UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

Guilherme Guerra Campos

Aprendizado de máquina e seu impacto em sistemas de detecção de intrusão: uma revisão sistemática de literatura

RECIFE
2025

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Guilherme Guerra Campos

Aprendizado de máquina e seu impacto em sistemas de detecção de intrusão: uma revisão sistemática de literatura

Monografia apresentada ao Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE), como requisito parcial para conclusão do Curso de Sistemas de Informação, orientada pelo(a) professor(a) Cleber Zanchettin.

RECIFE
2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Campos, Guilherme Guerra.
Aprendizado de máquina e seu impacto em sistemas de detecção de intrusão: uma revisão sistemática de literatura / Guilherme Guerra Campos. - Recife, 2025.
57 p. : il., tab.

Orientador(a): Cleber Zanchettin
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Sistemas de Informação - Bacharelado, 2025.
Inclui referências.

1. Sistema de detecção de intrusão. 2. Aprendizado de máquina. 3. Revisão Sistemática de Literatura. I. Zanchettin, Cleber. (Orientação). II. Título.

600 CDD (22.ed.)

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Guilherme Guerra Campos

Aprendizado de máquina e seu impacto em sistemas de detecção de intrusão: uma revisão sistemática de literatura

Monografia submetida ao corpo docente da Universidade Federal de Pernambuco, defendida e aprovada em 04 de agosto de 2025.

Banca Examinadora:

_____ Orientador(a)

Cleber Zanchettin

Doutor(a)

_____ Examinador(a)

Fernando Maciano de Paula Neto

Doutor(a)

Dedico este trabalho à minha esposa, companheira de todas as horas, cujo amor, paciência e apoio incondicional foram fundamentais em cada etapa desta jornada.

À minha mãe e aos meus irmãos, por sempre acreditarem no meu potencial e me incentivarem, mesmo nos momentos mais difíceis.

E às minhas três cadelas, que com sua presença silenciosa e afeto constante, tornaram os dias de estudo mais leves e aconchegantes.

AGRADECIMENTOS

A realização deste trabalho não seria possível sem o apoio e a presença de pessoas especiais que estiveram ao meu lado ao longo desta caminhada.

Agradeço, em primeiro lugar, à minha esposa, por seu amor, paciência e compreensão em todos os momentos. Sua presença foi essencial para que eu mantivesse o equilíbrio entre os desafios acadêmicos e a vida pessoal.

À minha mãe e aos meus irmãos, agradeço pelo incentivo constante, pelas palavras de motivação e pelo exemplo de dedicação e resiliência que sempre me inspiraram.

Agradeço também ao meu orientador, Professor Cleber Zanchettin, pela orientação precisa, pela confiança depositada em meu trabalho e pelos ensinamentos compartilhados ao longo deste processo.

Por fim, agradeço a todos os amigos, colegas e professores que, de alguma forma, contribuíram para a construção deste trabalho, seja por meio de conversas, apoio técnico ou simples gestos de encorajamento.

A todos, meu sincero obrigado.

“O real não está no início nem no fim, ele se mostra
para a gente é no meio da travessia.”
João Guimarães Rosa

RESUMO

A disseminação de serviços digitais ao longo dos anos tem sido cada vez mais abrangente e em todos os setores do nosso dia a dia. Porém, com essa evolução surgem formas cada vez mais sofisticadas de perpetrar ciberataques no meio digital, levando muitas vezes a prejuízos inestimáveis. A fim de identificar atividades potencialmente maliciosas, a literatura apresenta diversos sistemas de detecção de intrusão (IDS), que relatam atividade perigosa caso os padrões de rede e sistema sejam considerados anômalos. Esses sistemas (IDSs) podem utilizar algoritmos de aprendizado de máquina, que analisam dados continuamente comparando os padrões de amostras saudáveis e não saudáveis, permitindo traçar uma resposta em função da percepção de risco. Assim, o presente trabalho promove uma revisão de literatura de acordo com o método de Kitchenham e Charters (2007), buscando identificar quais algoritmos de aprendizado de máquina empregados, sua classificação quanto à abordagem, principais métricas de performance, bem como aspectos vantajosos e limitações de uso de algoritmos de aprendizado de máquina em IDSs. Busca fornecer, ainda, tendências e desafios dessas aplicações e, com isso, estruturar conhecimentos sobre algoritmos de aprendizado de máquina e sua aplicabilidade em sistemas de detecção de intrusão (IDS).

Palavras-chave: Sistema de detecção de intrusão, Aprendizado de máquina, Revisão Sistemática de Literatura.

ABSTRACT

The dissemination of digital services over the years has been increasingly comprehensive and in all sectors of our daily lives. However, with this evolution increasingly sophisticated ways of perpetrating cyberattacks in the digital environment emerge, often leading to inestimable losses. In order to identify potentially malicious activities, the literature introduces several intrusion detection systems (IDS), which report dangerous activity if network and system patterns are considered anomalous. These systems (IDSs) can use machine learning algorithms, which analyze data continuously by comparing the patterns of healthy and unhealthy samples, allowing a response to be drawn based on risk perception. Thus, the present work promotes a literature review according to the Kitchenham and Charters (2007) method, seeking to identify which machine learning algorithms are used, their classification regarding the approach, main performance metrics, as well as advantageous aspects and limitations of the use of machine learning algorithms in IDSs. It also seeks to provide trends and challenges of applications of this type, and, with this, structure knowledge about machine learning algorithms and their applicability in intrusion detection systems (IDS).

Keywords: Intrusion Detection System, Machine Learning, Systematic Literature Review.

Sumário

Lista de Figuras

Lista de Tabelas

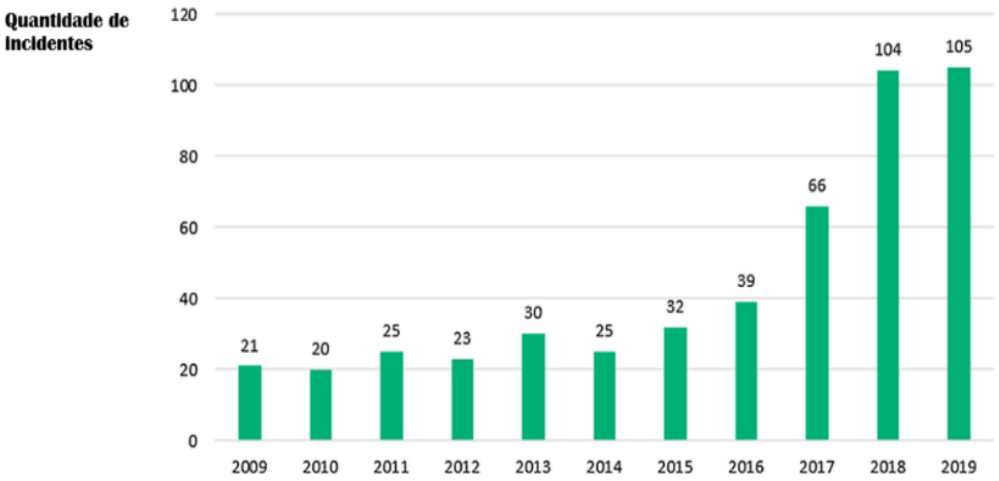
Tabela de Siglas

Sigla	Significado
IDS	Intrusion Detection System
ML	Machine Learning
TDA	Taxa de Detecção de Ataque
TAF	Taxa de Alarme Falso
NIDS	Network Intrusion Detection System
HIDS	Host Intrusion Detection System
KNN	K-nearest Neighbors
SVM	Support Vector Machine
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
FP	Falso Positivo
FN	Falso Negativo
LOF	Local Outlier Factor
TFP	Taxa de Falso Positivo
TVN	Taxa de Verdadeiro Negativo
TFN	Taxa de Falso Negativo
AUROC	Area Under the ROC Curve
PCA	Principal Component Analysis
AE	Autoencoder
VAE	Variational Autoencoder
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GMM	Gaussian Mixture Model
SOM	Self-Organizing Maps
DDoS	Distributed Denial of Service
DNS	Domain Name System
HTTPS	Hypertext Transfer Protocol Secure
APT	Ameaça Persistente Avançada
MQTT	Message Queuing Telemetry Transport
SDN	Software Defined Networking

1. Introdução

A evolução tecnológica, bem como a democratização dos meios digitais ocasionaram um grande crescimento na adoção de serviços e/ou produtos digitais nos últimos anos. Esse crescimento traz consigo diversas perspectivas positivas, como o aumento da capacidade de processamento e armazenamento. Entretanto, a evolução tecnológica também impõe grandes desafios no que tange a área de segurança. Meios cada vez mais sofisticados ameaçam a integridade, disponibilidade e confidencialidade dos dados e possibilitam prejuízos financeiros diversos, inclusive comprometendo a própria vida das pessoas [5]. Apesar de grandes esforços de representantes da área de segurança no contexto corporativo, como o uso de *Firewalls,* requisitos de segurança, sistemas de prevenção de ataques, autenticação em duplo fator e encriptação de dados, muitas organizações continuam sendo vítimas de ciberataques. A figura abaixo reforça que os prejuízos só crescem em função do tempo, gerando urgência para as entidades no combate a ciberataques. A figura considera a quantidade de incidentes acima de 1 milhão de dólares em milhões e, permite inferir, um crescimento contínuo ao longo dos anos.

Figura 1 - Evolução dos incidentes de ciberataques com perdas superiores a 1 milhão de dólares



Fonte: SIST [9]

Sistemas de detecção de intrusão (*Intrusion Detection System - IDS*) surgem nesse cenário como uma das diversas ferramentas existentes que protegem serviços e redes de ataques maliciosos. IDSs são softwares projetados para avaliar continuamente sistemas distribuídos diversos, que por meio da detecção de atividades anômalas, reportam essas ameaças para tratamento específico [6, 7]. Esses sistemas também fornecem dados relevantes para que futuras ocorrências sejam evitadas, garantindo ainda que tais ameaças se alastrem. Os IDSs se apresentam como uma ferramenta poderosa quando avaliam padrões de ataques previamente conhecidos em conjunto com os status atuais das redes e sistemas, gerando inclusive baixa taxa de falsos positivos. Porém, os IDSs apresentam limitações quando a abordagem busca prever ataques desconhecidos, uma vez que não estão atrelados a um padrão de assinatura conhecido e tais ataques podem se manifestar de várias formas. Essa dificuldade para detecção de ataques neste contexto gera consigo altas taxas de falsos positivos e reduz a performance das IDSs [4].

