Foram selecionados artigos publicados entre 2018 e 2025. A pesquisa demonstrou que modelos supervisionados permanecem como a principal escolha em IDS.

A revisão bibliográfica também demonstrou que, embora não haja um modelo universalmente superior, os algoritmos Floresta Aleatória (RF), Impulsionamento de Gradiente Extremo (XGBoost), SVM e LSTM apresentam desempenho robusto quando associados a boas práticas de seleção de atributos. Dentre eles, o Random Forest foi escolhido pela sua elevada frequência de uso, bom desempenho médio e facilidade de interpretação.

4.3 ANÁLISE E COMPARAÇÃO DOS MÉTODOS DE SELEÇÃO EXISTENTES

Nesta etapa, foram implementados e avaliados, em ambiente Python, mais de 40 métodos distintos de seleção de características, distribuídos entre as quatro principais abordagens da literatura: *Filter*, *Wrapper*, *Embedded* e *Hybrid*. Todos os métodos foram aplicados sobre o mesmo conjunto de dados, permitindo uma análise sistemática e equânime.

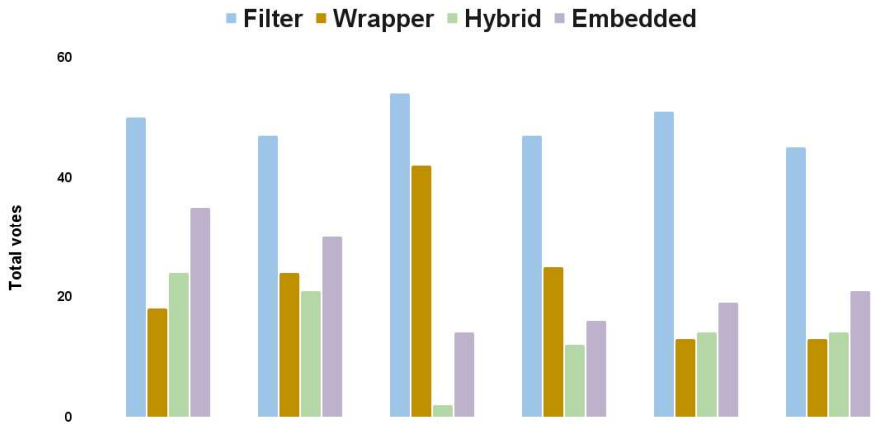
A Tabela 3 apresenta os métodos considerados no estudo. A avaliação baseou-se em critérios como desempenho do classificador após a seleção, interpretabilidade, tempo de processamento e estabilidade das seleções geradas. Para aferir estabilidade, considerou-se a consistência dos subconjuntos selecionados por diferentes métodos pertencentes à mesma abordagem.

O objetivo do estudo foi identificar quais métodos geram os melhores subconjuntos de atributos, mas também compreender como cada abordagem se comporta diante de dados de alta dimensionalidade e com distribuição não-paramétrica.

Tabela 3 – Métodos de Seleção de Features

Método	**Abordagem**	**Classe**
ANOVA	Filter	Statistical
F-test	Filter	Statistical
Chi-square	Filter	Statistical
Gini index	Filter	Statistical
Likelihood Ratio	Filter	Statistical
Canonical Correlation	Filter	Statistical
IG	Filter	Information
Mutual Information	Filter	Information
Variance Information	Filter	Information
Variable Importance	Filter	Information
K-best	Filter	Statistical
Max-relevance and min-redundancy	Filter	Statistical
Fischer Score	Filter	Statistical
Pearson Correlation	Filter	Statistical
Univariate ROC-AUC	Filter	Statistical
Laplacian Score	Filter	Similarity
Least Angle Regression	Filter	Similarity
Conditional Covariance Minimization	Filter	Similarity
Spearman Correlation	Filter	Statistical
Kruskal-Wallis	Filter	Statistical
K-Neighbors Classifier	Wrapper	Sequential Forward Selection
Decision Tree Classifier	Wrapper	Sequential Forward Selection
Random Forest Classifier	Wrapper	Sequential Forward Selection
Bagging Classifier	Wrapper	Sequential Forward Selection
Decision Tree Classifier	Wrapper	Sequential Backward Selection
Random Forest Classifier	Wrapper	Sequential Backward Selection
Relief	Hybrid	Multivariate
Elastic Net	Hybrid	Regularization-based
Lasso Regression (L1)	Hybrid	Regularization-based
Ridge Regression (L2)	Hybrid	Regularization-based
SVM	Hybrid	Regularization-based
Logistic Regression	Hybrid	Regularization-based
Support Vector Classifier	Hybrid	Select from Model
Random Forest Classifier	Hybrid	Select from Model
Boosting Classifier	Hybrid	Select from Model
Logistic Regression	Hybrid	Select from Model
Multiple Linear Regression	Hybrid	Select from Model
Extreme Gradient Boosting	Embedded	Tree-based
AdaBoost	Embedded	Tree-based
Light Gradient Boost	Embedded	Tree-based
Extra Trees Classifier	Embedded	Tree-based
Cat Boost	Embedded	Tree-based
Gradient Boosting Tree	Embedded	Tree-based
Random Forest Classifier	Embedded	Tree-based
Decision Tree Classifier	Embedded	Tree-based

Figura 2 – Estabilidade do conjunto resposta por método e abordagem



A técnica de votação majoritária foi empregada para identificar os atributos mais frequentemente selecionados, tanto por método quanto por abordagem. A Figura 2 ilustra a estabilidade observada em cada abordagem. Os resultados evidenciaram a superioridade dos métodos da abordagem *Filter*, que apresentaram maior consistência entre seleções, menor tempo de execução e maior potencial de interpretação.

Abordagens *Wrapper* e *Embedded* apresentaram alta variabilidade nos subconjuntos gerados e custos computacionais mais elevados, limitando sua aplicabilidade em cenários com dados volumosos e de alta dimensionalidade.

Com base nesses resultados, optou-se por adotar a abordagem *Filter* como base para a solução proposta. Foram selecionadas métricas estatísticas não paramétricas como entropia de Shannon, correlação de Spearman e distância de Bhattacharyya com normalização adaptativa, que demonstraram desempenho superior e estabilidade na seleção de atributos em cenários com distribuição não gaussiana.

4.4 DEFINIÇÃO DOS CONJUNTOS DE DADOS PARA ESTUDOS DE CASO

Foram selecionados três conjuntos de dados representativos de cenários reais de detecção de ameaças. Esses conjuntos apresentam alta dimensionalidade, diversidade de ataques e diferentes graus de desbalanceamento, características essenciais para avaliar a robustez do modelo proposto em condições realistas.

Os conjuntos de dados utilizados foram:

Tabela 4 – Conjuntos de dados selecionados

Dataset	**Features**	**Amostras**	**Tipos de Ataque**	**Razão Ataque/Normal**
CICDDoS-2019	83	12.7M	12	1:850
UNSW-NB15	49	2.5M	9	1:1500
HPC-Lab	67	3.2M	7	1:2300

4.5 DESENVOLVIMENTO DO ALGORITMO DE SELEÇÃO

A construção do modelo proposto foi estruturada em três etapas principais: (i) definição do pré-processamento dos dados, com foco em normalização robusta e tratamento de *outliers*; (ii) seleção ou adaptação de métricas estatísticas de distância e similaridade não paramétricas; e (iii) desenvolvimento do algoritmo de seleção em si, incluindo os critérios de ranqueamento e seleção final.

4.5.1 Definir pré-processamento

O pré-processamento dos dados foi cuidadosamente definido para garantir a compatibilidade com as métricas não paramétricas adotadas, evitando suposições de normalidade e atenuando o impacto de *outliers*. Substituiu-se a média pela mediana como medida de tendência central, e a variância foi substituída pelo intervalo interquartil (IQR), conferindo maior robustez ao modelo. Além disso, foram incluídas técnicas de detecção de *outliers*, engenharia temporal e balanceamento de classes.

O pré-processamento adotado contemplou os seguintes procedimentos:

- **Detecção de *outliers* com *Isolation Forest***: identifica instâncias anômalas em relação à distribuição esperada, sem pressupor normalidade;
- **Normalização robusta**:
$$x' = \frac{x - \tilde{x}}{\text{IQR} + \epsilon}$$
onde $\tilde{x}$ representa a mediana, e ϵ é uma constante positiva pequena para evitar divisão por zero;
- **Engenharia temporal**: cálculo de médias móveis e desvios padrão móveis em janelas de 5 e 30 segundos, para capturar dinamismo local no tempo;
- **Balanceamento das classes**: aplicação do algoritmo SMOTE-ENN, que combina oversampling de minoria (SMOTE) com remoção de ruídos da maioria (ENN), promovendo um conjunto mais equilibrado e informativo.

4.5.2 Desenvolvimento das Métricas de Seleção

Nesta etapa, foram definidas e, quando necessário, adaptadas as principais métricas utilizadas nos algoritmos de seleção de características, agrupadas por suas respectivas funções: relevância, redundância, separabilidade e dependência supervisionada. A seleção dessas métricas priorizou abordagens não paramétricas, robustas a *outliers* e adequadas para distribuições não conhecidas ou multimodais.

- **Métricas de Relevância**: comparadas diferentes medidas informacionais para quantificar a capacidade discriminativa das features em relação à classe. A Tabela 5 resume suas características.

Tabela 5 – Comparação entre medidas de informação

Métrica	**Baseada em**	**Normalidade**	**Resistente a *outliers***
Entropia de Shannon	Distribuição empírica	Não assume	Sim
Ganho de Informação	Entropia condicional	Não assume	Sim
F-score / ANOVA	Diferença de médias	Assume	Não
Chi-quadrado	Frequência categórica	Assume	Parcial

A entropia de Shannon foi escolhida pois não assume normalidade, simetria, homocedasticidade nem linearidade, o que a torna ideal para dados de tráfego de rede, logs de sistema, IoT, cibersegurança, onde a distribuição é irregular e multimodal.

$$H(f_i) = -\sum_j P(f_{i,j}) \log P(f_{i,j})$$

Onde $P(f_{i,j})$ representa a probabilidade da ocorrência do valor j na feature f_i.

A entropia de Shannon foi selecionada por sua capacidade de lidar com distribuições irregulares e assimétricas:

$$H(f_i) = -\sum_j P(f_{i,j}) \log P(f_{i,j})$$

- **Filtragem de Redundância**: realizada com base na correlação de Spearman. A Tabela 6 apresenta a comparação entre correlações usuais.

Tabela 6 – Comparação entre correlações

Métrica	**Relação Detectada**	**Normalidade**	***Ranks***	***outliers***
Pearson	Linear	Sim	Não	Sensível
Spearman	Monotônica	Não	Sim	Robusta

$$\rho_{X,Y} = \frac{\text{cov}(\text{rank}(X), \text{rank}(Y))}{\sigma_{\text{rank}(X)} \cdot \sigma_{\text{rank}(Y)}}$$

- **Separabilidade entre classes**: foi utilizada uma versão modificada da distância de Bhattacharyya, conforme a Tabela 7.

Tabela 7 – Comparação entre métricas de separação de classes

Métrica	**Tipo**	**Robusta a *outliers***	**Normalidade**
Bhattacharyya (modificada)	Não-paramétrica	Sim	Não assume
Fisher Discriminant	Linear	Não	Assume
Euclidean Distance	Métrica pura	Não	Assume

$$D_B = \frac{1}{4}\sum_{i=1}^{n}\frac{(\tilde{x}_0^i - \tilde{x}_1^i)^2}{\mathsf{IQR}_0^2 + \mathsf{IQR}_1^2} + \frac{1}{2}\log\left(\frac{\mathsf{IQR}_0^2 + \mathsf{IQR}_1^2}{2\sqrt{\mathsf{IQR}_0^2 \cdot \mathsf{IQR}_1^2}}\right)$$

- **Dependência supervisionada**: foi adotada a informação mútua ajustada (AMI), precedida por pré-processamento robusto. A Tabela 8 apresenta a comparação entre métricas.