A fim de melhorar o desempenho das IDSs, várias abordagens foram estudadas na literatura, entre elas o uso de modelos de aprendizado de máquina. Segundo Awad e Khanna (2015), aprendizado de máquina (ML) é um ramo da inteligência artificial que aplica algoritmos sistematicamente para sintetizar as relações básicas entre dados e informações. Os sistemas de aprendizado de máquina podem ser treinados para categorizar as condições mutáveis de um processo e modelar variações no comportamento operacional. À medida que os objetos de conhecimento evoluem sob a influência de novas tendências e padrões, os sistemas de ML podem identificar interrupções nos modelos existentes, redesenhar e até mesmo retreinar-se para se adaptar e evoluir com o novo conhecimento [8]. O uso de modelos de ML nos sistemas de detecção à intrusão tem intenção, sobretudo, de aumentar as taxas de detecção de ataques (TDA*)* em consonância com a redução da taxa de falsos positivos *(*TFP). Esses modelos de aprendizado de máquina podem ainda ser categorizados em função da abordagem de aprendizado conforme quadro abaixo:

Tabela 1 - Comparação entre aprendizado supervisionado e não supervisionado

Abordagem	Conceito
Supervisionada	Quando os dados de treinamento compreendem exemplos de vetores de entrada junto com seus vetores de destino correspondentes (Bishop, C. M., 2006).
Não Supervisionada	Quando os dados de treinamento consistem em um conjunto de vetores de entrada sem quaisquer valores alvo correspondentes (Bishop, C. M., 2006).

Fonte: elaborado pelo autor

1.1. Motivação

O presente trabalho é caracterizado como uma revisão sistemática de literatura, ou seja, se propõe a analisar diversos textos científicos a fim de realizar sínteses de evidências e analisar o desempenho de soluções de aprendizado de máquina como ferramenta de apoio aos sistemas de detecção de intrusão. A realização da pesquisa se propõe a fornecer características, informações, tendências e limitações de uso através de um amplo levantamento de estudo da literatura e execução de etapas do método proposto por Kitchenham e Charters (2007). A evolução tecnológica, bem como na evolução de ciberataques, justifica a concentração de conhecimentos da área para municiar estudos futuros e permitir pontos de partida para novas soluções. O trabalho, ademais, possibilita o levantamento de possíveis lacunas da literatura e a redução de viés de estudos quando estes são analisados isoladamente. Assim, a revisão de literatura auxilia, sobretudo, no processo de tomada de decisões, proporcionando conclusões mais completas e confiáveis. Como este estudo visa abarcar o conhecimento acerca de sistemas de detecção de intrusão, percebe-se a necessidade de avaliar modelos com apoio de aprendizado de máquina, a fim de mensurar seu desempenho também frente a dois desafios constantemente discutidos na literatura: a robustez adversarial e o *concept drift*. A robustez adversarial refere-se à capacidade do modelo resistir a manipulações intencionais nos dados de entrada feitas por atacantes com o objetivo de enganar o sistema, tornando-se essencial para garantir que os IDS mantenham a sua eficácia mesmo diante de tentativas de evasão sutis, porém maliciosas [87]. Já o *concept drift* trata da mudança natural nos padrões de dados ao longo do tempo, exigindo que os modelos sejam adaptáveis e continuem relevantes em ambientes dinâmicos, onde novos tipos de ataques e comportamentos surgem constantemente [88].

Tais conceitos na análise reforçam a necessidade de compreender melhor os limites atuais das soluções de aprendizado de máquina, bem como orientar pesquisas futuras no sentido de desenvolver sistemas mais resilientes, adaptáveis e realistas para contextos de segurança cibernética.

1.2. Objetivos

Analisar a aplicação de algoritmos de aprendizado de máquina em sistemas de detecção de intrusão, visando identificar os tipos de algoritmos mais utilizados, as abordagens

predominantes (NIDS, HIDS ou híbridos), os principais conjuntos de dados usados como base e a relação entre o desempenho dos modelos e os algoritmos aplicados.

São objetivos específicos deste projeto:

- Identificar quais tipos de algoritmos de aprendizado de máquina são mais empregados;
- Detectar quais tipos de sistemas de detecção de intrusão são mais utilizados (NIDS, HIDS ou híbrido);
- Levantar quais *datasets* são mais utilizados;
- Identificar performance dos modelos em função dos algoritmos de aprendizado de máquina utilizados;
- Identificar quais tipos específicos de ataques/ameaças são mais frequentemente detectados;
- Levantar limitações e desafios de uso de algoritmos de aprendizado de máquina em sistemas de detecção a intrusão;

2. Conceitos

Neste capítulo, são introduzidos alguns conceitos mais aprofundados a fim de subsidiar conhecimentos do ramo.

2.1.Aprendizado de máquina

Entende-se por aprendizado de máquina o ramo da inteligência artificial voltado ao desenvolvimento de técnicas computacionais que promovem o autoaperfeiçoamento por meio da identificação de padrões nos dados. Seu uso é especialmente indicado em cenários onde o problema exige a escolha, entre inúmeras hipóteses, daquela que melhor se ajusta aos dados observados e às classificações previamente estruturadas pelo idealizador do modelo [11]. Nesse contexto, alguns conceitos fundamentais são indispensáveis para a compreensão aprofundada dos mecanismos que sustentam os modelos de aprendizado de máquina.

Um dos principais elementos de algoritmos de aprendizado de máquina é o **classificador**, que corresponde ao método responsável por atribuir uma categoria ou classe a uma nova instância de entrada sem rótulo. Por meio de inferência estatística, o classificador determina a saída mais adequada com base em padrões aprendidos previamente, sendo comum o uso de técnicas como *K-Nearest Neighbors* (KNN), árvores de decisão e *Support Vector Machines* (SVM) [8]. Associado ao desempenho do classificador, o **custo** é uma métrica que avalia os desvios entre os valores previstos pelo modelo e os valores reais, funcionando como parâmetro para quantificar a eficácia do processo preditivo [8]. Para que esse processo ocorra, é necessário o uso de um **conjunto de dados**, ou seja, uma coleção de dados brutos que serve de base para o treinamento e validação do modelo [8]. O coração do aprendizado de máquina reside no algoritmo de indução, que infere saídas com base nos dados fornecidos durante a fase de treinamento. Essa capacidade de generalização é justamente o que permite ao modelo identificar padrões ou realizar previsões a partir de exemplos [8]. A estrutura resultante desse processo é o **modelo**, que sintetiza os dados com o objetivo de realizar descrições ou previsões. Cada modelo é ajustado para atender às especificidades de uma aplicação, podendo possuir um conjunto finito ou infinito de parâmetros, conforme a natureza do problema [8]. Outro conceito recorrente em algoritmos de aprendizado de máquina que possibilita a metrificação e avaliação da performance do modelo é a **matriz de erro**, que permite analisar sua capacidade de classificar corretamente as instâncias com base em categorias como falso positivo, verdadeiro positivo, falso negativo e verdadeiro negativo. Kohavi e Provost (1998) traz a ilustração abaixo, onde classifica os

dados atuais e previstos do modelo. A partir dessas classificações é possível testar a performance do modelo através das relações entre essas variáveis a fim de avaliar acurácia, precisão, recall, etc. (relações aprofundadas na seção 2.3 deste trabalho).

Figura 2 - Estrutura da matriz de erro aplicada em modelos de classificação

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Kohavi e Provost (1998)

Embora essas definições se apliquem a uma ampla gama de contextos, os desafios se intensificam quando o objetivo é detectar intrusões em sistemas de segurança cibernética.

2.1.1. Agrupamento

Segundo Awad e Khanna (2015), agrupamento é uma técnica que levanta agrupamentos de uma coleção de dados baseado nas similaridades de seus atributos e/ou características. Agrupamento é um tipo de algoritmo interativo que se baseia na ideia de prova e erro que opera no âmbito da similaridade (ou dissimilaridade em contextos específicos) [8]. Normalmente, um algoritmo de agrupamento para quando um critério de término é atingido. O desafio na implementação deste tipo de algoritmo reside na identificação de uma função que correlacionem dois itens como um valor numérico [8].

2.1.2. Redução de dimensionalidade

Redução da dimensionalidade pode ser descrita como uma técnica de transformação de um espaço de alta dimensão em um de baixa dimensão, assegurando preservação máxima da topologia do espaço de dados original [13]. Este processo de sintetização das características (ou dimensões) permite simplificação dos modelos, redução de ruídos dos dados e melhor visualização. Com a evolução cada vez maior do volume e processamento dos

dados, esta técnica tem recebido atenção especial de estudos, sobretudo aqueles que possuem dimensões muito complexas.

2.1.3. Detecção de Anomalias

Detecção de anomalias é uma técnica essencial em problemas onde o objetivo é identificar padrões que divergem significativamente do comportamento esperado. Segundo Chandola, Banerjee e Kumar (2009), anomalias são observações que não seguem o padrão geral dos dados e, por isso, podem representar eventos raros ou críticos, como fraudes, falhas em sistemas ou atividades suspeitas em redes. Essa abordagem é especialmente útil quando há predominância de dados normais e escassez de exemplos rotulados como anômalos. Algoritmos como *Isolation Forest*, *Local Outlier Factor* (LOF) e *One-Class* SVM são frequentemente empregados, pois não requerem dados previamente rotulados para realizar a identificação das anomalias. Conforme Schölkopf et al. (2021), algoritmos baseados em isolamento, como o *Isolation Forest,* têm se mostrado eficazes em detectar anomalias em grandes volumes de dados com baixo custo computacional. Com o aumento de dados em tempo real e sistemas automatizados, o campo de detecção de anomalias continua sendo amplamente explorado em áreas como segurança cibernética, Internet das Coisas (IoT) e monitoramento industrial (Goldstein & Uchida, 2016).

2.1.4. Regras de Associação

As regras de associação são técnicas utilizadas para encontrar relacionamentos frequentes entre variáveis em grandes conjuntos de dados. De acordo com Tan, Steinbach e Kumar (2018), o processo de descoberta de associações envolve a identificação de itens ou eventos que ocorrem frequentemente em conjunto, a partir dos quais são geradas regras de associação baseadas em métricas como suporte, confiança e *lift*. Algoritmos clássicos como *Apriori*, *Eclat* e *FP-Growth* são amplamente utilizados nesse contexto. Estudos recentes (Liu et al., 2020) apontam que técnicas baseadas em regras de associação têm sido adaptadas para ambientes de dados em fluxo contínuo, tornando-as relevantes em aplicações modernas como motores de recomendação, análise de comportamento de usuários e sistemas de recomendação de conteúdo. A simplicidade e interpretabilidade dessas regras tornam essa abordagem valiosa em contextos de apoio à decisão, especialmente em sistemas comerciais, financeiros e de marketing (Agrawal et al., 2021).