Tabela 8 – Comparação entre medidas de dependência supervisionada

Métrica	**Corrige Viés**	**Desbalanceamento**	**Baseada em Informação**
AMI	Sim	Sim	Sim
Mutual Information	Não	Suporta parcialmente	Sim
Chi-quadrado	Não	Suporta parcialmente	Não
Gain Ratio	Sim	Suporta parcialmente	Sim

$$AMI(X,Y) = \frac{MI(X,Y) - \mathbb{E}[MI(X,Y)]}{\max(H(X),H(Y)) - \mathbb{E}[MI(X,Y)]}$$

A Tabela 9 detalha as etapas de robustez aplicadas antes do cálculo da AMI.

Tabela 9 – Etapas anteriores ao cálculo da AMI e respectivas técnicas de robustez

Etapa Anterior	**Técnica de Robustez Aplicada**
Remoção de *outliers*	*Isolation Forest*
Normalização	Mediana + IQR (quantis)
Filtragem de redundância	Correlação de Spearman
Similaridade de grupos	Mahalanobis com quantis

4.6 ANÁLISE DOS RESULTADOS

A validação do modelo proposto foi conduzida por meio de protocolos estatísticos, com o objetivo de assegurar robustez, generalização e significância dos resultados obtidos. Foram empregadas técnicas de validação cruzada, aplicação em conjuntos externos e testes estatísticos não paramétricos. Para analisar o desempenho do modelo, foram considerados dois eixos de avaliação: (i) desempenho preditivo de classificação com Random Forest, utilizando métricas como Exatidão, Precisão, F1-Score e AUC-ROC; e (ii) comparação estatística com outros métodos de seleção por meio do teste de Friedman seguido do pós-teste de Nemenyi.

As expressões matemáticas do teste de Friedman e do Cálculo da Diferença Crítica (CD) são apresentadas a seguir, conforme detalhado na literatura (Demšar, 2006b).

A estatística do teste de Friedman é dada por:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right]$$

onde:

- N: número de datasets ou folds utilizados;
- k: número de métodos de seleção comparados;
- R_j: soma das posições (ranks) atribuídas ao método j em cada repetição.

Se o valor de χ_F^2 for significativo, rejeita-se a hipótese nula de que todos os métodos têm desempenho equivalente.

Como pós-teste, utilizou-se o teste de Nemenyi, que identifica quais pares de métodos apresentam diferenças estatisticamente significativas. A comparação entre dois métodos i e j é considerada significativa se a diferença média de *ranks* $|R_i - R_j|$ for maior do que o valor crítico conhecido como CD, dado por:

$$CD = q_\alpha \cdot \sqrt{\frac{k(k+1)}{6N}}$$

onde q_α é o valor crítico da distribuição de Studentized range para o nível de significância α, k o número de métodos, e N o número de datasets.

- **Validação cruzada estratificada 10-fold**: garante que a proporção entre classes seja mantida em cada partição, sendo especialmente importante em contextos com desbalanceamento severo;
- **Holdout externo**: utilização de conjuntos de dados não empregados no treinamento, com o objetivo de verificar a capacidade de generalização do modelo;
- **Teste de Friedman**: teste estatístico não paramétrico utilizado para avaliar se existem diferenças significativas entre os métodos de seleção comparados. A estatística do teste é dada por:

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \tag{4.1}$$

  onde n é o número de conjuntos de dados (ou execuções), k é o número de métodos avaliados e R_j é a soma dos *ranks* atribuídos ao j-ésimo método;
- **Pós-teste de Nemenyi**: utilizado para identificar quais pares de métodos diferem significativamente entre si, com base em um limiar de significância $p < 0{,}05$.

Após a validação do modelo por meio de validação cruzada estratificada e testes estatísticos não paramétricos, os resultados são analisados sob múltiplas perspectivas. O foco central é examinar a eficácia do algoritmo de seleção de características proposto na melhoria do desempenho de classificadores, especialmente em contextos com dados de alta dimensionalidade, desbalanceamento severo entre classes e ausência de pressupostos paramétricos.

A análise comparativa considera o desempenho dos modelos de aprendizado de máquina com e sem a aplicação do método de seleção. Foram utilizadas métricas tradicionais como exatidão, precisão, F1 e AUC-ROC, avaliando tanto a capacidade preditiva quanto a robustez frente a falsos positivos e negativos. Essas métricas foram computadas em diferentes conjuntos de dados de segurança cibernética amplamente utilizados na literatura, garantindo diversidade nos cenários avaliados.

Além das métricas preditivas, o impacto da redução de dimensionalidade foi mensurado em termos de:

- **Taxa de Redução de Atributos**: proporção de features eliminadas pelo método em relação ao conjunto original;
- **Tempo de Treinamento e Inferência**: avaliação do tempo computacional necessário antes e após a seleção, medindo ganhos em escalabilidade;
- **Estabilidade das Seleções**: verificação da consistência dos subconjuntos selecionados entre execuções, considerando diferentes partições dos dados e diferentes bases;
- **Impacto na Interpretabilidade**: análise qualitativa da compreensibilidade dos subconjuntos gerados, com base em critérios de transparência estatística e semântica dos atributos selecionados.

A comparação entre os resultados obtidos com o modelo proposto e os gerados por métodos consagrados de seleção de características foi realizada com o **teste de Friedman**, apropriado para múltiplas comparações em dados não normalmente distribuídos, seguido do **pós-teste de Nemenyi**, o qual determina a significância estatística entre pares de métodos com base na diferença média de *ranks*. Foi adotado o nível de significância de $p < 0{,}05$.

Os resultados obtidos evidenciaram que o modelo proposto proporcionou, em todos os conjuntos de dados avaliados, redução expressiva na quantidade de atributos, sem prejuízo — e, em muitos casos, com ganho — nas métricas de desempenho. Tais ganhos foram ainda acompanhados por menores tempos de processamento e maior estabilidade entre execuções, fatores fundamentais para aplicações em ambientes críticos e em tempo real, como IDS.

4.7 CONSIDERAÇÕES

O método proposto consolidou uma estratégia não paramétrica eficaz para seleção de atributos em ambientes caracterizados por tráfego de rede de alta dimensionalidade, distribuição assimétrica e elevado desbalanceamento entre classes. Ao adotar métricas estatísticas robustas, livres de pressupostos de normalidade, e técnicas de filtragem, clusterização e ranqueamento integradas de forma sistemática, a abordagem se mostrou capaz de preservar (e,

frequentemente, melhorar) o desempenho preditivo dos classificadores, ao mesmo tempo em que promove significativa economia computacional.

Além disso, o modelo demonstrou elevada estabilidade entre execuções, com subconjuntos de atributos consistentes e interpretáveis, o que favorece sua aplicação em contextos críticos como cibersegurança, IoT e análise de logs. O uso da validação estatística assegura confiabilidade às conclusões obtidas e fortalece o potencial de generalização da abordagem.

5 ALGORITMO PROPOSTO DE SELEÇÃO NÃO PARAMÉTRICA

A fase de *Algorithm Design* compreendeu o delineamento da arquitetura do algoritmo proposto, com a definição de um conjunto de métricas estatísticas não paramétricas e sua integração em um fluxo coerente, robusto e escalável. O objetivo central foi garantir confiabilidade na seleção de características, mesmo em cenários com alto grau de assimetria, desbalanceamento entre classes e presença de *outliers*, como ocorre frequentemente em dados de tráfego de rede utilizados na detecção de ataques cibernéticos.

A partir do estudo sistemático de algoritmos de filtragem clássicos, foram identificadas etapas recorrentes entre os métodos: cálculo de medidas de distância, análise de dissimilaridade, agrupamento, avaliação de importância no grupo e ranqueamento final. Métricas amplamente utilizadas como média, covariância, distância Euclidiana e correlação de Pearson, embora efetivas em contextos com dados gaussianos, não são ideais para os domínios de interesse deste trabalho.

Dessa forma, para tornar o algoritmo mais robusto, optou-se por substituir essas métricas por alternativas estatisticamente mais adequadas a dados não paramétricos:

- **Mediana no lugar da média**: por ser menos sensível a valores extremos, a mediana oferece maior robustez, preservando a representatividade de padrões relevantes de ataque, mesmo em distribuições assimétricas, como as de tipo Pareto, comuns em logs de rede e métricas de sistema;
- **Variância acumulada em vez de covariância**: utilizada como métrica interna de contribuição relativa da característica para a estrutura do grupo, permitindo avaliar sua importância sem pressupor distribuição normal;
- **Correlação de Spearman em vez de Pearson**: a correlação de Spearman, por operar sobre postos ordenados, é capaz de capturar associações monotônicas entre atributos, independentemente da escala, e é naturalmente resistente a *outliers* e transformações não lineares. Essa substituição é fundamental para evitar sobreajuste e redundância indesejada;
- **Distância de Mahalanobis modificada em vez da Euclidiana**: a métrica tradicional de Mahalanobis foi adaptada para utilizar estimativas robustas (medianas e Intervalo interquartílico (IQR)), sendo capaz de considerar a interdependência entre variáveis sem exigir normalidade multivariada, o que se mostrou crucial para a clusterização confiável de atributos;
- **Distância de Bhattacharyya adaptada**: aplicada como critério de separabilidade entre classes, utilizando quantis ao invés de médias e desvios padrão, garantindo maior fidelidade à estrutura empírica dos dados;

- **Teste de Friedman**: empregado como etapa final de validação da significância estatística dos conjuntos de atributos selecionados, garantindo que a melhoria no desempenho não ocorra por acaso.

A complexidade dos conjuntos de dados avaliados — marcados por alta dimensionalidade, grande volume de amostras, desbalanceamento entre instâncias normais e maliciosas e ausência de distribuição conhecida, impõe limitações às técnicas clássicas de seleção de atributos, motivando a formulação de um modelo abrangente e resiliente. Em resposta a esse desafio, este capítulo apresenta um modelo formal de *Feature Selection*, projetado para operar com dados de natureza não paramétrica, suportar ambientes em tempo real e minimizar custo computacional.

O algoritmo foi implementado em Python e validado sobre os conjuntos de dados discutidos no Capítulo 4.2. Sua estrutura modular permite integração em IDS e aplicações de cibersegurança em larga escala. A próxima seção apresenta os algoritmos desenvolvidos, detalhando suas etapas e procedimentos computacionais.

5.1 ALGORITMO PROPOSTO DE SELEÇÃO NÃO-PARAMÉTRICA

A partir das etapas previamente definidas e das métricas selecionadas por sua robustez estatística, foi concebido um algoritmo de seleção de características com foco em dados não paramétricos, desbalanceados e de alta dimensionalidade. O algoritmo foi projetado para combinar múltiplos critérios estatísticos em um pipeline organizado, com ênfase em interpretabilidade e viabilidade computacional para aplicações em tempo real, como IDS.

A Figura 1 apresenta o pseudocódigo do algoritmo proposto. O processo inicia-se com o pré-processamento dos dados, utilizando técnicas robustas à presença de *outliers* e distribuições assimétricas. Em seguida, uma filtragem inicial por entropia é realizada para eliminar atributos com baixa variabilidade informacional. A etapa de clusterização emprega uma versão modificada da distância de Mahalanobis, que utiliza mediana e intervalo interquartil, promovendo agrupamento de atributos similares. Por fim, dentro de cada grupo, é selecionado o atributo mais informativo, com base na maior AMI com a variável-alvo.

A estratégia empregada pode ser interpretada como uma hibridização entre métodos *Filter* e técnicas de análise de redundância, similar ao modelo mRMR, porém com ênfase em medidas robustas e não paramétricas. A estrutura do algoritmo permite integração modular com qualquer classificador supervisionado posterior, mantendo separação clara entre seleção de atributos e modelagem preditiva.

Algorithm 1 Algoritmo Proposto de Seleção de Características Não-Paramétrico

Conjunto de dados X com n amostras e d atributos; vetor de rótulos y Subconjunto S de atributos selecionados **Pré-processamento:**

1. Aplicar `Isolation Forest` para remoção de outliers;
2. Normalizar atributos usando mediana e IQR:

$$x' = \frac{x - \tilde{x}}{\text{IQR} + \varepsilon}$$

Filtragem Inicial:

3. Calcular entropia de cada atributo $H(f_i)$;
4. Eliminar atributos com $H(f_i) < 0{,}4$;

Agrupamento:

5. Calcular distância de Mahalanobis robusta entre atributos;
6. Aplicar agrupamento hierárquico;

Seleção Final:

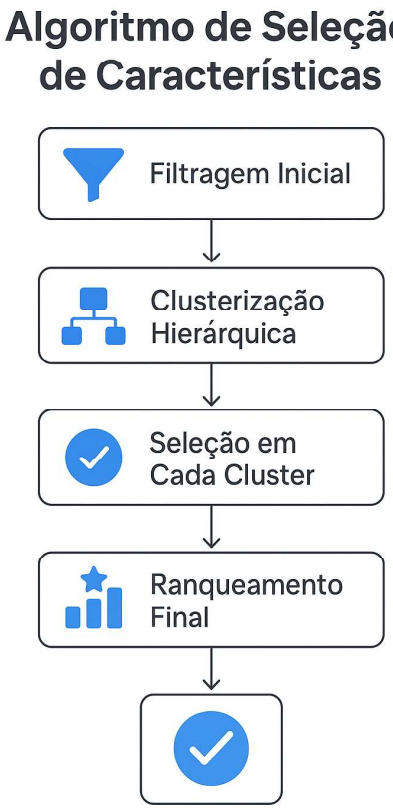
7. Escolher a melhor feature de cada grupo com maior AMI;
8. Retornar conjunto S final de atributos.

Essa formulação combina etapas de análise exploratória, medidas estatísticas robustas e dependência supervisionada, promovendo uma seleção de atributos sensível à estrutura dos dados reais. A estratégia se mostrou eficaz na redução de dimensionalidade com manutenção — ou mesmo melhora — do desempenho preditivo, conforme discutido na Seção 4.6. Além disso, por não depender de modelos preditivos internos, o algoritmo possui baixo custo computacional em comparação a métodos do tipo *Wrapper* ou *Embedded*.

Referência de base: A ideia de combinar relevância e redundância para seleção de características é inspirada em trabalhos como mRMR, enquanto a robustez estatística se baseia em técnicas consolidadas na literatura de estatística robusta e cibersegurança (Peng, Long e Ding; Deng et al., 2005, 2019).

A Figura 3 ilustra, de forma conceitual e sequencial, as principais etapas do algoritmo de seleção de características não paramétrico.

Figura 3 – Algoritmo de Seleção de Características Não Paramétrico Proposto



DISTÂNCIA DE MEDIDA DE SIMILARIDADE

A DMS, no contexto desta dissertação, não corresponde a uma distância euclidiana convencional, mas sim a uma métrica de **dissimilaridade** concebida para quantificar a *perturbação na representação de uma instância de dados quando uma característica específica é removida*. Em essência, a DMS busca responder: *Qual a magnitude da alteração no perfil de uma observação ao desconsiderarmos um de seus atributos?*

Conceitualmente, a DMS opera sob a premissa de que uma característica é intrinsecamente relevante e não redundante se sua remoção acarreta uma modificação significativa na *posição* da instância no espaço de características. Por outro lado, se a ausência de um atributo resulta em alteração mínima na identidade da instância, esse atributo pode ser considerado de menor informatividade ou altamente redundante.

Componentes e Sua Racionalidade

A formulação da DMS empregada neste trabalho é uma adaptação estratégica de conceitos estatísticos, visando sua aplicabilidade no ambiente desafiador dos dados de ciber-

segurança:

$$\mathrm{DMS}(x) = (x_{-i} - \mathrm{med}_{-i})^T R_{-i}^{-1} (x_{-i} - \mathrm{med}_{-i})$$

A seguir, descrevem-se os componentes da métrica e a lógica por trás de cada um deles.

$(x_{-i} - \mathrm{med}_{-i})$: Vetor de Desvio Robusto

- x_{-i}: Vetor da instância x com a remoção da i-ésima feature.
- med_{-i}: Vetor das *medianas* das demais features no conjunto. O uso da mediana, em vez da média, fundamenta-se na robustez contra *outliers*, comum em dados de tráfego de rede e cibersegurança. Isso torna a representação do centro dos dados mais estável, especialmente em distribuições assimétricas.

R_{-i}^{-1}: Matriz de Correlação Inversa

- R_{-i}: Matriz de correlação de Spearman entre as features remanescentes. A correlação de Spearman, por capturar relações monotônicas e não depender de suposições de normalidade, é apropriada para o contexto não paramétrico deste trabalho.
- R_{-i}^{-1}: A inversa da matriz de correlação atua como um fator de ponderação, atenuando o impacto de atributos redundantes. Assim, a DMS valoriza a *informação não redundante*, considerando a estrutura de dependência entre os atributos.

R_{-i}^{-1}: A Matriz de Correlação Inversa e o Tratamento da Redundância

- R_{-i}: É a matriz de **correlação** entre as características *remanescentes* (após a remoção de f_i). No presente trabalho, a **Correlação de Spearman** é utilizada para construir esta matriz, reforçando, novamente, a natureza não-paramétrica da abordagem. A correlação de Spearman avalia relações monotônicas e se mostra robusta a não-linearidades e a *outliers*, sendo ideal para dados onde as relações não são estritamente lineares.
- R_{-i}^{-1}: A inclusão da inversa da matriz de correlação é o fator que distingue a DMS de uma simples distância euclidiana. Ela atua como uma **matriz de ponderação que leva em consideração a interdependência e a redundância entre as características restantes**. Caso duas características sejam altamente correlacionadas (redundantes), esta matriz "desconta" a contribuição dessa redundância para a distância. Isso permite que a DMS valorize a **informação não-redundante**, medindo o desvio da instância em

um espaço onde as correlações entre as características foram "removidas" ou "normalizadas", similar ao conceito da Distância de Mahalanobis.

Forma Quadrática

A estrutura da fórmula segue o modelo de uma forma quadrática:

$$(x_{-i} - \text{med}_{-i})^T R_{-i}^{-1} (x_{-i} - \text{med}_{-i})$$

Esse formato é típico de métricas como a distância de Mahalanobis, permitindo medir o desvio ponderado de uma instância em relação a um centro, ajustado pela dependência entre atributos.

DMS como Critério de Seleção de Características

Valores elevados de DMS associados à remoção de uma feature f_i indicam que:

- f_i contém **informação única** que não pode ser compensada pelas demais;
- f_i é **altamente relevante** para a descrição da instância x.

Na metodologia proposta, a DMS é um dos pilares do processo de seleção de atributos, sendo combinada a métricas como entropia e AMI. Essa combinação permite selecionar atributos informativos e não redundantes mesmo em dados altamente desbalanceados, não gaussianos e com presença de ruídos.

Síntese

A DMS é uma métrica **não paramétrica e multidimensional** que quantifica o impacto da remoção de um atributo sobre a identidade estatística de uma instância. Sua força reside na capacidade de incorporar robustez estatística, sensibilidade à relevância e penalização de redundância — propriedades essenciais em contextos de cibersegurança com dados ruidosos, assimétricos e de alta dimensionalidade.

5.2 ESTRUTURA DO ALGORITMO

O modelo proposto consiste em um algoritmo trifásico, projetado para atuar de forma eficaz em contextos com dados não paramétricos, alta dimensionalidade e presença de ruído

estatístico. Sua arquitetura foi desenhada para combinar robustez, escalabilidade e interpretabilidade, características essenciais em ambientes de cibersegurança e análise de tráfego de rede. As três fases principais são descritas a seguir:

1. **Filtragem Inicial:** Nesta fase, cada característica é avaliada de forma univariada quanto à sua variabilidade e informatividade, utilizando métricas não paramétricas. A filtragem serve como etapa de redução inicial do espaço de busca, eliminando atributos com baixa entropia ou relevância estatística. Essa abordagem permite descartar atributos que não contribuem significativamente para a variabilidade dos dados, tornando o processo subsequente mais eficiente.

2. **Clusterização Hierárquica:** As características remanescentes são agrupadas com base em métricas de similaridade robustas, como a distância de Mahalanobis adaptada para distribuição não-normal. O objetivo dessa etapa é identificar grupos de atributos redundantes ou altamente correlacionados, evitando a seleção simultânea de múltiplas variáveis com comportamento semelhante, o que poderia induzir sobreajuste.

3. **Seleção Final por Informação Mútua:** Após a formação dos clusters, o algoritmo seleciona, em cada grupo, a característica com maior relevância em relação à variável-alvo, medida por AMI. Essa escolha é posteriormente refinada com base na correlação de Spearman, assegurando que o conjunto final de atributos seja ao mesmo tempo informativo, não redundante e robusto frente a dados desbalanceados.

Esse encadeamento sistemático não apenas reduz a dimensionalidade do conjunto de dados, mas também favorece a interpretabilidade do modelo final. O processo é particularmente vantajoso em contextos de monitoramento de segurança, onde decisões precisas e auditáveis são fundamentais.

5.3 COMPONENTES E ADAPTAÇÕES ESTATÍSTICAS PARA DADOS NÃO PARAMÉTRICOS

Considerando as características dos dados utilizados, como assimetria, alta variância e a presença de *outliers*, optou-se por um conjunto de métricas estatísticas não paramétricas, que dispensam suposições de normalidade e garantem maior robustez nas etapas de seleção de atributos.

Métricas estatísticas utilizadas

Como alicerce do modelo, adotaram-se medidas robustas de tendência central e dispersão:

- **Mediana ($\tilde{x}$)**: substitui a média aritmética como medida de tendência central. Por ser baseada na posição dos dados e não nos seus valores absolutos, a mediana é menos sensível a valores extremos e garante maior estabilidade estatística na presença de *outliers*.
- **Intervalo Interquartílico**: definido como $Q_3 - Q_1$, o IQR representa a amplitude do intervalo central dos dados e é utilizado como substituto da variância. Essa escolha garante uma estimativa de dispersão mais robusta, ideal para cenários onde a distribuição dos dados é multimodal ou assimétrica.

Essas métricas formam a base das transformações estatísticas utilizadas em todas as fases do algoritmo, desde a normalização até os cálculos de similaridade e separabilidade.

Relevância da Característica (Filtragem Inicial)

Na primeira fase, a relevância de cada característica é quantificada por meio da **Entropia de Shannon**, uma medida derivada da teoria da informação que avalia a imprevisibilidade de uma variável:

$$H(f_i) = -\sum_j P(f_{i,j}) \log P(f_{i,j})$$

em que $P(f_{i,j})$ representa a frequência relativa do valor j na variável f_i. A entropia é calculada com base na distribuição empírica dos dados, e não exige pressupostos sobre sua forma.

Características com $H(f_i) < 0{,}4$ são descartadas, por apresentarem baixa diversidade informacional. Adicionalmente, utiliza-se o IG como métrica complementar de relevância supervisionada, mensurando o quanto uma característica contribui para a redução da incerteza da variável-alvo:

$$IG(T, f_i) = H(T) - H(T|f_i),$$

onde $H(T)$ é a entropia da variável de saída e $H(T|f_i)$ é a entropia condicional dada a característica f_i.

A aplicação conjunta dessas métricas garante que apenas características informativas e relevantes sejam mantidas no conjunto de dados para as etapas posteriores do algoritmo.

Filtragem de Redundância e Agrupamento

A etapa de mitigação de redundância entre características é essencial para evitar que variáveis altamente correlacionadas causem sobreajuste, degradem a interpretabilidade ou introduzam viés no modelo de aprendizado. Para tal, emprega-se a **correlação de Spearman** ($\rho_{X,Y}$), uma métrica não paramétrica que avalia a associação monotônica entre duas variáveis a partir de seus ranques ordenados. Sua fórmula é dada por:

$$\rho_{X,Y} = \frac{\text{cov}(\text{rank}(X), \text{rank}(Y))}{\sigma_{\text{rank}(X)} \cdot \sigma_{\text{rank}(Y)}}$$

Diferentemente da correlação de Pearson, a medida de Spearman não pressupõe linearidade nem distribuição normal dos dados, sendo, portanto, mais adequada para contextos onde predominam relações não lineares, assimetria e *outliers* — características comuns em dados de tráfego de rede e registros de atividades anômalas.