2.2. Performance do modelo

Uma das questões que o atual trabalho pretende identificar é como a performance de um modelo pode ser medida. Para problemas de classificação, o resultado da classificação pode assumir um valor correto ou incorreto [13]. Conforme apresentado no tópico 2.1.7, as condições de uma matriz de erro podem ser as seguintes:

1) Verdadeiro positivo (VP): os ataques reais são classificados como ataques;
2) Verdadeiro negativo (VN): os dados atuais normais são classificados como normal;
3) Falso positivo (FP): os dados atuais normais são classificados como ataque;
4) Falso negativo (FN): ataques reais são classificados como dados normais;

Baseado em como essas variáveis se relacionam, é possível traçar meios de metrificar a aderência do modelo quanto a sua performance. As métricas mais utilizadas incluem **Precisão**, **Acurácia**, **Recall**, **F1-Score**, **Taxa de Falsos Positivos (TFP), Taxa de Falsos Negativos (TFN), Taxa de Verdadeiros Negativos (TVN) e Área Sob a Curva ROC (AUC/AUROC)**, métricas essas utilizadas nesse trabalho para realizar os comparativos entre os artigos utilizados para análise.

- Precisão - mede a proporção de alertas de intrusão corretamente classificados em relação ao total de alertas gerados pelo sistema. Um valor alto indica que poucos eventos legítimos são erroneamente classificados como ataques (falsos positivos), sendo crucial para evitar sobrecarga operacional.
- Acurácia - representa a taxa global de classificações corretas (tanto ataques quanto tráfego normal) em relação ao total de observações. Embora útil, pode ser enganosa em conjuntos de dados desbalanceados, onde a classe majoritária (normal) domina a métrica.
- *Recall* - quantifica a capacidade do sistema de detectar ataques reais, calculando a proporção de intrusões identificadas corretamente em relação ao total de intrusões existentes. Um recall alto é essencial para minimizar riscos de segurança.
- *F1-Score* - harmoniza precisão e recall por meio da média harmônica, sendo ideal para cenários com desbalanceamento de classes. Valores próximos a 1 indicam um equilíbrio entre a redução de falsos positivos e a detecção de ataques.

- Taxa de Falso Positivo (TFP) - mede a proporção de tráfego normal incorretamente classificado como malicioso. Em ambientes críticos, uma TFP elevada pode levar a alertas desnecessários e desgaste operacional.
- Taxa de Falso Negativo (TFN) - reflete a incapacidade de detectar ataques reais, representando riscos diretos à segurança. Sistemas com TFN alta são considerados ineficazes, pois permitem que intrusões passem despercebidas.
- Taxa de Verdadeiro Negativo (TVN) indica a eficiência em classificar corretamente o tráfego normal. Em IDS, uma TVN alta complementa a redução de falsos positivos.
- AUC/AUROC - avalia o desempenho geral do modelo independentemente do limiar de classificação. Valores próximos a 1 indicam excelente distinção entre classes (ataque vs. normal), enquanto valores próximos a 0.5 sugerem desempenho aleatório.

2.3. Sistema de detecção de intrusão

Sistema de detecção de intrusão se apresenta como uma ferramenta proativa para detectar e classificar intrusões, ataques e/ou violações da política de segurança automaticamente em uma camada de rede ou hospedeiro [10]. Basicamente funciona como um instrumento de análise que detecta possível atividade maliciosa que ultrapassa os limites de segurança. Esse monitoramento pode ser apoiado por usos de tecnologia de aprendizado de máquina e se desdobra em dois tipos: HIDS (*host intrusion detection system)* e NIDS (*network intrusion detection system).*

2.3.1 IDS de Host (HIDS)

Os IDSs baseado em *host* são sistemas que monitoram os computadores hospedeiros locais através de arquivos de logs. Os dados desses arquivos de logs são coletados através de sensores locais. Esses logs podem ser de vários tipos, entre eles: de software, de sistema, de recurso de disco, de informações de contas de usuários, entre outros [10].

2.3.2. IDS de Rede (NIDS)

Os IDSs baseado em rede são sistemas que monitoram o tráfego de dados via comportamento em rede [10]. Para coleta destes dados, normalmente se utilizam equipamentos de espelhamento de dispositivos de rede como *switches*, roteadores e grampo de rede. Esta coleta visa analisar essencialmente ataques e possíveis ameaças que possam surgir através da troca de pacotes de redes [10].

3. Trabalhos Relacionados

Esta seção será responsável por realizar um apanhado de estudos relevantes sobre o uso de aprendizado de máquina em sistemas de detecção de intrusão, demonstrando o que os estudos mais recentes abordam sobre o tema, quais formas e tendências de uso tem sido percebido e quais desafios requerem atenção para resolução futura.

De acordo com Yu Liu et al [15], existe uma relação inversamente proporcional quanto a taxa de acurácia e a abordagem de aprendizado não supervisionada. Como os algoritmos de aprendizado supervisionado possuem uma saída previsível, estes normalmente apresentam performance superior à abordagem não supervisionada. A performance da abordagem não supervisionada normalmente apresenta índices mais altos de falsos positivos e menor acurácia do modelo, posto que a compreensão, agrupamento e classificação dos dados não rotulados é mais complexa e tem como consequência uma saída imprevisível.

Por outro lado, a abordagem não supervisionada é mais indicada em contextos cuja dimensão seja complexa e dinâmica, pois permite conhecer ataques e/ou vulnerabilidades desconhecidas e que não esteja atrelada a um padrão de assinatura conhecido [16]. Na medida que os sistemas evoluem, os esforços para traçar completamente todas as vulnerabilidades que um sistema está exposto é contraproducente, logo, o uso de modelos de ML não supervisionados podem ser aplicados de forma contínua para identificar novas formas de se perpetrar ataques [16].

Nassif et al [17] elaboraram uma revisão sistemática para detecção de anomalias utilizando aprendizado de máquina. Esta revisão se pauta em 4 perspectivas: as aplicações de detecção de anomalia, técnicas de aprendizado de máquina, métricas de performance para modelos de aprendizado de máquina e classificação da detecção de anomalia. Halbouni et al [18] também promovem uma revisão sistemática, porém, estimulando o debate acerca dos tipos de algoritmos de ML e DL são usados para proteger dados de comportamento malicioso, além de estabelecer amplo debate sobre as formas de implementação, aplicabilidade e uso e as limitações do conjunto de dados. Segundo Halbouni et al [18], existe uma tendência no uso de abordagens semi-supervisionadas, ou seja, com características supervisionadas e não supervisionadas. Tal uso permite reduzir ruídos e vieses das duas abordagens de aprendizado.

Destacam-se também limitações e desafios levantados por autores diversos quanto ao uso de modelos de ML. Segundo Mirkovic e Reiher [19], ruídos de dados (ou *outliers*) podem levar a erros de detecção quando são identificados como ameaça e fazem parte do comportamento normal do modelo. Sharif et al [20] discorre que a interpretação dos dados

também é um elemento a ser considerado na concepção de um modelo de ML, pois a classificação errônea é uma das características que reduzem a performance do modelo. Debar [21], por sua vez, comenta que por mais que o processo de redução da dimensionalidade seja um processo favorável para simplificação do modelo de ML, este processo de generalização também leva à perda de performance se utilizado em diferentes ambientes, logo, o autor traz consigo a preocupação da generalização de dados em ambientes diversos.

4. Metodologia

O presente trabalho se apoia na estruturação de uma revisão sistemática de literatura preconizada por Kitchenham e Charters [22]. Este trabalho propõe realizar uma análise crítica da literatura científica relacionada ao uso de algoritmos de aprendizado de máquina em sistemas de detecção de intrusões (IDS). O foco principal é identificar e sistematizar quais algoritmos têm sido mais empregados, os contextos de aplicação, as bases de dados utilizadas e, principalmente, os critérios de avaliação de desempenho utilizados nos estudos. A comparação entre os estudos visa compreender a efetividade e os limites dos algoritmos analisados, bem como apontar tendências e lacunas na literatura atual. Isto posto, esse trabalho utilizará como base o método descrito por Kitchenham e Charters inclui as seguintes etapas: planejamento, seleção de estudos, coleta de dados relevantes e análise e investigação. Os próximos tópicos detalham melhor as etapas da metodologia adotada.

4.1 Perguntas da pesquisa

A revisão sistemática de literatura visa organizar e examinar estudos na literatura que utilizam aprendizado de máquina na abordagem de aprendizado como ferramenta de apoio aos sistemas de detecção de intrusão (IDS). A fim de orientar as análises sobre os resultados coletados, as seguintes perguntas são descritas para embasar a intenção do estudo:

1. Quais tipos de algoritmos de aprendizado de máquina são mais empregados nos modelos?
2. Quais abordagens de sistemas de detecção de intrusão são mais utilizados (Sistemas de detecção de intrusão de rede, de *host* ou híbrido)?
3. Quais as taxas de precisão, acurácia, sensibilidade, *F1-Score*, Taxa de Falsos Positivos, *AUC/AUROC*, Taxa de Falsos Negativos, Taxa de Verdadeiros Negativos dos estudos levantados?
4. Quais tipos específicos de ataques/ameaças são mais frequentemente detectados em modelos de aprendizado de máquina?
5. Quais principais conjunto de dados utilizados?
6. Quais as principais limitações/desafios encontrados na literatura?

4.2 Extração e síntese dos dados

Para cada estudo selecionado, serão extraídas as seguintes informações principais:

- Título do artigo
- Autores e ano de publicação
- Algoritmos de aprendizado de máquina
- Métricas de performance
- Conjunto de dados utilizados
- Principais conclusões do estudo
- Limitações e desafios identificados no estudo

Esses dados serão organizados em tabelas comparativas, facilitando a visualização dos aspectos analisados. Além disso, será realizada uma análise qualitativa dos conteúdos, discutindo padrões recorrentes, lacunas na literatura e tendências em relação ao uso de algoritmos de modelos de aprendizado de máquina em sistemas de detecção de intrusão.

4.3 Protocolo de revisão

Após a etapa da formulação de perguntas da pesquisa foi possível traçar o protocolo de revisão. Tal protocolo serve como guia para definição de alguns critérios básicos para apoiar a busca dos artigos, como a definição das bases de dados, palavras-chaves e *string* de busca.