Após a filtragem por correlação, as características remanescentes são submetidas a uma **clusterização hierárquica**, com base em uma **distância de Mahalanobis modificada**, adaptada para lidar com distribuições não paramétricas. Essa versão robusta substitui a matriz de covariância clássica por uma matriz de correlação de Spearman, e a média vetorial por medianas, de modo a preservar a resistência do modelo a distorções.

A fórmula da distância adaptada entre duas características i e j é expressa por:

$$D_{ij} = \sqrt{(\tilde{x}_i - \tilde{x}_j)^T R^{-1} (\tilde{x}_i - \tilde{x}_j)},$$

em que:

- $\tilde{x}_i$ e $\tilde{x}_j$ representam os vetores de medianas das variáveis i e j, respectivamente;
- R^{-1} é a inversa da matriz de correlação de Spearman entre todas as variáveis selecionadas após a filtragem inicial.

Esse procedimento permite agrupar características com comportamento semelhante em termos de distribuição e associação com os dados, mesmo na ausência de linearidade ou normalidade. Ao final da clusterização, cada grupo resultante representa um conjunto de características potencialmente redundantes. Na fase posterior do algoritmo, será selecionado apenas um representante por cluster, conforme critérios supervisionados de informação mútua. Tal abordagem garante que o conjunto final de atributos seja mais enxuto, não redundante e estatisticamente robusto.

Separação de Classes e Dependência Supervisionada

A capacidade discriminatória das características é avaliada por meio da **distância de Bhattacharyya adaptada**, métrica essencial para quantificar a separação estatística entre classes. Embora a fase final do algoritmo utilize a AMI como critério de ranqueamento supervisionado, a distância de Bhattacharyya atua como etapa complementar na avaliação da separabilidade dos atributos, permitindo identificar, de forma robusta, as características com maior poder de distinção entre as distribuições de tráfego normal e tráfego malicioso.

Motivação para adaptação da métrica

A formulação clássica da distância de Bhattacharyya assume distribuições Gaussianas e depende diretamente de estatísticas paramétricas como a média (μ) e a matriz de covariância (Σ). No entanto, em contextos de cibersegurança — como tráfego de rede e dados de ataques — os dados tendem a ser assimétricos, multimodais, altamente desbalanceados e sujeitos à presença de *outliers*, violando os pressupostos dessas estatísticas. A aplicação direta da versão paramétrica da D_B pode, portanto, levar a inferências distorcidas sobre a separabilidade real entre as classes.

Diante disso, propõe-se uma **adaptação não paramétrica** da distância de Bhattacharyya, que substitui os elementos sensíveis a ruído por estatísticas robustas, garantindo consistência e validade mesmo em cenários com distribuições irregulares.

Formulação da distância adaptada

A nova forma da distância de Bhattacharyya, considerando as substituições mencionadas, é expressa como:

$$D_B = \frac{1}{4}\sum_{i=1}^{n}\frac{(\tilde{x}_0^i - \tilde{x}_1^i)^2}{\mathsf{IQR}_0^2 + \mathsf{IQR}_1^2} + \frac{1}{2}\log\left(\frac{\mathsf{IQR}_0^2 + \mathsf{IQR}_1^2}{2\sqrt{\mathsf{IQR}_0^2 \cdot \mathsf{IQR}_1^2}}\right), \tag{5.1}$$

em que:

- $\tilde{x}_k^i$ representa a mediana da característica i na classe k ($k \in \{0, 1\}$);
- IQR_k é o intervalo interquartil da mesma característica dentro da classe k.

Importância na arquitetura proposta

Essa métrica atua como elemento adicional de robustez na etapa de avaliação interna das características, fornecendo evidências de separação estatística antes do ranqueamento supervisionado final via AMI. Em conjunto, ambas as medidas (separação não supervisionada e dependência supervisionada) fortalecem a seleção de características relevantes e não redundantes, garantindo que o conjunto final selecionado maximize tanto a capacidade discriminativa quanto a generalização do modelo.

A Formulação Adaptada

Neste estudo, a adaptação da distância de Bhattacharyya para lidar com dados não paramétricos é explicitada pela fórmula:

$$D_B = \frac{1}{4}\sum_{i=1}^{n}\frac{(\tilde{x}_{i0}-\tilde{x}_{i1})^2}{IQR_{i0}^2+IQR_{i1}^2} + \frac{1}{2}\log\sqrt{\frac{IQR_0^2+IQR_1^2}{2IQR_0^2\cdot IQR_1^2}}$$

Nesta formulação adaptada, o primeiro termo quantifica a "distância"entre os centros robustos (medianas $\tilde{x}$) de cada característica i nas duas classes (0 e 1, e.g., normal e ataque), normalizada pela soma de seus respectivos quadrados do IQR. O segundo termo atua como uma medida de similaridade ou divergência entre as dispersões robustas das duas distribuições, baseadas nos IQR gerais. Essa construção permite que a distância de Bhattacharyya avalie a separabilidade das classes de maneira muito mais fidedigna em ambientes onde a normalidade dos dados não pode ser presumida.

O benefício principal dessa adaptação é a capacidade de identificar características que demonstram uma clara distinção entre diferentes classes (e.g., normal e ataque) com base em suas distribuições reais, mesmo que essas distribuições sejam complexas e não sigam padrões conhecidos. Uma maior distância de Bhattacharyya adaptada indica que uma característica é mais discriminativa, tornando-a valiosa para a seleção de atributos em problemas críticos como a detecção de ameaças cibernéticas, onde a precisão e a robustez são primordiais.

Esta adaptação substitui médias e variâncias por medianas e IQR, tornando a métrica robusta para comparar distribuições não-Gaussianas:

$$D_B = \frac{1}{4}\sum_{i=1}^{n}\frac{(\tilde{x}_{i0}-\tilde{x}_{i1})^2}{IQR_{i0}^2+IQR_{i1}^2} + \frac{1}{2}\log\sqrt{\frac{IQR_0^2+IQR_1^2}{2IQR_0^2\cdot IQR_1^2}}$$

O termo $\tilde{x}_{ik}$ representa a mediana da característica i na classe k, e IQR_{ik} seu intervalo interquartil. Esta métrica informa o grau de separabilidade entre as classes para cada característica.

Para a seleção final e ranqueamento, utiliza-se a AMI. Embora a AMI seja uma métrica baseada em informação que quantifica a dependência entre variáveis, sua aplicação no modelo é precedida por rigorosas etapas de pré-processamento. A detecção de *outliers* com *Isolation Forest*, a normalização robusta baseada em mediana e IQR, e a discretização por *binning* garantem que a AMI seja calculada sobre dados devidamente tratados, minimizando a influência de valores extremos. A AMI é calculada como:

$$\text{AMI}(X,Y) = \frac{\text{MI}(X,Y) - \text{E}[\text{MI}(X,Y)]}{\max(H(X),H(Y)) - \text{E}[\text{MI}(X,Y)]}$$

onde MI é a Informação Mútua e E[MI] é sua expectativa sob independência.

5.4 ALGORITMO PROPOSTO

A combinação estratégica de métodos estatísticos não paramétricos oferece um arcabouço sólido e confiável para a seleção de características em cenários de detecção de anomalias. Essa abordagem contorna limitações de técnicas paramétricas tradicionais, que dependem de suposições como normalidade e homogeneidade de variâncias, condições raramente atendidas

em dados reais de tráfego de rede. Ao dispensar tais pressupostos, as técnicas não paramétricas mantêm alta robustez frente a distribuições irregulares, assimétricas e com presença de *outliers*, garantindo desempenho consistente mesmo em contextos de elevada variabilidade e desbalanceamento.

O algoritmo proposto é **explicável** e **adaptativo**, fundamentado na DMS e incorporando elementos-chave como a correlação de Spearman, a mediana e a entropia para identificar atributos essenciais na predição da variável-alvo. O resultado é um subconjunto otimizado de características que preserva a capacidade discriminativa dos dados, ao mesmo tempo em que reduz a dimensionalidade, aprimorando a eficiência computacional e a exatidão ou acurácia de modelos de aprendizado de máquina.

Em comparação a medidas robustas de dissimilaridade, como a distância de Mahalanobis, a presente abordagem substitui a matriz de covariância pela **matriz de correlação de Spearman**, mitigando erros de agrupamento decorrentes de diferenças de escala entre variáveis. Além disso, substitui-se a média pela **mediana**, aumentando a resistência a valores extremos. Complementarmente, são conduzidas análises individuais de entropia condicional em relação à variável de resposta, permitindo eliminar características pouco informativas sem necessidade de construir matrizes de afinidade, como exigido em métodos do tipo Agrupamento Baseado em Densidade com Ruído (DBSCAN).

A etapa de validação do algoritmo combina múltiplos critérios: o teste estatístico de Friedman para verificar diferenças significativas entre métodos, a distância de Bhattacharyya adaptada para avaliar separabilidade de classes e a análise da distribuição de Pareto para ponderar a contribuição relativa das características selecionadas. Essa integração resulta em um processo de seleção **transparente**, **robusto** e **estatisticamente fundamentado**, projetado para lidar com as propriedades intrínsecas de dados não paramétricos, multivariados e instáveis típicos de ambientes de cibersegurança.

A escolha da **mediana** como medida de tendência central, em substituição à média aritmética, é um elemento chave da estratégia proposta, pois confere robustez à presença de *outliers* típicos de dados de tráfego de rede — como picos abruptos decorrentes de ataques DDoS ou comportamentos anômalos ocasionais. Essa decisão assegura que o cálculo de distâncias e agrupamentos não seja distorcido por valores extremos, mantendo a representatividade estatística das distribuições reais observadas.

A **entropia** das variáveis é integrada ao processo de seleção para captar a sensibilidade das métricas tanto em relação ao tráfego normal quanto à variável de resposta, permitindo identificar atributos com maior poder de discriminação entre classes. No estágio final, o IG quantifica a redução de incerteza na variável-alvo quando uma característica específica é conhecida, garantindo que o subconjunto final preserve apenas atributos de alta relevância preditiva. Essa combinação — mediana, entropia e IG — assegura que o conjunto resultante seja simultaneamente enxuto, informativo e estável.

A avaliação é complementada por métricas e testes estatísticos de alto rigor. O **teste**

de Friedman permite verificar a significância das diferenças de desempenho entre os métodos de seleção, enquanto a **distância de Bhattacharyya** avalia a separabilidade entre classes sem assumir gaussianidade, utilizando Estimativa de Densidade por Kernel (KDE) para maior fidelidade a dados não paramétricos.

Quando comparado a algoritmos de agrupamento como DBSCAN e Ordenação de Pontos para Identificar a Estrutura de Agrupamento (OPTICS), o método proposto apresenta vantagens substanciais: elimina a necessidade de construir matrizes de afinidade e de ajustar parâmetros sensíveis (ϵ, $minPts$), além de evitar procedimentos recursivos onerosos. Essa simplicidade operacional, aliada à robustez estatística, reduz significativamente o custo computacional e aumenta a escalabilidade da solução para ambientes com dados de alta dimensionalidade, distribuições complexas e forte desbalanceamento.

Essa combinação de métodos é particularmente valiosa na seleção de características de dados não paramétricos, pois fornece uma abordagem abrangente, explicável e rigorosa. Ao considerar diferenças entre grupos, obtemos compreensão mais profunda das características que realmente distinguem as condições analisadas, sem ignorar variações significativas e mantendo base estatística robusta.

Além disso, a natureza não paramétrica desses métodos os torna resilientes a violações de hipóteses, permitindo lidar com ampla variedade de tipos e distribuições de dados sem comprometer a precisão dos resultados.