O estudo usou como base os repositórios descritos na imagem abaixo devido a sua relevância e qualidade acadêmica:

Tabela 2 - Bases de dados científicas utilizadas na pesquisa

Nome
Scopus
IEEE
ACM Digital Library
Web of Science
ScienceDirect

Fonte: elaborado pelo autor

Segundo as recomendações de Kitchenham e Charters (2007), foram realizados testes pilotos para simular os resultados das buscas. A partir de tais testes, a listagem de grupos de categorias e suas palavras-chave conforme tabela abaixo:

Tabela 3 - Palavras-chave organizadas por categoria temática

Categoria	Palavras-chave
Tema principal	intrusion detection

Tecnologia aplicada	machine learning
Tipo de aprendizagem	unsupervised learning
Técnicas e algoritmos	anomaly detection, clustering, one-class, autoencoder, isolation forest

Fonte: elaborado pelo autor

Por meio da definição das palavras-chave foi possível montar a *string* de busca que irá permear as buscas iniciais levando em consideração correspondência no título, resumo e palavras-chave. A *string* também possui condicionais booleanos (AND e OR) para melhor abrangência e assertividade e é descrita na tabela abaixo:

Tabela 4 - Estratégia de busca estruturada com operadores booleanos

***String* de busca**
("intrusion detection") AND ("machine learning") AND ("unsupervised learning") AND ("anomaly detection" OR "clustering" OR "one-class" OR "autoencoder" OR "isolation forest")

Fonte: elaborado pelo autor

4.4 Seleção de estudos

A partir dos critérios estabelecidos na etapa de planejamento foram realizadas buscas nos repositórios citados em maio de 2025. Como a quantidade do retorno das buscas foi consideravelmente grande, foram estabelecidos critérios de seleção dos estudos conforme proposto pela metodologia de Kitchenham e Charters (2007).

4.4.1 Estudos publicados entre 2020 e 2025

A área de detecção de intrusão baseada em aprendizado de máquina é altamente dinâmica e evolui rapidamente. Restringir a busca a estudos publicados nos últimos 5 anos garante que os dados, técnicas e conclusões refletem as tecnologias atuais, algoritmos recentes e ameaças modernas, mantendo a relevância e a atualidade da revisão.

4.4.2 Estudos primários

Estudos primários são aqueles que apresentam dados originais e resultados experimentais. A escolha por esse tipo de estudo assegura que a revisão seja baseada em evidências empíricas diretas, evitando duplicação de dados e fornecendo uma análise fundamentada em resultados concretos de experimentação.

4.4.3 Estudos relacionados com o tema

Selecionar apenas estudos diretamente relacionados ao tema sistemas de detecção de intrusão assegura o foco e a relevância científica da revisão, evitando incluir publicações que não contribuam efetivamente para responder à pergunta de pesquisa ou aos objetivos do trabalho.

4.4.4 Estudos com acesso pelo portal capes

A seleção de estudos para esta revisão sistemática foi limitada àqueles disponíveis por meio do Portal de Periódicos da CAPES, plataforma mantida pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Este critério foi adotado devido à disponibilidade de acesso institucional oferecido pela Universidade Federal de Pernambuco (UFPE), que mantém parceria ativa com o portal e suas bases indexadoras. Através da autenticação via CAFe (Comunidade Acadêmica Federada), foi possível consultar e recuperar artigos completos nas bases de dados escolhidas na tabela 2.

4.4.5 Estudos sem duplicações entre as bases

A remoção de duplicatas entre diferentes bases de dados evita a contagem repetida de evidências, que poderia viciar os resultados da revisão. Esse critério assegura a precisão da análise quantitativa e qualitativa, garantindo que cada estudo contribua de forma única para os achados da pesquisa.

4.4.6 Estudos publicados em periódicos

Estudos publicados em periódicos passam por revisão por pares, o que assegura maior rigor metodológico, validade dos dados e confiabilidade nos resultados. Esse filtro também ajuda a garantir que os trabalhos incluídos apresentem discussões aprofundadas e fundamentadas, contribuindo para maior rigor de qualidade no aprofundamento da revisão sistemática.

4.5 Critérios de avaliação

Após seleção dos artigos remanescentes baseados nos critérios de seleção supracitados, foi realizada a leitura das seções **resumo, introdução e conclusão**. Cada estudo foi avaliado quanto aos critérios definidos abaixo e representados em 3 escalas, a saber: não conforme (nota 0), parcialmente conforme (nota 0,5) e conforme (nota 1,0).

Tabela 5 - Critérios utilizados para avaliação da qualidade dos estudos selecionados

Critérios de avaliação
Qualidade do *dataset*
Se reportou taxas de performance
Citação direta de algoritmos clássicos de ML
Reprodutibilidade dos dados

Fonte: elaborado pelo autor

A soma das notas de cada critério definido para o estudo representa o indicador final da etapa de avaliação. Para este trabalho, foram considerados os estudos cuja nota fosse **igual ou superior a 2,0.**

4.5.1 Critérios de Avaliação dos Estudos Incluídos

Para garantir a relevância, qualidade e aplicabilidade dos estudos selecionados na revisão sistemática, foram definidos quatro critérios principais de avaliação. Cada artigo foi analisado à luz desses critérios, permitindo uma comparação padronizada entre os trabalhos e uma identificação mais clara das contribuições científicas no contexto de detecção de intrusão utilizando técnicas de aprendizado de máquina.

4.5.2. Base de Dados Utilizada

Foi verificado se o estudo utilizou conjuntos de dados amplamente reconhecidos na área de segurança da informação e detecção de intrusão. Exemplos típicos incluem o NSL-KDD, CICIDS2017 e UNSW-NB15. O uso dessas bases é essencial para garantir a comparabilidade dos resultados e a validação adequada dos modelos propostos. Estudos que utilizam dados proprietários ou não acessíveis ao público foram considerados com menor grau de reprodutibilidade e aplicabilidade prática.

4.5.3. Métricas de Avaliação Relatadas

Considerou-se a robustez da avaliação dos modelos por meio das métricas apresentadas. Estudos que relataram métricas como F1-score, AUC-ROC (Área sob a curva ROC) e taxa de falsos positivos (TFP) foram valorizados, pois essas medidas são amplamente aceitas para avaliar a eficácia de modelos de classificação, especialmente em contextos de segurança cibernética onde falsos positivos podem comprometer a usabilidade do sistema. A ausência dessas métricas ou a utilização exclusiva de acurácia, por exemplo, foi considerada uma limitação metodológica.

4.5.4. Comparação com Baselines Clássicos

Foi analisado se o estudo realizou comparações com métodos de referência, tais como *Isolation Forest*, *PCA (Principal Component Analysis)*, ou *One-Class SVM*. A presença de tais comparações é fundamental para contextualizar o desempenho do modelo proposto frente a técnicas já consolidadas, permitindo uma avaliação crítica de sua efetividade e uma mensuração objetiva do avanço científico.

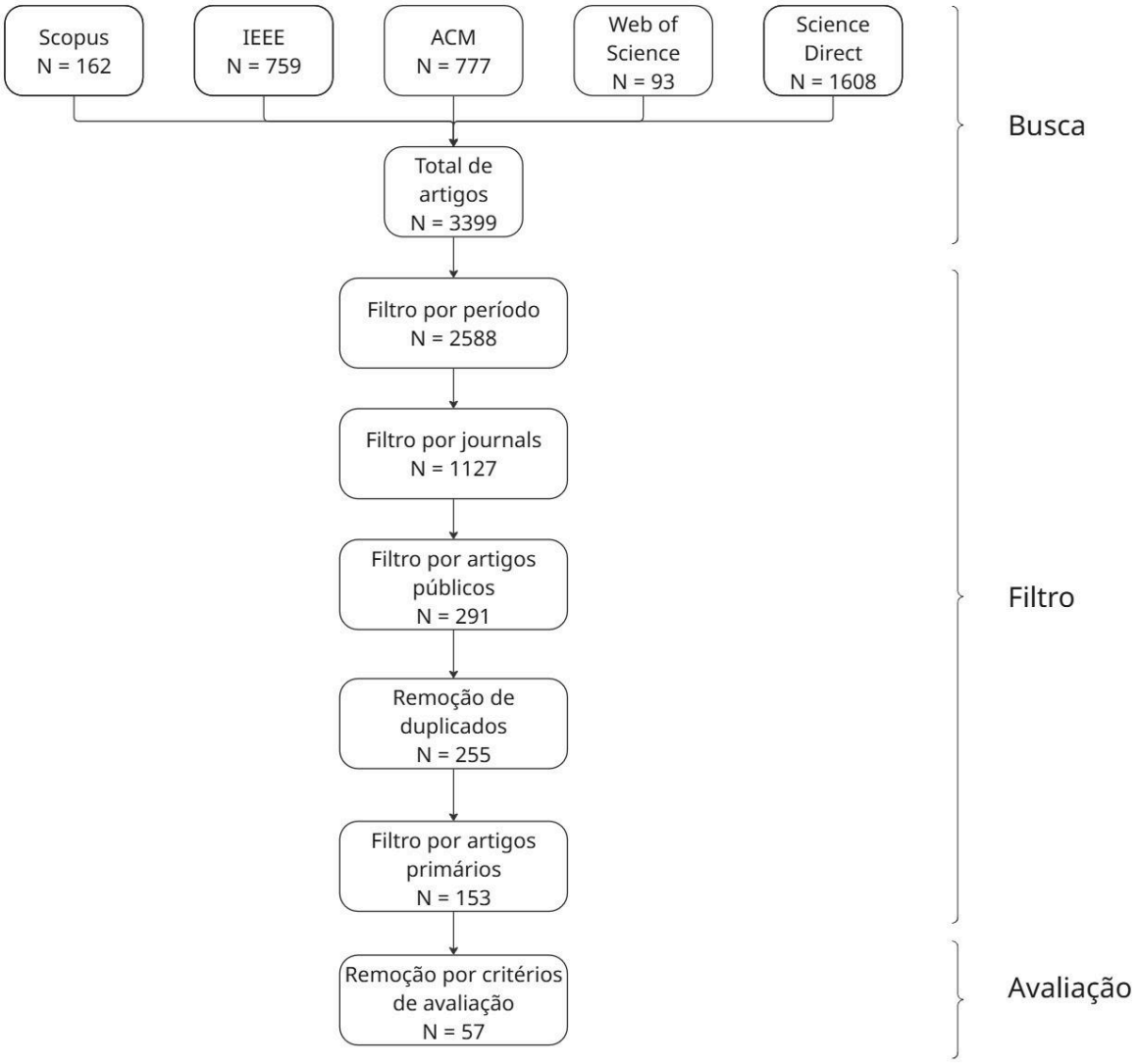
4.5.5. Reprodutibilidade do Estudo

A transparência e a possibilidade de replicação dos resultados são pilares da pesquisa científica. Assim, foi avaliado se o artigo fornece acesso ao código-fonte, scripts de treinamento/teste ou links para repositórios públicos (ex.: GitHub, Kaggle). Também foi considerada a disponibilização dos dados utilizados no experimento. A reprodutibilidade é especialmente importante em pesquisas de aprendizado de máquina, pois pequenos ajustes nos dados ou nos parâmetros podem levar a variações significativas nos resultados.

4.6 Detalhamento da revisão sistemática de literatura

Após realização de todos os mecanismos de busca e filtro, bem como da aplicação dos critérios de seleção e a avaliação, foram selecionados 57 artigos finais para análise. A ilustração abaixo fornece detalhamento do quantitativo de estudos selecionados após cada etapa do estágio de planejamento da revisão sistemática de literatura:

Figura 3 - Fluxograma do protocolo da revisão sistemática de literatura



Fonte: elaborado pelo autor

Após todos os critérios de filtro e avaliação, a relação final de artigos que serão trabalhados neste estudo estão discriminados na tabela abaixo:

Tabela 6 - Lista dos estudos incluídos na revisão sistemática com seus identificadores

Identificador	Título	Ano
A1	A Comparative Analysis of Supervised and Unsupervised Models for Detecting Attacks on the Intrusion Detection Systems	2023
A2	A convolutional autoencoder architecture for robust network intrusion detection in embedded systems	2024

A3	A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework	2023
A4	A dual-tier adaptive one-class classification IDS for emerging cyberthreats	2025
A5	A framework for detecting zero-day exploits in network flows	2024
A6	A hybrid unsupervised clustering-based anomaly detection method	2021
A7	A novel bi-anomaly-based intrusion detection system approach for industry 4.0	2023
A8	A privacy-preserving Self-Supervised Learning-based intrusion detection system for 5G-V2X networks	2025
A9	A semi-supervised soft prototype-based autonomous fuzzy ensemble system for network intrusion detection	2025
A10	A sequential deep learning framework for a robust and resilient network intrusion detection system	2024
A11	A stacking ensemble of deep learning models for IoT intrusion detection	2023
A12	AEC_GAN: Unbalanced Data Processing Decision-Making in Network Attacks Based on ACGAN and Machine Learning	2023
A13	An Adaptive Intrusion Detection System for Evolving IoT Threats: An Autoencoder-FNN Fusion	2025
A14	An Attack-Resilient Architecture for the Internet of Things	2020
A15	An Effective Anomaly Detection Approach based on Hybrid Unsupervised Learning Technologies in NIDS	2024
A16	An Ensemble Learning Based Intrusion Detection Model for Industrial IoT Security	2023
A17	An Investigation of Learning Model Technologies for Network Traffic Classification Design in Cyber Security Exercises	2023
A18	An unsupervised approach for the detection of zero-day distributed denial of service attacks in Internet of Things networks	2024
A19	Anomaly Detection in Network Traffic Using Advanced Machine Learning Techniquxes	2025
A20	Automatic decision tree-based NIDPS ruleset generation for DoS/DDoS attacks	2024
A21	BIDS: An efficient Intrusion Detection System for in-vehicle networks using a two-stage Binarised Neural Network on low-cost FPGA	2024
A22	can-train-and-test: A curated CAN dataset for automotive intrusion detection	2024
A23	Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree	2023
A24	CPS-GUARD: Intrusion detection for cyber-physical systems and IoT devices using outlier-aware deep autoencoders	2023
A25	Detection of advanced persistent threat: A genetic programming approach	2024
A26	Enabling semi-supervised learning in intrusion detection systems	2025
A27	Evaluating ML-based anomaly detection across datasets of varied integrity: A case study	2024
A28	Fast anomaly detection with locality-sensitive hashing and hyperparameter autotuning	2022

A29	FlowGANAnomaly: Flow-Based Anomaly Network Intrusion Detection with Adversarial Learning	2024
A30	FSDC: Flow Samples and Dimensions Compression for Efficient Detection of DNS-over-HTTPS Tunnels	2024
A31	Generalizing intrusion detection for heterogeneous networks: A stacke d-unsupervise d fe derate d learning approach	2023
A32	Graph-embedded subspace support vector data description	2023
A33	Histogram-based network traffic representation for anomaly detection through PCA	2025
A34	Improving Performance of Autoencoder-Based Network Anomaly Detection on NSL-KDD Dataset	2021
A35	Intrusion Detection Based on Autoencoder and Isolation Forest in Fog Computing	2020
A36	Intrusion Detection Based on Sequential Information Preserving Log Embedding Methods and Anomaly Detection Algorithms	2021
A37	Intrusion Detection in IoT Systems Using Denoising Autoencoder	2024
A38	Lurking in the shadows: Unsupervised decoding of beaconing communication for enhanced cyber threat hunting	2025
A39	Malware classification for the cloud via semi-supervised transfer learning	2020
A40	Minority Resampling Boosted Unsupervised Learning With Hyperdimensional Computing for Threat Detection at the Edge of Internet of Things	2021
A41	Network anomaly detection methods in IoT environments via deep learning: A Fair comparison of performance and robustness	2023
A42	Network traffic classification based on federated semi-supervised learning	2024
A43	Neural AutoForensics: Comparing Neural Sample Search and Neural Architecture Search for malware detection and forensics	2022
A44	OIL-AD: An anomaly detection framework for decision-making sequences	2025
A45	On the improvement of the isolation forest algorithm for outlier detection with streaming data	2021
A46	Real-Time Anomaly Detection in Edge Streams	2022
A47	Res-TranBiLSTM: An intelligent approach for intrusion detection in the Internet of Things	2023
A48	SC-MLIDS: Fusion-based Machine Learning Framework for Intrusion Detection in Wireless Sensor Networks	2025
A49	Simulation of multi-stage attack and defense mechanisms in smart grids	2025
A50	STFT-TCAN: A TCN-attention based multivariate time series anomaly detection architecture with time-frequency analysis for cyber-industrial systems	2024
A51	SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security	2024
A52	The good, the bad, and the algorithm: The impact of generative AI on cybersecurity	2025
A53	The rise of machine learning for detection and classification of malware: Research developments, trends and challenges	2020

A54	Towards a Secure Internet of Things: A Comprehensive Study of Second Line Defense Mechanisms	2020
A55	Towards an Optimized Ensemble Feature Selection for DDoS Detection Using Both Supervised and Unsupervised Method	2022
A56	Unsupervised Algorithms to Detect Zero-Day Attacks: Strategy and Application	2021
A57	Which algorithm can detect unknown attacks? Comparison of supervised, unsupervised and meta-learning algorithms for intrusion detection	2023

Fonte: elaborado pelo autor

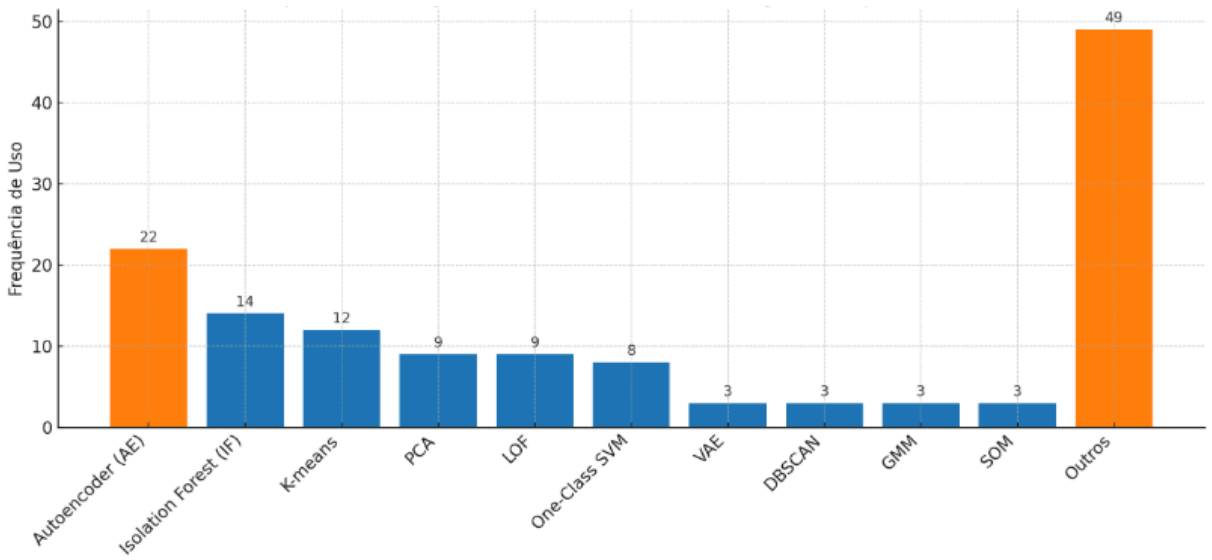
5. Resultados

Nesta seção, são apresentados os resultados quantitativos e qualitativos da revisão sistemática, com foco nos algoritmos de aprendizado de máquina mais empregados, sua frequência de uso e aplicações em sistemas de detecção de intrusão. A análise revela as tendências predominantes, como a adoção de *Autoencoders* e *Isolation Forest*, bem como métodos complementares, como *K-means e PCA*, destacando suas vantagens, limitações e cenários de uso. Além disso, discute-se a relação entre a escolha do algoritmo e a eficácia na detecção de anomalias, oferecendo insights para pesquisas futuras nessa área em constante evolução.