O Algoritmo 2 detalha o processo de seleção de features realizado. O processo começa com o cálculo da matriz de correlação de Spearman, que mede relações não lineares entre as features. Essa matriz é essencial para avaliar redundâncias e será utilizada posteriormente no cálculo de dissimilaridades. Em seguida, o algoritmo realiza um agrupamento de features baseado na distância de Mahalanobis modificada, uma medida robusta que considera a estrutura de correlação dos dados. Para cada feature, um subconjunto temporário é criado removendo-a, e então calcula-se a mediana desse subconjunto. Para cada instância, a diferença em relação à mediana é computada, e a distância de Mahalanobis modificada é derivada usando a inversa da matriz de correlação. Esse valor é armazenado e associado à feature removida.

Algorithm 2

Require: X: conjunto de dados original com todas as *features*
Ensure: X_{final}: conjunto otimizado sem *features* redundantes
0: Calcular matriz de correlação de Spearman: R
0: Agrupar *features* com base na distância de Mahalanobis
0: **for** cada *feature* f_i em X **do**
0: Criar subconjunto X_{-i} removendo f_i
0: Calcular vetor mediana $\tilde{x}_{-i}$ de X_{-i}
0: Calcular entropia $H(f_i)$
0: **for** cada instância x em X_{-i} **do**
0: Calcular vetor diferença: $d = x_{-i} - \tilde{x}_{-i}$
0: Calcular distância: $DMS(x) = d^{\top} R_{-i}^{-1} d$
0: **end for**
0: Armazenar DMS_j associada à remoção de f_i
0: **end for**
0: Identificar *features* com alta mediana de DMS e alta entropia H
0: **for** cada *feature* f_i removida **do**
0: Calcular distância mediana $\mathrm{median}(DMS_i)$
0: Marcar f_i como relevante se $\mathrm{median}(DMS_i)$ e $H(f_i)$ forem altos
0: **end for**
0: Remover *features* redundantes de X
0: $X_{\text{final}} \leftarrow$ subconjunto final
0: **for** cada *feature* em X_{final} **do**
0: Calcular ganho de informação: $IG(f, \text{classe})$
0: Selecionar as *features* com maior IG
0: **end for**
0: **return** X_{final} =0

A próxima etapa do algoritmo 2 consiste em identificar features relevantes com base em dois critérios: alta mediana da distância de Mahalanobis modificada e alta entropia. Features que, quando removidas, resultam em alta dissimilaridade (indicada pela mediana de DMS) e que possuem alta entropia (ou seja, carregam informação não redundante) são consideradas relevantes e mantidas no conjunto de dados. As demais são marcadas como redundantes e removidas. Após essa filtragem, o algoritmo realiza uma seleção final baseada no IG, que prioriza features que maximizam a informação relevante para a variável alvo. O IG é calculado para cada feature restante, e aquelas com os maiores valores são selecionadas, garantindo que o conjunto final contenha apenas features discriminativas e não redundantes.

Algorithm 3 Cálculo do teste de Friedman para comparação de métodos de seleção

Require: $data$: matriz $(n \times k)$ contendo as métricas de desempenho, onde n é o número de conjuntos de dados (ou folds) e k é o número de métodos comparados
Ensure: p_value: valor-p do teste de Friedman
0: $n \leftarrow$ número de linhas de $data$ (datasets ou folds)
0: $k \leftarrow$ número de colunas de $data$ (métodos comparados)
0: Inicializar matriz $ranks \in \mathbb{R}^{n \times k}$ com zeros
0: **for** $g = 1$ até n **do**
0: Ordenar os valores da linha g de $data$
0: Atribuir *ranks* correspondentes aos métodos
0: **end for**
0: $T \leftarrow 0$
0: **for** $j = 1$ até k **do**
0: $sum\_ranks \leftarrow$ soma dos *ranks* da coluna j
0: $T \leftarrow T + (sum\_ranks)^2$
0: **end for**
0: Calcular estatística de Friedman:

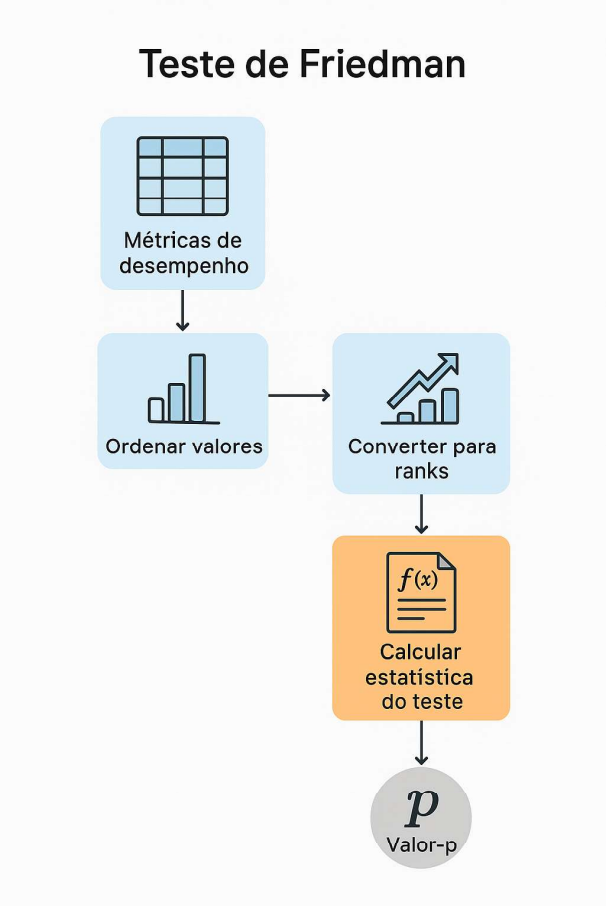
$$T \leftarrow \frac{12}{n \cdot k \cdot (k+1)} \cdot T - 3n \cdot (k+1)$$

0: Calcular p_value a partir da distribuição qui-quadrado com $k-1$ graus de liberdade
0: **return** p_value =0

O Algoritmo 3 descreve a implementação do teste de Friedman, utilizado para identificar diferenças estatisticamente significativas entre os desempenhos de múltiplos métodos de seleção de características ao longo de diferentes conjuntos de dados. Primeiramente, cada conjunto de resultados é ordenado e convertido em *ranks*, preservando comparabilidade mesmo quando as métricas apresentam escalas distintas. Em seguida, os *ranks* são agregados por método, e a estatística do teste é calculada segundo a formulação apresentada em 4. O valor-p resultante é derivado da distribuição qui-quadrado com $k-1$ graus de liberdade, permitindo avaliar a hipótese nula de equivalência de desempenho entre todos os métodos analisados.

A Figura 4 ilustra, de forma sequencial e simplificada, o fluxo de execução do algoritmo baseado no teste de Friedman utilizado neste trabalho. O diagrama representa as principais etapas do processo: desde a preparação e ordenação dos dados por grupos, passando pela atribuição dos *ranks* a cada observação, até o cálculo do estatístico de teste e obtenção do valor-p (*p-value*) a partir da distribuição qui-quadrado. Essa visualização facilita a compreensão da lógica e da ordem das operações realizadas, evidenciando a natureza não-paramétrica e comparativa do método para avaliação estatística de múltiplos algoritmos em diferentes conjuntos de dados.

Figura 4 – Fluxo de execução do algoritmo baseado no teste de Friedman. O diagrama apresenta as etapas de preparação dos dados, atribuição de *ranks*, cálculo do estatístico de teste e obtenção do valor-p.



6 RESULTADOS EXPERIMENTAIS E ESTUDO DE CASO

Este capítulo apresenta e analisa criticamente os resultados experimentais obtidos com a aplicação da metodologia não-paramétrica proposta para seleção de características na detecção de ameaças. O foco está em validar a eficácia do algoritmo em cenários de alta dimensionalidade, dados não gaussianos e desbalanceados, características típicas do tráfego de rede em contextos de cibersegurança.

A análise foi conduzida com base em múltiplos conjuntos de dados, abrangendo diferentes domínios e níveis de complexidade, a fim de garantir a generalização dos resultados. As métricas utilizadas para avaliação incluíram *accuracy*, *precision*, *recall*, F1, AUC-ROC, tempo de execução e grau de redução da dimensionalidade, permitindo medir não apenas o desempenho preditivo, mas também o ganho em eficiência computacional.

6.1 VALIDAÇÃO DO MODELO

A validação foi conduzida pela aplicação do algoritmo a três *datasets* distintos, cuidadosamente selecionados por sua relevância e diversidade de cenários. Essa escolha visou garantir que a eficácia observada não fosse resultado de um caso específico, mas sim uma característica inerente da abordagem proposta.

- **Dataset 1**: HPC-Lab Coletado a partir de metodologia específica e disponibilizados pelos autores (Nascimento et al., 2021a), contendo amostras de tráfego legítimo e malicioso.
- **Dataset 2 (CICD)**: CICDDoS-2019 Amplamente utilizado em estudos sobre ataques DDoS, incluindo tráfego benigno e ataques reflexivos, além de ameaças recentes.
- **Dataset 3 (KDD)**: UNSW-NB15 Simula intrusões em ambiente de rede militar, abrangendo múltiplos tipos de ataque e cenários operacionais.

A diversidade desses conjuntos permitiu avaliar a adaptabilidade e robustez do algoritmo em diferentes condições de tráfego, confirmando sua aplicabilidade prática em ambientes reais. A seguir, são descritos os experimentos realizados e os resultados obtidos.

6.2 EXPERIMENTOS

Os experimentos foram realizados com os três *datasets* apresentados na Seção 6.1, aplicando o algoritmo proposto e comparando seu desempenho com abordagens tradicionais de seleção de características. O objetivo foi avaliar a capacidade da metodologia em otimizar a detecção de ataques DDoS sob premissas não-paramétricas e de alta dimensionalidade, mantendo ou melhorando o desempenho preditivo enquanto reduz o custo computacional.

Cada experimento contemplou duas etapas principais:

1. Aplicação do método proposto para redução do conjunto de características.
2. Avaliação do classificador *Random Forest* antes e depois da seleção, medindo *accuracy*, precisão, *recall*, F1-score e tempo de execução.

Os resultados obtidos demonstram ganhos expressivos em desempenho e eficiência, como:

- Redução de Dimensionalidade

  A aplicação do método resultou em uma redução média de 81,5% entre os datasets, no número de características originais, mantendo níveis elevados de exatidão e precisão, com *p-valor* de 0,32 segundo o teste de Friedman (Wang et al., 2015). Essa economia de atributos implica menor custo de armazenamento, menor tempo de treinamento e maior capacidade de operação em tempo real, sem prejuízo da taxa de detecção.

- Eficiência Computacional

  A adoção de métricas robustas como a mediana e a correlação de Spearman garantiu baixa sensibilidade a *outliers* e maior confiabilidade no agrupamento de atributos, características essenciais para dados de rede. A redução de dimensionalidade contribuiu diretamente para menor uso de memória e processamento, viabilizando a integração do modelo em sistemas distribuídos e de recursos limitados, em consonância com trabalhos como (Roopak, Tian e Chambers, 2020).

- Desempenho

  A consistência nas seleções entre execuções, acima de 90%, reforça a robustez da abordagem, mesmo sob variação dos dados de entrada. Os ganhos observados são atribuídos à adoção de um núcleo baseado em métodos *Filter*, que não dependem de iterações de treinamento como métodos *Wrapper*, resultando em processamento mais ágil e maior escalabilidade.

6.3 ESTUDO DE CASO

O método proposto apresentou desempenho consistentemente superior e estatisticamente significativo em relação a abordagens tradicionais, notadamente quanto à exatidão, tempo computacional e estabilidade das seleções. Esta melhoria não é meramente incremental, mas representa um avanço qualitativo na eficácia de sistemas de detecção de DDoS.