5.1. Quais tipos de algoritmos de aprendizado de máquina são mais empregados nos modelos?

Com o objetivo de identificar os algoritmos mais recorrentes em sistemas de detecção de intrusão, foi realizada uma análise quantitativa da frequência de uso dos métodos relatados nos estudos incluídos nesta revisão sistemática. Os resultados apontam uma clara preferência por algumas técnicas específicas, conforme apresentado na figura abaixo:

Figura 4 - Frequência de utilização dos algoritmos de aprendizado de máquina nos estudos analisados



Fonte: elaborado pelo autor

O *Autoencoder (AE)* destaca-se como o algoritmo mais utilizado, presente em 22 estudos (16,3%). Essa predominância pode ser explicada pela capacidade dos *autoencoders* em aprender representações compactas dos dados e identificar padrões de reconstrução anômalos, o que os torna altamente eficazes para tarefas de detecção de anomalias. Na sequência, o *Isolation Forest (IF)* aparece em 14 trabalhos (10,4%), sendo valorizado por sua eficiência em ambientes de alta dimensionalidade e sua abordagem baseada em partições aleatórias que isolam pontos discrepantes com facilidade. Algoritmos de clusterização também figuram com destaque, especialmente o *K-means*, identificado em 12 estudos (8,9%). Apesar de sua simplicidade, o *K-means* é amplamente utilizado como referência ou combinado com outras técnicas para identificação de agrupamentos anômalos. Métodos tradicionais de redução de dimensionalidade e detecção de *outliers*, como o *Principal Component Analysis (PCA)* e o *Local Outlier Factor (LOF)*, aparecem com a mesma frequência (9 estudos, 6,7% cada), evidenciando seu papel na pré-processamento e na identificação de instâncias atípicas em grandes volumes de dados. O *One-Class SVM*, outro algoritmo clássico para detecção de anomalias, foi identificado em 8 estudos (5,9%), sendo frequentemente usado em cenários com dados altamente desbalanceados ou quando apenas exemplos de comportamento normal estão disponíveis. Outros algoritmos relevantes incluem *Variational Autoencoder (VAE), DBSCAN, Gaussian Mixture Model (GMM) e Self-Organizing Maps (SOM)*, cada um com 3 ocorrências (2,2%), representando abordagens mais especializadas ou híbridas.

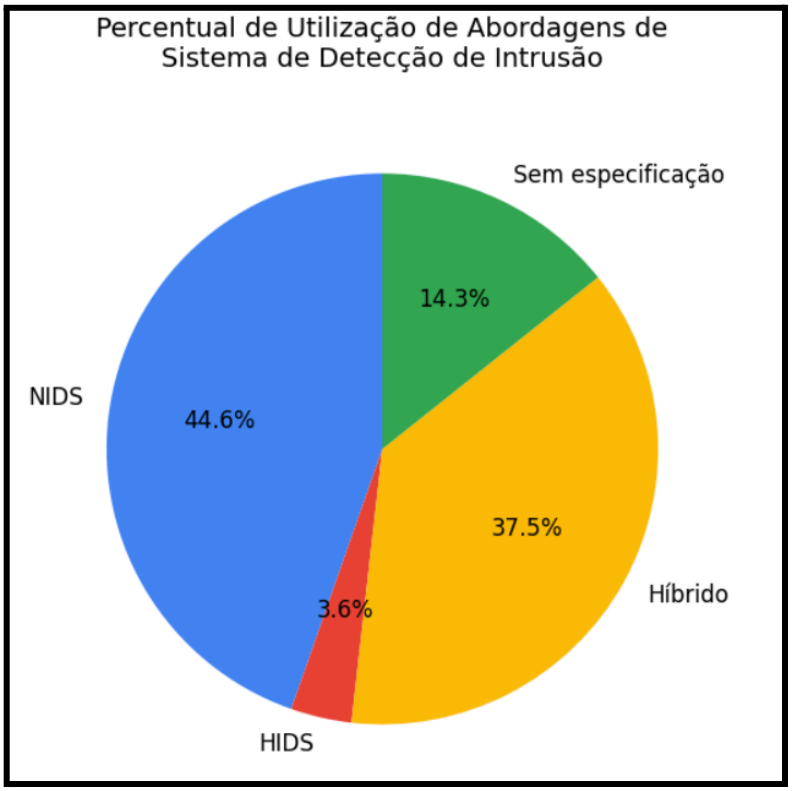
Por fim, foi registrada uma diversidade considerável de 36 outros algoritmos individuais, totalizando 49 menções (36,3%), o que reforça a natureza exploratória da área e o interesse da comunidade científica em testar novas abordagens para o problema. Essa distribuição demonstra que, embora existam técnicas amplamente consolidadas, a detecção de intrusão continua sendo um campo aberto a experimentações, onde a combinação de diferentes métodos pode trazer avanços significativos.

5.2. Quais abordagens de sistemas de detecção de intrusão são mais utilizados (NIDS, HIDS ou híbrido)?

A análise dos estudos incluídos na revisão sistemática permitiu identificar três principais abordagens de sistemas de detecção de intrusão utilizados em conjunto com algoritmos de aprendizado de máquina: Sistemas de detecção de intrusão de redes (NIDS), Sistemas de detecção de intrusão de *host* (HIDS) e abordagens híbridas. Além disso, uma

parte dos estudos não especificou claramente o tipo de sistema empregado. Conforme o gráfico abaixo, a abordagem mais recorrente foi a NIDS, presente em 44,64% dos estudos. Esse tipo de sistema opera a partir da análise de tráfego em rede, sendo amplamente adotado por sua capacidade de monitorar múltiplos dispositivos e identificar ataques distribuídos. Sua popularidade também se deve à disponibilidade de bases de dados públicas com tráfego de rede simulado (como *CICIDS2017 e NSL-KDD*), o que facilita a experimentação e avaliação de modelos.

Figura 5 - Distribuição das abordagens de IDS (NIDS, HIDS ou híbrido) nos estudos revisados



Fonte: elaborado pelo autor

Em segundo lugar, destacam-se os sistemas híbridos, encontrados em 37,50% dos artigos analisados. Estes sistemas combinam características de NIDS e HIDS, visando uma detecção mais robusta e abrangente, cobrindo tanto o tráfego da rede quanto atividades específicas do host. Os sistemas HIDS, por sua vez, apareceram em apenas 3,57% dos estudos, especificamente nos trabalhos A36 e A38. Esse número reduzido pode estar relacionado à dificuldade de obtenção de dados sensíveis a nível de host, além das questões de privacidade e limitações de generalização para ambientes diversos. Por fim, 14,29% dos

estudos não especificaram a abordagem utilizada, o que limita a compreensão sobre o contexto de aplicação dos modelos propostos. Esses resultados indicam que, embora os sistemas baseados em rede continuem sendo os mais populares, há um interesse crescente por soluções híbridas, que buscam aumentar a abrangência e precisão dos sistemas de detecção de intrusão.

5.3. Quais as principais métricas dos estudos levantados?

Para avaliar a eficácia dos modelos de detecção de intrusão nos estudos revisados, foram extraídas e analisadas as principais **métricas de desempenho** utilizadas na literatura: **Precisão**, **Acurácia**, **Recall**, **F1-Score**, **Taxa de Falsos Positivos (TFP)**, **Taxa de Falsos Negativos (TFN)**, **Taxa de Verdadeiros Negativos (TVN)** e **AUC/AUROC**. A Tabela 7 detalha os valores encontrados por estudo, e os dados foram posteriormente agregados para análise estatística.

Tabela 7 - Métricas de desempenho dos modelos analisados por estudo

Identificador	Precisão	Acurácia	Recall	F1-Score	TFP	TFN	TVN	AUC/AUROC
A1	-	0,987	0,989	-	0,013	0,011	-	-
A2	0,995	1,000	0,880	0,934	0,010	0,120	0,990	-
A3	0,379	0,873	0,935	0,912	0,060	-	0,940	-
A4	-	0,859	-	0,913	-	-	-	-
A5	-	0,984	-	-	-	-	-	-
A6	-	-	-	-	-	-	-	-
A7	-	-	-	-	-	-	-	-
A8	-	-	-	-	-	-	-	-
A9	-	0,910	-	-	-	-	-	-
A10	-	-	-	-	-	-	-	-
A11	0,997	0,997	0,997	0,997	-	-	-	-
A12	0,990	-	0,990	-	-	-	-	-
A13	-	0,996	-	-	-	-	-	1,000
A14	-	-	-	0,870	-	-	-	-
A15	0,849	-	0,969	0,905	-	-	-	-

A16	1,000	1,000	1,000	1,000	-	-	-	0,925
A18	0,933	0,945	0,953	0,943	-	-	-	0,945
A19	0,880	0,850	0,850	0,850	-	-	-	-
A20	0,997	0,997	0,997	0,997	-	-	-	0,999
A21	0,995	0,997	0,978	0,986	-	-	-	-
A22	-	-	-	-	-	-	-	-
A23	-	-	-	-	-	-	-	-
A24	0,996	0,991	0,986	0,991	-	-	-	0,996
A25	0,379	0,873	0,935	0,912	0,060	-	0,940	-
A26	0,999	0,999	0,999	0,999	-	-	-	-
A27	-	-	-	-	-	-	-	-
A28	-	-	-	-	-	-	-	0,921
A29	-	-	-	-	-	-	-	-
A30	0,889	-	1,000	0,941	0,000	0,000	-	-
A31	-	-	-	0,840	-	-	-	-
A32	-	-	-	-	-	-	-	-
A33	0,996	0,991	0,986	0,991	-	-	-	0,996
A34	-	0,921	-	0,923	-	-	-	0,959
A35	0,970	0,954	0,938	0,954	-	-	-	-
A36	-	-	-	-	-	-	-	0,980
A37	-	-	-	-	-	-	-	-
A38	0,980	0,990	-	-	0,010	-	-	0,980
A39	-	-	-	-	-	-	-	-
A40	0,986	0,987	0,987	0,987	-	-	-	-
A41	-	-	-	-	-	-	-	-
A42	-	-	-	-	-	-	-	-
A43	-	-	-	-	-	-	-	-
A44	0,996	-	1,000	0,998	-	-	-	-
A45	-	-	-	-	-	-	-	-
A46	-	-	-	-	-	-	-	0,987

A47	-	-	-	-	-	-	-	-
A48	0,998	0,998	0,998	0,998	-	-	-	-
A50	0,994	-	0,997	0,996	-	-	-	-
A51	1,000	1,000	1,000	1,000	-	-	-	-
A52	-	-	-	-	-	-	-	-
A53	-	-	-	-	-	-	-	-
A54	-	-	-	-	-	-	-	-
A55	0,872	0,872	0,937	0,879	-	-	-	-
A56	0,997	-	0,994	0,995	-	-	-	-
A57	-	0,908	0,760	-	-	-	-	-

Fonte: elaborado pelo autor

As médias dessas métricas, calculadas com base nos estudos que forneceram os valores correspondentes, estão apresentadas na tabela 8 abaixo:

Tabela 8 - Média das métricas de todos os estudos

Métrica	Média	Desvio Padrão
Precisão	0,9194	0,1729
Acurácia	0,9551	0,0539
Recall	0,9622	0,0576
F1-Score	0,9504	0,0519
Taxa de Falsos Positivos	0,0254	0,0269
Taxa de Falsos Negativos	0,0437	0,0663
Taxa de Verdadeiros Negativos	0,9569	0,0287
AUC/AUROC	0,9715	0,0296

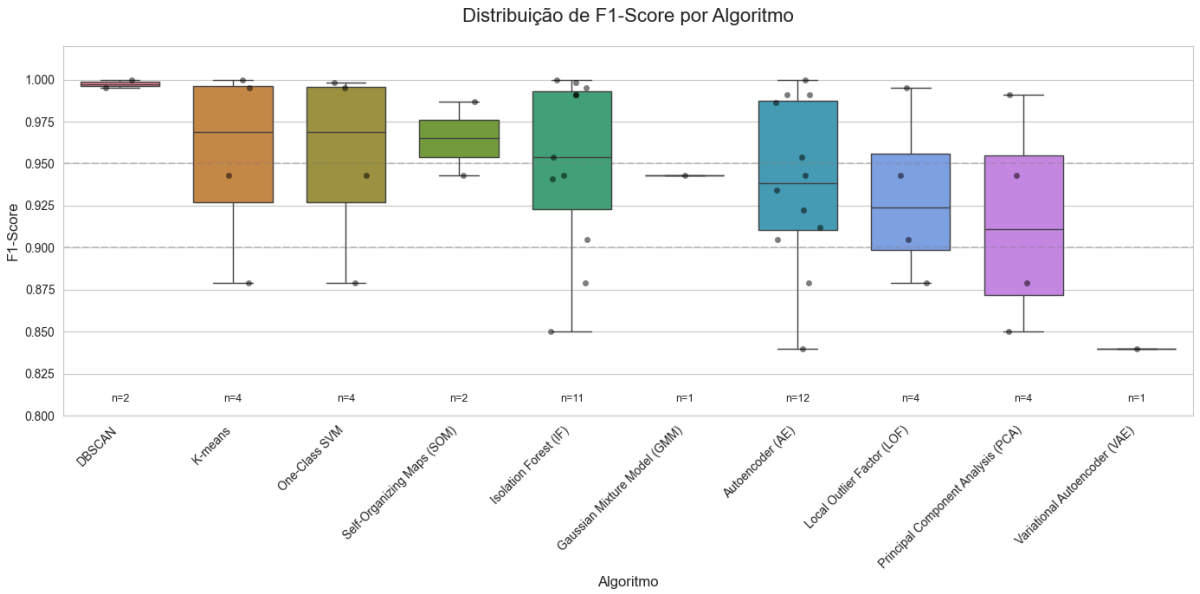
Fonte: Elaborado pelo autor

A alta precisão média (0,9194) evidencia que a maioria dos alertas emitidos pelos modelos correspondeu, de fato, a eventos maliciosos reais. Isso é fundamental para evitar sobrecarga operacional e desgaste de recursos humanos em ambientes de segurança. Da mesma forma, a acurácia média (0,9551) aponta para um bom desempenho global na classificação correta tanto de eventos normais quanto de intrusões. A sensibilidade (recall),

com valor médio de 0,9622, indica que os modelos foram eficazes em identificar a maior parte das intrusões presentes nos dados. Já o F1-Score médio de 0,9504 confirma o equilíbrio entre precisão e recall, sendo particularmente relevante em contextos com dados desbalanceados — algo comum em sistemas de detecção de intrusão. As taxas médias de erro também apresentaram desempenho positivo: a taxa de falsos positivos foi de apenas 2,54%, enquanto a taxa de falsos negativos foi de 4,37%, sugerindo que os modelos não apenas evitam alarmes falsos, como também são eficazes na detecção de comportamentos anômalos. Complementarmente, a taxa de verdadeiros negativos (0,9569) indica que os modelos foram eficientes em reconhecer atividades legítimas da rede. A taxaAUC/AUROC, que mede a capacidade de discriminação entre classes, apresentou valor médio de 0,9715, demonstrando excelente performance dos modelos, mesmo sob cenários desbalanceados.

Quando analisado o desempenho do F1-Score em função dos diversos algoritmos clássicos percebe-se pouca variabilidade no resultado desta métrica, embora alguns algoritmos tenha pouca representabilidade como *DBSCAM, Gaussian Mixture Model (GMM) e Variational Autoencoder (VAE)*. Outros algoritmos possuem alto índice de desempenho F1, inclusive com amostras mais robustas como Isolation Forest e Autoencoder (AE).

Figura 6 - Desempenho médio do F1-Score por algoritmo de aprendizado de máquina



Fonte: elaborado pelo autor

Em conjunto, esses resultados apontam que os modelos analisados são promissores para aplicação em ambientes reais, oferecendo alto desempenho na identificação de ameaças, com baixa incidência de falsos alertas. No entanto, é importante destacar que nem todos os estudos reportaram todas as métricas analisadas, o que limita comparações diretas em alguns casos.

5.4 Quais tipos específicos de ataques/ameaças são mais frequentemente detectados por modelos de aprendizado de máquina?

Os modoelos têm se mostrado particularmente eficazes na identificação de determinadas categorias de ataques e ameaças cibernéticas, conforme evidenciado pela literatura especializada. A tabela abaixo mostra os principais tipos de ameaças e ataques:

Tabela 9 - Tipos de ataques e ameaças detectados por modelos de aprendizagem de máquina

Tipo de Ataque/Ameaça	Identificador(es) do(s) Artigos
Ataques de Negação de Serviço (DoS/DDoS)	A1, A2, A4, A5, A6, A7, A8, A10, A11, A18, A20, A21, A22, A33, A41, A47, A52, A54, A55
Ataques de Força Bruta	A2, A4, A8, A10, A11, A26, A47, A56

Ataques de Varredura/Porta	A2, A4, A8, A10, A11, A33, A41, A47, A56
Ataques de Infiltração/Exfiltração	A2, A4, A5, A8, A10, A11, A26
Ataques de Spoofing	A21, A22
Ataques Web	A4, A5, A8, A11, A20, A26, A47
Outros Ataques Específicos	A1, A2, A4, A5, A11, A12, A22, A24, A30, A36, A39, A47
Ameaças Avançadas Persistentes (APT)	A3, A25
Ataques Dia Zero (Zero-Day)	A5, A18, A7, A54, A56, A57
Malware/Botnet	A11, A26, A37, A38, A39, A41, A47, A53, A54
Outliers/Anomalias Genéricas	A2, A4, A7, A9, A13, A14, A15, A16, A17, A19, A23, A24, A27, A28, A29, A31, A32, A34, A35, A37, A40, A41, A44, A45, A46, A48, A49, A50, A53, A54

Fonte: elaborado pelo autor

Entre os ataques mais frequentemente detectados por essas abordagens, destacam-se os ataques de negação de serviço (DoS/DDoS), amplamente documentados nos artigos A1, A2, A4, A5, A6, A7, A8, A10, A11, A18, A20, A21, A22, A33, A41, A47, A52, A54 e A55. Esses ataques são facilmente identificáveis devido às suas características de tráfego anormal, como volumes excessivos de requisições ou pacotes com padrões incomuns.

Outra categoria relevante é a dos ataques de força bruta (A2, A4, A8, A10, A11, A26, A47, A56) e varredura de portas (A2, A4, A8, A10, A11, A33, A41, A47, A56), que geram múltiplas tentativas de acesso ou conexões sequenciais, configurando anomalias detectáveis por meio de algoritmos de clustering ou análise de frequência.

Ataques mais sofisticados, como infiltração/exfiltração de dados (A2, A4, A5, A8, A10, A11, A26), spoofing (A21, A22) e ameaças avançadas persistentes (APTs) (A3, A25), também são identificados por essas técnicas, uma vez que envolvem comportamentos atípicos na comunicação de rede ou acesso a sistemas. Além disso, ataques *zero-day* (A5, A7, A18, A54, A56, A57) e *malware/botnets* (A11, A26, A37, A38, A39, A41, A47, A53, A54) são frequentemente detectados devido às suas assinaturas comportamentais incomuns.

Observa-se também uma lacuna significativa na abordagem de ataques furtivos, como os que envolvem infiltração/exfiltração de dados ao longo do tempo, *beaconing* ou comunicação intermitente de comando e controle (C2). Tais comportamentos são característicos das chamadas Ameaças Persistentes Avançadas (APT) ou de ataques do tipo

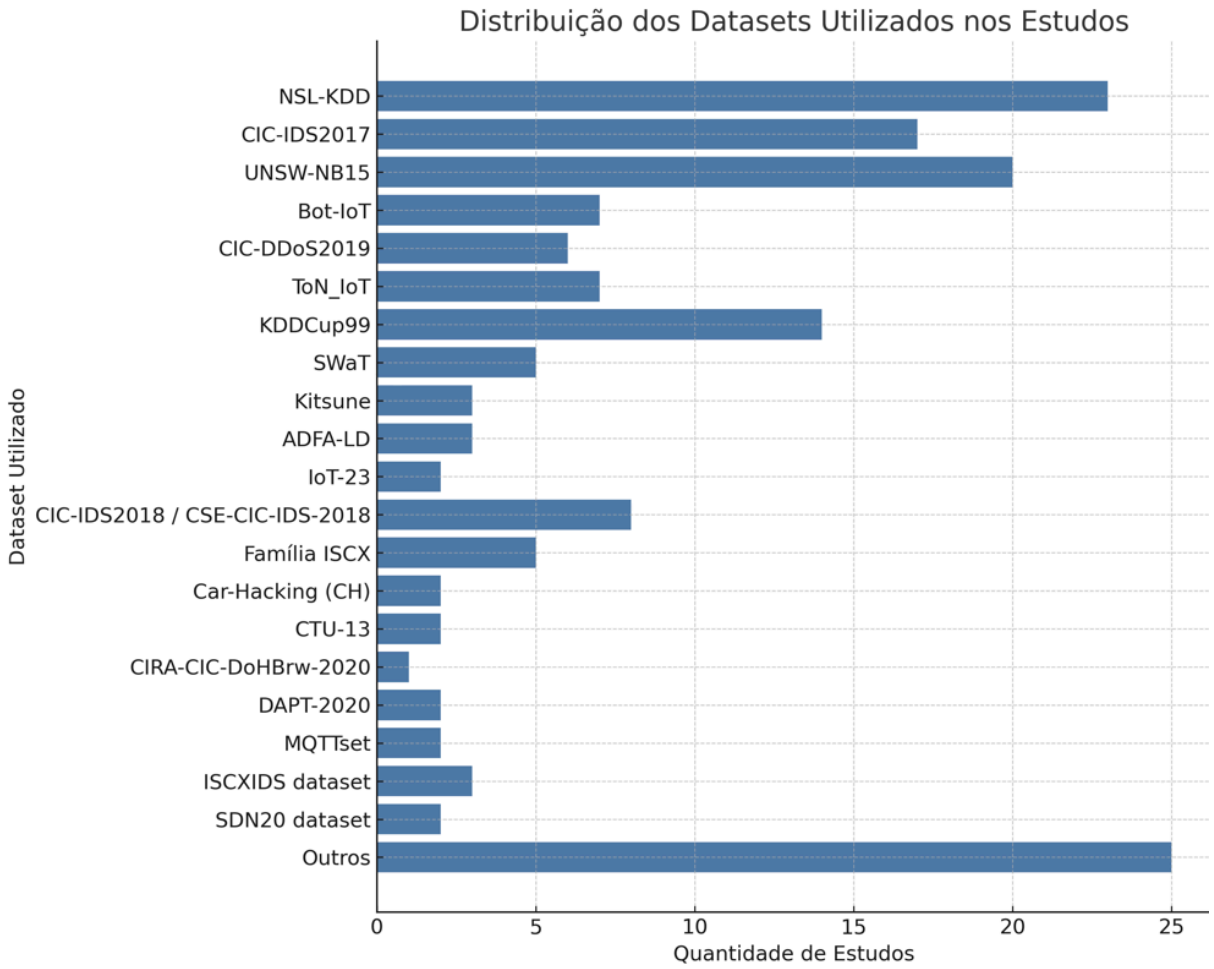
"*low-and-slow*", os quais buscam evitar a detecção por mecanismos baseados em volume ou em janelas temporais curtas. Apenas três estudos (A3, A25 e A38) mencionam explicitamente a avaliação de modelos frente a esses padrões de tráfego mais sutis e persistentes. Essa concentração em cargas evidentes compromete a capacidade de generalização dos modelos de detecção, além de negligenciar cenários reais em que o atacante opta por movimentações laterais silenciosas, roubo gradual de dados ou exploração de vulnerabilidades em longo prazo. Portanto, destaca-se a necessidade de futuras pesquisas focarem em ataques furtivos de baixa taxa de pacote, propondo abordagens mais sensíveis a variações temporais prolongadas e menos dependentes de picos volumétricos.