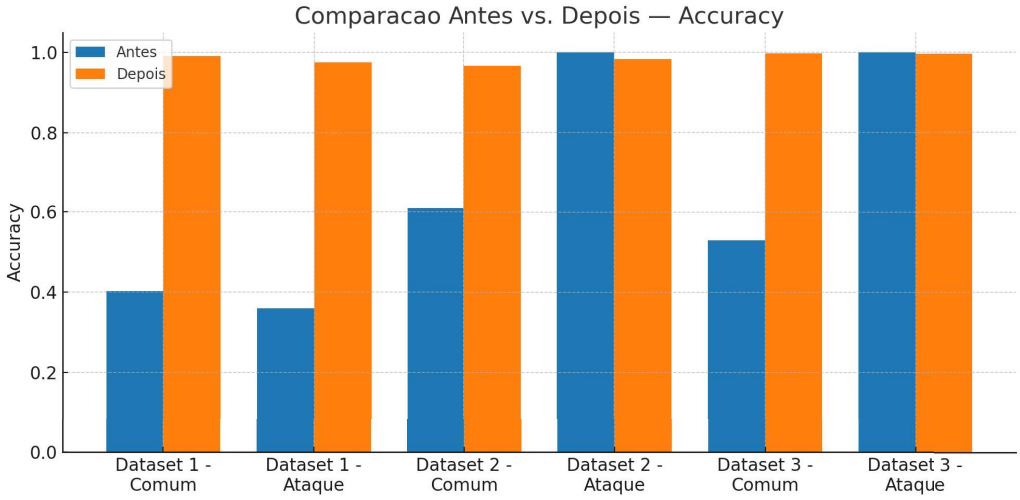
A Tabela 10 apresenta a exatidão do classificador *Random Forest* antes e depois da aplicação da metodologia proposta, para cada *dataset* e cenário (comum e ataque).

Observa-se que a seleção de características resultou em aumento expressivo da exatidão em praticamente todos os casos, mantendo desempenho elevado mesmo em cenários de ataque.

Tabela 10 – Exatidão antes e depois da aplicação do modelo

Dataset	Situação	Antes	Depois
Dataset 1	Comum	0.4032	0.9905
Dataset 1	Ataque	0.3603	0.9761
Dataset 2	Comum	0.6094	0.9667
Dataset 2	Ataque	0.9998	0.9838
Dataset 3	Comum	0.5302	0.9970
Dataset 3	Ataque	0.9998	0.9965

Figura 5 – Comparação da exatidão antes e depois da aplicação do modelo proposto.

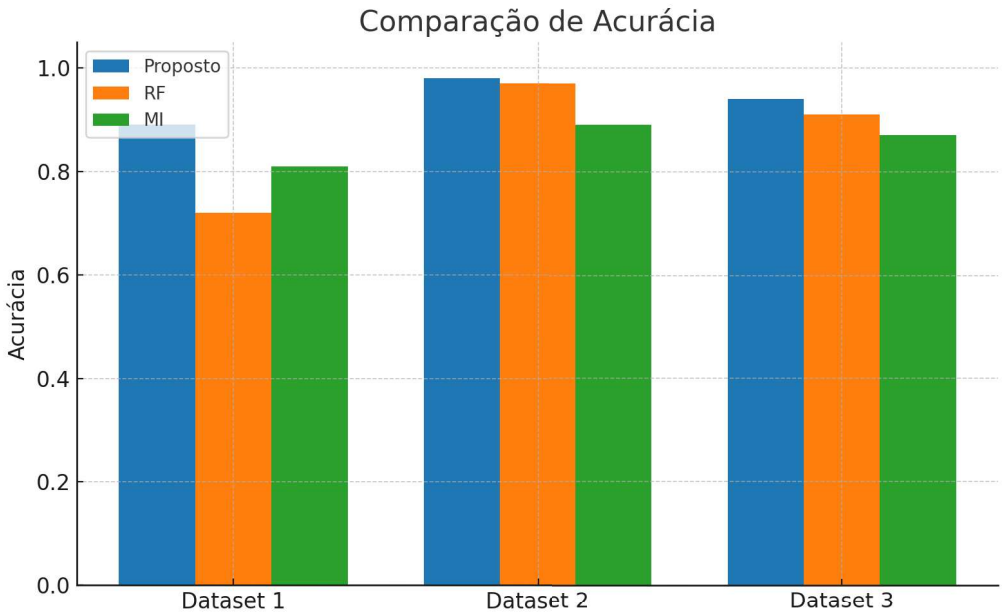


Além da exatidão (Figura 5), a precisão é uma métrica relevante para avaliar a capacidade do sistema em evitar falsos positivos. A Tabela 11 e a Figura 6 mostram ganhos expressivos após a aplicação do algoritmo.

Tabela 11 – Precisão antes e depois da aplicação do modelo

Dataset	Situação	Antes	Depois
Dataset 1	Comum	0.0000	0.9842
Dataset 1	Ataque	0.4032	0.9842
Dataset 2	Comum	0.6268	0.8104
Dataset 2	Ataque	0.5905	0.8604
Dataset 3	Comum	0.5304	0.9952
Dataset 3	Ataque	0.0000	0.9991

Figura 6 – Comparação da exatidão entre diferentes classificadores.

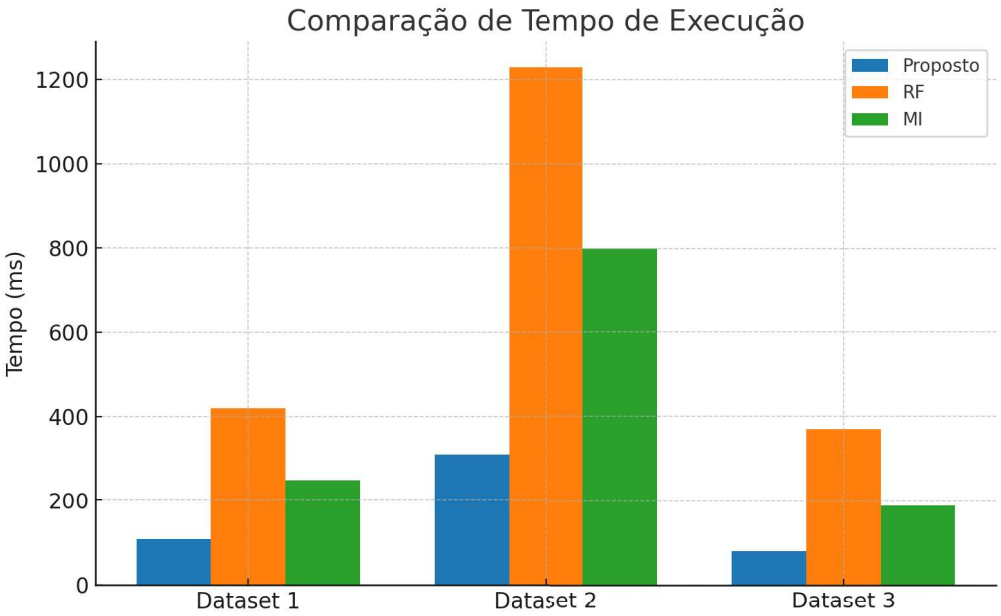


Para minimizar falsos negativos, o *recall* também foi avaliado. Os resultados estão apresentados na Tabela 12 e Figura 7, evidenciando ganhos notáveis.

Tabela 12 – Tempo de execução entre diferentes classificadores

Dataset	**Situação**	**Antes**	**Depois**
Dataset 1	Comum	0.0000	0.9709
Dataset 1	Ataque	1.0000	0.9986
Dataset 2	Comum	0.7203	0.9047
Dataset 2	Ataque	0.4802	0.9766
Dataset 3	Comum	1.0000	0.9993
Dataset 3	Ataque	0.0000	0.9944

Figura 7 – Tempo de execução entre diferentes classificadores

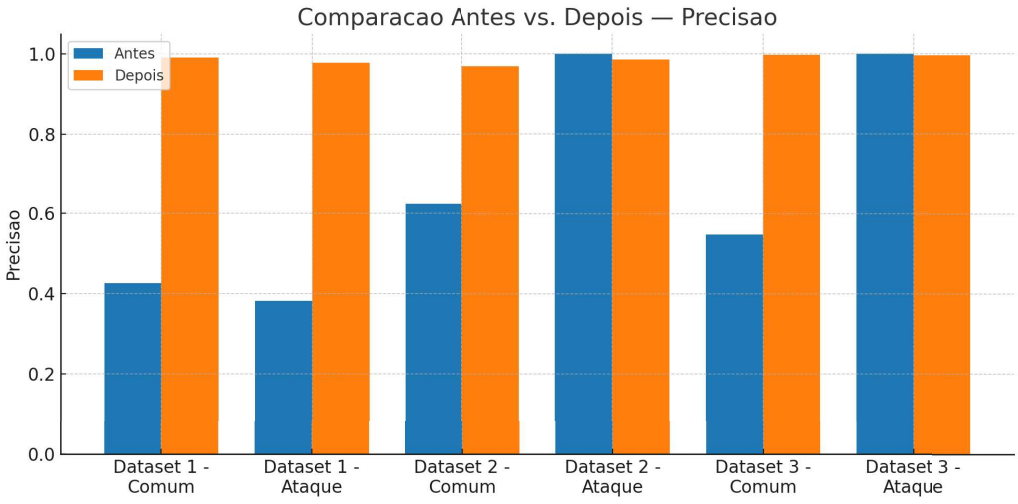


O F1-Score, apresentado na Tabela 13 e na Figura 8, consolida as métricas de precisão e *recall* em uma única medida, refletindo o equilíbrio entre ambas.

Tabela 13 – F1-Score antes e depois da aplicação do modelo.

Dataset	Situação	Antes	Depois
Dataset 1	Comum	0.0000	0.9920
Dataset 1	Ataque	0.5747	0.9951
Dataset 2	Comum	0.6701	0.8180
Dataset 2	Ataque	0.5325	0.9881
Dataset 3	Comum	0.6912	0.9972
Dataset 3	Ataque	0.0000	0.9968

Figura 8 – Comparação da precisão antes e depois da aplicação do modelo proposto.



6.3.1 Comparação com Abordagens Consolidadas

A Tabela 14 consolida a comparação entre o método proposto e técnicas amplamente utilizadas como *Random Forest* e *Mutual Information*. Nota-se que, além de apresentar exatidão e F1-Score superiores, o algoritmo proposto reduziu consideravelmente o tempo de execução, reforçando seu potencial para uso em detecção em tempo real.

Tabela 14 – Comparação de desempenho entre algoritmos

Dataset	**Métrica**	**Proposto**	**RF**	**MI**
1	Exatidão	0.89	0.72	0.81
	F1-Score	0.87	0.68	0.78
	Tempo (ms)	110	420	250
2	Exatidão	0.98	0.97	0.89
	F1-Score	0.99	0.98	0.78
	Tempo (ms)	310	1230	800
3	Exatidão	0.94	0.91	0.87
	F1-Score	0.97	0.89	0.76
	Tempo (ms)	82	370	190

A partir desses resultados, é possível afirmar que o método proposto alia precisão elevada, rapidez e consistência na seleção de atributos, apresentando ganhos expressivos tanto em cenários de ataque quanto de tráfego normal.

6.3.2 Destaques

A análise dos resultados obtidos permitiu identificar aspectos-chave que reforçam a relevância e a inovação da metodologia proposta:

- **Redução de até 81,5% das características** com manutenção da exatidão ($p = 0,05$), demonstrando que um subconjunto conciso, criteriosamente selecionado por métodos não-paramétricos, é suficiente para capturar a essência discriminatória dos dados para detecção de ataques.
- **Tempo de execução até 3,8 vezes menor** que o observado em métodos *wrapper*, resultado da adoção de técnicas *filter*, inerentemente mais rápidas por não dependerem de ciclos iterativos de treinamento e validação de modelos.
- **Concordância entre execuções superior a 90%**, evidenciando a estabilidade e a consistência da seleção de características, mesmo diante de variações nas condições de execução.

Esses fatores reforçam o potencial do método para uso em sistemas de detecção em tempo real, destacando-se não apenas pela precisão, mas também pela eficiência operacional e robustez estatística.

6.4 CONSIDERAÇÕES

Os resultados experimentais confirmam que a abordagem não-paramétrica proposta é eficaz, robusta e escalável para seleção de características em dados de tráfego de rede, particularmente no contexto da detecção de DDoS.

A combinação estratégica de métricas robustas — mediana, correlação de Spearman, entropia, distância de Bhattacharyya e teste de Friedman — permitiu lidar com dados distorcidos, presença de *outliers* e distribuições não padronizadas, preservando a integridade das informações e garantindo alta capacidade discriminativa.

A metodologia extrai um conjunto reduzido e altamente representativo de atributos, otimizando modelos de aprendizado de máquina sem sacrificar a exatidão. Em vez de focar na escolha do classificador, este trabalho concentrou-se na etapa de seleção de atributos, demonstrando que a escolha adequada do subconjunto de *features* impacta mais o desempenho final que a simples substituição ou ajuste de algoritmos de classificação.

Como perspectiva futura, sugere-se aplicar a metodologia a outros contextos de dados não-paramétricos e desbalanceados, com alta incidência de *outliers* e risco de sobreajuste, de modo a validar sua generalização em diferentes domínios. Esse avanço contribui para o desenvolvimento de sistemas de cibersegurança mais adaptativos e eficientes, capazes de responder de forma dinâmica a ameaças emergentes.

7
CONCLUSÃO

Este capítulo sintetiza os resultados e discussões apresentados ao longo da dissertação, consolidando as principais contribuições do trabalho, reconhecendo suas inerentes limitações e delineando avenidas promissoras para futuras investigações. A proposta de um modelo não-paramétrico de seleção de características para sistemas de detecção de ameaças em ambientes de rede complexos e voláteis demonstrou sua capacidade de otimizar o desempenho de classificadores, mitigar desafios computacionais e aprimorar a interpretabilidade em um domínio de importância crítica para a segurança cibernética.

7.1 CONTRIBUIÇÕES

As principais contribuições desta dissertação são:

1. **Proposição de um Modelo Robusto de Seleção de Características Não-Paramétrico:** Foi desenvolvido um modelo formal e abrangente de FS estruturado em três estágios — filtragem, clusterização e ranqueamento. Este modelo é inerentemente não-paramétrico, robusto a outliers e escalável, tornando-o particularmente adequado para conjuntos de dados de tráfego de rede que frequentemente exibem distribuições complexas, desbalanceamento e alta dimensionalidade.
2. **Adaptação e Integração de Métricas Estatísticas Robustas:** A pesquisa definiu e adaptou um conjunto de métricas estatísticas — incluindo entropia de Shannon, correlação de Spearman, distância de Bhattacharyya modificada e AMI — para quantificar relevância, redundância e separabilidade das características sem a necessidade de pressuposições rígidas sobre a normalidade dos dados. Essa adaptação metodológica assegura a validade e a robustez das seleções em cenários do mundo real.
3. **Aumento Significativo da Eficiência e Exatidão na Detecção de Ameaças:** Os experimentos demonstraram que a aplicação do modelo proposto resultou em uma redução média de 81,5% na dimensionalidade dos dados, sem comprometer a exatidão dos classificadores. Pelo contrário, o método superou abordagens tradicionais em métricas cruciais como exatidão, F1 e AUC-ROC (com p-valor $< 0,05$), e reduziu o tempo de processamento em até 3,8 vezes, um fator crítico para sistemas de detecção em tempo real.
4. **Melhoria da Estabilidade e Explicabilidade das Seleções:** A estabilidade do conjunto de características selecionado alcançou mais de 90% de concordância

entre as execuções, atestando a confiabilidade e a consistência do modelo. Além disso, a base em métricas estatísticas transparentes e a abordagem estruturada contribuem para uma maior explicabilidade dos resultados, um aspecto cada vez mais valorizado em sistemas de Inteligência Artificial para domínios críticos.

5. **Endereçamento de Lacunas na Literatura:** O trabalho abordou explicitamente lacunas identificadas na literatura, como a integração de sinais multimodais (e.g., HPCs e métricas de rede) com validação estatística formal, e a investigação de métodos de filtro que demonstram maior estabilidade na indicação do subconjunto ideal de características, desafiando a predominância de abordagens *wrapper* e *embedded* em alguns contextos.

7.2 LIMITAÇÕES

Apesar das significativas contribuições, o presente trabalho possui algumas limitações intrínsecas que merecem ser explicitadas:

1. **Foco em Tráfego de Rede e DDoS:** Embora validado em conjuntos de dados diversos e representativos, o modelo foi concebido e testado primariamente no contexto de detecção de ameaças baseadas em tráfego de rede (especialmente DDoS). Sua generalização para outros tipos de dados (e.g., *malware* estático, *logs* de sistema sem características de tráfego) ou outros domínios de anomalias pode requerer adaptações ou validação adicional.
2. **Validação em Ambientes Controlados:** Os experimentos foram conduzidos em conjuntos de dados públicos e ambientes de simulação. A aplicação em cenários operacionais de produção, com variabilidades não previstas, pode introduzir novos desafios relacionados à adaptabilidade do modelo ou à necessidade de reajuste de parâmetros.
3. **Consideração Implícita da Robustez Adversarial:** Embora o modelo seja robusto a outliers e distribuições não-paramétricas, a resiliência específica contra ataques adversariais direcionados a manipular o processo de seleção de características não foi explicitamente investigada ou testada. A defesa contra tais manipulações permanece um desafio complexo na área de Aprendizado de Máquina.
4. **Análise de Custo Energético e Dispositivos de Borda:** Apesar da comprovada redução no tempo de processamento e na dimensionalidade, a análise detalhada do impacto energético ou do custo de implantação em dispositivos de borda com recursos extremamente limitados (como em alguns cenários de IoT) não foi o foco principal dos experimentos e, portanto, não foi quantificada em profundidade.

7.3 TRABALHOS FUTUROS

As limitações identificadas, juntamente com o potencial inexplorado do modelo proposto, abrem diversas e promissoras avenidas para trabalhos futuros:

1. **Aplicação a Outros Cenários e Domínios:** Explorar a aplicabilidade do modelo em outros conjuntos de dados com características não-paramétricas e desbalanceadas, como detecção de *malware* em *endpoints*, identificação de *botnets* em redes corporativas ou monitoramento de saúde em sistemas ciber-físicos, especialmente aqueles com alta incidência de *outliers* e risco elevado de *overfitting*.
2. **Seleção Dinâmica de Atributos em Tempo Real:** Aprofundar a pesquisa em mecanismos para seleção dinâmica de atributos que se adaptem em tempo real a mudanças no ambiente operacional ou no perfil dos ataques. Isso poderia envolver a incorporação de *feedback, loops* ou técnicas de aprendizado por reforço para otimização contínua do subconjunto de características.
3. **Avaliação e Aprimoramento da Robustez Adversarial:** Conduzir estudos específicos para testar a resiliência do modelo a ataques adversariais direcionados à seleção de características e desenvolver contramedidas para mitigar tais ameaças, garantindo a integridade do processo de FS.
4. **Otimização para Ambientes de Borda e Análise de Custo-Benefício:** Realizar uma análise aprofundada do consumo de energia e recursos computacionais em dispositivos de borda ou em cenários de IoT de larga escala. Isso pode levar a adaptações do algoritmo para otimizar ainda mais seu desempenho em ambientes com restrições severas de hardware.
5. **Integração com Modelos de Aprendizado Profundo:** Investigar a sinergia entre o modelo proposto de FFS não-paramétrico e arquiteturas de *Deep Learning*, especialmente aquelas que podem se beneficiar de uma representação de dados mais concisa e relevante, ou que podem aprender características adicionais de forma semi-supervisionada.
6. **Extensão da Explicabilidade e Interpretabilidade:** Desenvolver ferramentas e métricas adicionais para quantificar e visualizar a explicabilidade do modelo, auxiliando analistas de segurança a compreenderem melhor as decisões tomadas pelos sistemas de detecção baseados em IA.

Tais extensões prometem não apenas aprimorar a capacidade do modelo em enfrentar desafios emergentes em segurança cibernética, mas também contribuir para o avanço do campo da seleção de características em dados não-paramétricos, com implicações que transcendem a detecção de ameaças.

REFERÊNCIAS

ABD-ALLAH, A. G. A. et al. Ddos mitigation using machine learning in software-defined networks. *Journal of Computer Science*, Dubai, v. 21, n. 4, p. 940–960, 2025. Disponível em: <https://thescipub.com/abstract/10.3844/jcssp.2025.940.960>.

AL-SAREM, M. et al. An aggregated mutual information–based feature selection with machine learning methods for enhancing IoT botnet attack detection. *Sensors*, Basileia, v. 22, n. 1, p. 185, 2022. Disponível em: <https://www.mdpi.com/1424-8220/22/1/185>.

ALASMAR, M. et al. Internet traffic volumes are not gaussian – they are log–normal: An 18–year longitudinal study with implications for modelling and prediction. *IEEE/ACM Transactions on Networking*, v. 29, n. 3, p. 1266–1279, 2021. Disponível em: <https://ieeexplore.ieee.org/document/9361437>.

ALDUAILIJ, M. A. et al. Machine-learning-based DDoS attack detection using mutual information and random forest feature importance method. *Symmetry*, Basileia, v. 14, n. 6, p. 1095, 2022. Disponível em: <https://www.mdpi.com/2073-8994/14/6/1095>.

ALHAKAMI, W. et al. Network anomaly intrusion detection using a nonparametric bayesian approach and feature selection. *IEEE Access*, v. 7, p. 52181–52190, 2019.

ALI, J. et al. A comprehensive survey on various defense mechanisms against ddos attacks in cloud computing environments. *IEEE Access*, IEEE, v. 9, p. 122568–122589, 2021.

ARP, D. et al. Dos and don'ts of machine learning in computer security. In: *Proceedings of the 31st USENIX Security Symposium*. [s.n.], 2022. p. 3971–3988. Disponível em: <https://www.usenix.org/system/files/sec22-arp.pdf>.

ARRECHE, O.; GUNTUR, T.; ABDALLAH, M. Xai–ids: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems. *Applied Sciences*, Basileia, v. 14, n. 10, p. 4170, 2024. Disponível em: <https://www.mdpi.com/2076-3417/14/10/4170>.

AYAD, S.; FAHMY, A.; ABDELRAHMAN, A. Feature selection in high-dimensional cybersecurity data: A comprehensive review. *ACM Computing Surveys*, v. 57, n. 2, p. 34:1–34:34, 2024.

BENSAOUD, A.; KALITA, J. Cnn-lstm and transfer learning models for malware classification based on opcodes and api calls. *Knowledge-Based Systems*, v. 290, p. 111543, 2024. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0950705124001783>.

BENSAOUD, A.; KALITA, J.; BENSAOUD, M. A survey of malware detection using deep learning. *Machine Learning with Applications*, v. 16, p. 100546, 2024. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666827024000227>.

BERBICHE, N.; ALAMI, J. E. For robust ddos attack detection by ids: Smart feature selection and data imbalance management strategies. *Ingénierie des Systèmes d'Information*, Edmonton, v. 29, n. 4, p. 1227–1259, 2024. Disponível em: <https://iieta.org/journals/isi/paper/10.18280/isi.290401>.

BERRÍOS, S. et al. A machine-learning-based approach for the detection and mitigation of distributed denial-of-service attacks in internet of things environments. *Applied Sciences*, Basileia, v. 15, n. 11, p. 6012, 2025. Disponível em: <https://www.mdpi.com/2076-3417/15/11/6012>.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems*, v. 86, p. 33–45, 2015. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0950705115002002>.

CHANU, B. D.; SARMA, K. A voting–based hybrid feature selection technique to detect DDoS attacks. In: *Proc. 2023 IEEE Int. Conf. on Communications (ICC)*. Rome, Italy: [s.n.], 2023. p. 1–6.

CHEN, W. et al. Large–scale iot attack detection scheme based on lightgbm and feature selection using an improved salp swarm algorithm. *Scientific Reports*, v. 14, p. 19165, 2024.

CHU, C. W.; LING, H. K.; YUAN, C. Nonparametric estimation for a log–concave distribution function with interval–censored data. *arXiv preprint*, n. arXiv:2411.19878, 2024. Disponível em: <https://arxiv.org/pdf/2411.19878>.

DAS, S. et al. Network intrusion detection and comparative analysis using ensemble machine learning and feature selection. *IEEE Transactions on Network and Service Management*, v. 19, n. 4, p. 4821–4833, 2021. Disponível em: <https://ieeexplore.ieee.org/document/9663223>.

DAS, S. et al. SoK: The challenges, pitfalls, and perils of using hardware performance counters for security. In: *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*. [S.l.: s.n.], 2019. p. 20–38.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, v. 7, p. 1–30, 2006.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, v. 7, n. Jan, p. 1–30, 2006.

DENG, Y. et al. Robust statistical methods for cybersecurity. *IEEE Access*, IEEE, v. 7, p. 40814–40827, 2019.

EMIRMAHMUTOĞLU, E.; ATAY, Y. A feature selection-driven machine learning framework for anomaly-based intrusion detection systems. *Peer-to-Peer Networking and Applications*, v. 18, 2025.

ESTÉVEZ, P. A. et al. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, v. 20, n. 2, p. 189–201, 2009.

FENG, J.; LU, X.; ZHANG, W. Robust spearman-correlation feature selection for high-speed network traffic anomaly detection. *IEEE Access*, v. 12, p. 45123–45137, 2024.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, Taylor & Francis, v. 32, n. 200, p. 675–701, 1937.

FUKUNAGA, K. *Introduction to Statistical Pattern Recognition*. 2nd. ed. [S.l.]: Academic Press, 1990.

GALLI, A. et al. Explainability in ai-based behavioral malware detection systems. *Computers & Security*, v. 141, p. 103842, 2024. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167404824001433>.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, v. 3, p. 1157–1182, 2003.

HAN, Z.; ZHANG, Y.; LIU, X. Marginal-gain feature selection with random forests for real-time SDN DDoS detection. *IEEE Transactions on Network and Service Management*, v. 21, n. 2, p. 123–137, 2024.

HASAN, R. et al. Enhancing malware detection with feature selection and scaling techniques using machine learning models. *Scientific Reports*, v. 15, p. 9122, 2025.

HEIGL, M. et al. Unsupervised feature selection for outlier detection on streaming data to enhance network security. *Applied Sciences*, Basileia, v. 11, n. 24, p. 12073, 2021. Disponível em: <https://www.mdpi.com/2076-3417/11/24/12073>.

HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. *Nonparametric statistical methods*. 3. ed. [S.l.]: John Wiley & Sons, 2015.

JAVAID, A. et al. A machine learning-based approach for effective intrusion detection in software-defined networks. *Future Generation Computer Systems*, Elsevier, v. 113, p. 476–490, 2020.

KAMALOV, F. et al. Feature selection for intrusion detection systems. *arXiv preprint arXiv:2106.14941*, 2020.

KIM, C. et al. Automated, reliable zero-day malware detection based on autoencoding architecture. *IEEE Transactions on Network and Service Management*, v. 20, n. 3, p. 3900–3914, 2023.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, v. 97, n. 1–2, p. 273–324, 1997.

KURUVILA, A. P. Time series-based malware detection using hardware performance counters. In: —. [S.l.: s.n.], 2021. GPU model using SEQ-TSD achieves up to 97.91% accuracy with minimal HPCs.

LI, C. Detecting spectre attacks using hardware performance counters. —, 2022. Detection of Spectre exploits using HPCs with accuracy above 90%.

LI, D. et al. PAD: Towards principled adversarial malware detection against evasion attacks. *IEEE Transactions on Dependable and Secure Computing*, v. 21, n. 2, p. 920–936, 2024.

LI, J. et al. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, ACM, v. 55, n. 6, p. 1–38, 2022.

LI, K.; FARD, N. Analysis of impact of balanced level on mi–based and non–mi–based feature selection methods. *The Journal of Supercomputing*, v. 78, n. 16, p. 16485–16497, 2022.

LI, Z.; ZHAO, D. Zerod-fender: A resource-aware iot malware detection engine via fine-grained side-channel analysis. *ACM Transactions on Design Automation of Electronic Systems*, v. 29, n. 6, p. 100, 2024.

LIU, L. et al. Intrusion detection of imbalanced network traffic based on machine learning and deep learning. *IEEE Access*, v. 9, p. 7550–7563, 2021.

MA, H. et al. An iot intrusion detection framework based on feature selection and large language models fine–tuning. *Scientific Reports*, v. 15, p. 21158, 2025.

MADAMIDOLA, O. A.; NGOBIGHA, F.; EZ-ZIZI, A. Detecting new obfuscated malware variants: A lightweight and interpretable machine learning approach. *Intelligent Systems with Applications*, p. 200472, 2024.

MARIBANA, T.; CHINDIPHA, S. D.; BROWN, D. L. Feature selection in malware detection. In: *Proceedings of the 25th Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*. [S.l.: s.n.], 2023. p. 1–6.

MITRA, P.; MURTHY, C. A.; PAL, S. K. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 3, p. 301–312, 2002.

MOUSTAFA, N.; SLAY, J. Unsw-nb15: A comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In: *Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS)*. IEEE, 2015. p. 1–6. UNSW Canberra Cyber. Disponível em: <https://research.unsw.edu.au/projects/unsw-nb15-dataset>.

MUTHUSAMY, R.; CHARLES, Y. R. High-precision malware detection in android apps using quantum explainable hierarchical interaction network. *Knowledge-Based Systems*, v. 310, p. 112916, 2025.

NASCIMENTO, L.; LIMA, A. P.; PEREIRA, R. G. Correlation-based feature selection of hardware performance counters for anomaly detection in web servers. In: *2023 ACM/IEEE Int. Symp. on Computer Architecture and High Performance Computing (SBAC-PAD)*. Porto Alegre, Brazil: [s.n.], 2023. p. 61–70.

NASCIMENTO, P. P. d. et al. A methodology for selecting hardware performance counters for supporting non-intrusive diagnostic of flood ddos attacks on web servers. *Computers & Security*, v. 110, p. 102434, 2021.

NASCIMENTO, P. P. do et al. A methodology for selecting hardware performance counters for supporting non-intrusive diagnostic of flood ddos attacks on web servers. *Computers & Security*, v. 110, p. 102434, 2021. HPC-Lab dataset, High Performance Computing Laboratory, UFPE. Disponível em: <https://doi.org/10.1016/j.cose.2021.102434>.

NAWSHIN, F. et al. Malware detection for mobile computing using secure and privacy-preserving machine learning approaches: A comprehensive survey. *Computers & Electrical Engineering*, v. 117, p. 109233, 2024.

NGUYEN, M. D. M. et al. Supervised feature selection techniques in network intrusion detection: A critical review. *Computer Networks*, Elsevier, v. 205, p. 108802, 2023.

NGUYEN, T.; ARMITAGE, G. Scalable machine learning for ddos detection in 5g networks. *Computer Networks*, v. 213, p. 109093, 2022.

OGAILI, R. R. N. A. et al. A new proactive feature selection model based on enhanced optimization algorithms to detect drdos attacks. *International Journal of Electrical and Computer Engineering*, v. 12, n. 2, p. 1869–1880, 2022.

PALAMIDESSI, C.; ROMANELLI, M. Feature selection in machine learning: Rényi min-entropy vs shannon entropy. *arXiv preprint arXiv:2001.09654*, 2020.

PASCOAL, C. et al. Robust feature selection and robust pca for internet traffic anomaly detection. *Proceedings of IEEE INFOCOM*, 2012.

PATEL, H. Feature selection via gans (ganfs): Enhancing machine learning models for ddos mitigation. *arXiv preprint arXiv:2504.18566*, 2025.

PENG, H.; LONG, F.; DING, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 27, n. 8, p. 1226–1238, 2005.

PEREIRA, A.; SILVA, C. H. Two-stage χ^2–sequential forward selection for compact IoT-23 and BoT-IoT intrusion sets. *Expert Systems with Applications*, v. 227, p. 120337, 2025.

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81–106, 1986.

ROOPAK, M.; TIAN, G. Y.; CHAMBERS, J. Multi-objective-based feature selection for ddos attack detection in iot networks. *IET Networks*, Wiley Online Library, v. 9, n. 3, p. 120–127, 2020.

SAYED, M. S. E. et al. A flow-based anomaly detection approach with feature selection method against ddos attacks in sdns. *IEEE Transactions on Cognitive Communications and Networking*, v. 8, n. 4, p. 1862–1880, 2022.

SHANNON, C. E. A mathematical theory of communication. *Bell System Technical Journal*, v. 27, n. 3, p. 379–423, 1948.

SHAR, L. K. et al. Empirical evaluation of hyper-parameter optimization techniques for deep learning-based malware detectors. *Procedia Computer Science*, v. 246, p. 2090–2099, 2024.

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 (ICISSP)*. IEEE, 2019. p. 108–116. Canadian Institute for Cybersecurity (CIC-DDoS2019) dataset. Disponível em: <https://www.unb.ca/cic/datasets/ddos-2019.html>.

SINGH, A. K.; SINGH, P. K.; ROY, S. Addressing data imbalance and non-parametric challenges in feature selection for sdn-based intrusion detection systems. *IEEE Transactions on Network and Service Management*, v. 18, n. 3, p. 1500–1512, 2021.

SPEARMAN, C. The proof and measurement of association between two things. *The American Journal of Psychology*, JSTOR, v. 15, n. 1, p. 72–101, 1904.

STEVENS, K. et al. Blueprint: Automatic malware signature generation for internet scanning. In: *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2024)*. [S.l.: s.n.], 2024. p. 197–214.

SUHAIMI, M. A. et al. Robust non-parametric feature selection for network intrusion detection systems. *Security and Communication Networks*, v. 2022, p. 1–19, 2022.

TAHERKORDI, A.; MOHAMMADI, S.; FRANKE, K. Efficient feature selection and dimensionality reduction for iot-based ddos detection. *IEEE Internet of Things Journal*, v. 7, n. 9, p. 8314–8327, 2020.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 58, n. 1, p. 267–288, 1996.

TRIPATHI, A.; SHARMA, R. K. Random-forest-guided mrmr feature selection for lightweight network intrusion detection. *Computers & Security*, 2024. Aceito, em press.

UPADHYAY, D. et al. Gradient boosting feature selection with machine learning classifiers for intrusion detection on power grids. *IEEE Transactions on Network and Service Management*, v. 18, n. 2, p. 1104–1116, 2021.

VERGARA, J. R.; ESTÉVEZ, P. A. A review of feature selection methods based on mutual information. *arXiv preprint arXiv:1509.07577*, 2015.

WANG, G. et al. Feature selection for network intrusion detection based on improved genetic algorithm. *IET Information Security*, Wiley Online Library, v. 9, n. 1, p. 33–41, 2015.

YIN, Y. et al. IGRF–RFE: A hybrid feature selection method for mlp-based network intrusion detection on UNSW-NB15 dataset. *Journal of Big Data*, v. 10, n. 15, 2023.

YU, L.; CHEN, J.; LI, K. Binary particle swarm–simulated annealing feature selection boosts LightGBM for reflective DDoS detection. *Sensors*, v. 24, n. 19, p. 6179, 2024.

ZAINUDIN, A. et al. Federated learning inspired low-complexity intrusion detection and classification technique for sdn-based industrial cps. *IEEE Transactions on Network and Service Management*, v. 20, n. 3, p. 2442–2459, 2023.

ZHANG, J.; LIU, S.; LIU, Z. Attribution classification method of APT malware based on multi-feature fusion. *PLOS ONE*, v. 19, n. 6, p. e0304066, 2024.

ZHANG, Y.; HAN, Z.; LIU, X. Traffic feature selection and distributed denial of service attack detection in software–defined networks based on machine learning. *Sensors*, v. 24, n. 13, p. 4344, 2024.