Por fim, os modelos são eficientes na detecção de anomalias genéricas e *outliers* (A2, A4, A7, A9, A13, A14, A15, A16, A17, A19, A23, A24, A27, A28, A29, A31, A32, A34, A35, A37, A40, A41, A44, A45, A46, A48, A49, A50, A53, A54), que podem indicar ataques ainda não classificados ou variações de ameaças conhecidas.

5.5 Quais principais conjuntos de dados utilizados?

A pesquisa em detecção de intrusões e anomalias em redes depende fortemente do conjuntos de dados (*datasets*) que representem tráfego realista, contendo tanto comportamentos legítimos quanto maliciosos. Conforme a literatura analisada, diversos *datasets* têm sido amplamente utilizados em estudos recentes, cada um com características específicas que os tornam adequados para diferentes cenários de ataques. Abaixo segue um gráfico que demonstra os diversos *datasets* utilizados em função da sua frequência da lista final de artigos selecionados:

Figura 7 - Frequência de utilização dos principais conjuntos de dados nos estudos analisados



Fonte: elaborado pelo autor

O *NSL-KDD* (A1, A6, A12, A15, A16, A19, A23, A2, A29, A31, A34, A35, A36, A4, A40, A41, A47, A5, A51, A54, A55, A56, A57) é um dos mais consolidados, sendo uma versão melhorada do antigo *KDDCup99*. Ele contém registros de tráfego de rede rotulados como normais ou ataques, incluindo categorias como *DoS*, varredura de portas e invasões. Sua popularidade deve-se à sua ampla adoção como benchmark para avaliação de algoritmos de detecção de intrusões.

Outro *dataset* relevante é o *CIC-IDS2017* (A11, A16, A19, A2, A24, A26, A27, A29, A31, A33, A37, A40, A47, A8, A52, A54, A57), que simula ataques modernos, como força bruta, *DDoS* e infiltração, em um ambiente mais recente e diversificado. Da mesma forma, o *UNSW-NB15* (A12, A16, A19, A2, A29, A31, A33, A39, A4, A40, A41, A45, A46, A8, A9, A51, A52, A54, A55, A57) apresenta uma variedade de ataques contemporâneos, sendo

frequentemente utilizado para avaliar métodos de aprendizado de máquina em cenários desbalanceados.

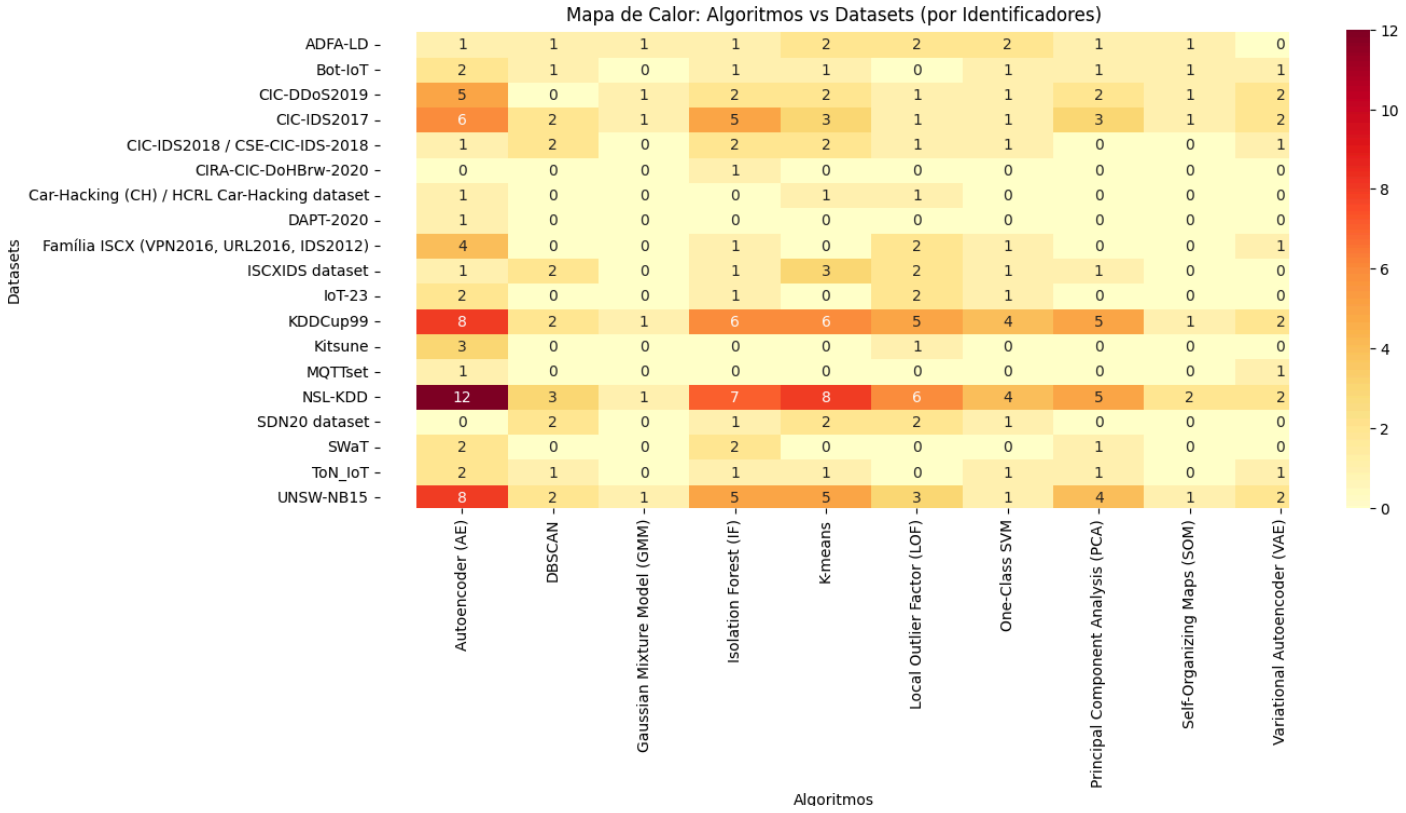
Para ambientes de Internet das Coisas (IoT), destacam-se o *Bot-IoT* (A16, A23, A31, A40, A48, A8, A51) e o *ToM IoT* (A11, A16, A2, A23, A31, A4, A48), que incluem tráfego gerado por dispositivos comprometidos, como *bots* em redes IoT. Além disso, o *CIC-DDoS2019* (A1, A18, A24, A29, A31, A4) é especializado em ataques distribuídos de negação de serviço, enquanto o *Kitsune* (A10, A2, A41) é voltado para detecção de malware em redes.

Em cenários industriais, o *SWaT* (A11, A14, A24, A33, A50) é amplamente utilizado para avaliar ameaças a sistemas de controle industrial (ICS), enquanto o *Car-Hacking Dataset* (A21, A22) foca em ataques a redes veiculares. Além desses, *datasets* como *CTU-13* (A38, A46), *CIRA-CIC-DoHBrw-2020* (A30) e *DAPT-2020* (A25, A3) são empregados em pesquisas sobre *botnets*, tráfego *DNS* sobre *HTTPS* e ameaças persistentes avançadas (APAs), respectivamente.

Conjuntos como *MQTTset* (A31, A47), *ISCXIDS* (A54, A56, A57) e *SDN20* (A56, A57) são utilizados em contextos específicos, como comunicações que utilizam protocolo *MQTT (Message Queuing Telemetry Transport)* e *Software Defined Networking (SDN)*.

A partir da contagem cruzada entre os dados apresentados na Figura 4 e 7, observa-se também um padrão recorrente na escolha dos algoritmos de aprendizado em função do conjunto de dados utilizado. A Figura 8 apresenta um mapa de calor que cruza a frequência de utilização de algoritmos de aprendizado de máquina com os diferentes *datasets* empregados nos artigos analisados. A contagem é baseada no número de identificadores de artigos que utilizaram simultaneamente um determinado algoritmo e um *dataset* específico.

Figura 8 - Mapa de calor da ocorrência entre algoritmos e conjunto de dados nos estudos selecionados



Fonte: elaborado pelo autor

Observa-se que o dataset *NSL-KDD* destaca-se significativamente por sua alta recorrência com diversos algoritmos, especialmente com *Autoencoder (AE), K-means, Local Outlier Factor (LOF) e Isolation Forest (IF)*, acumulando até 12 ocorrências com *AE*. Isso reforça a posição do *NSL-KDD* como um benchmark clássico amplamente utilizado em estudos de detecção de intrusão, especialmente em abordagens baseadas em aprendizado. Outro dataset frequentemente utilizado é o *UNSW-NB15*, que também apresenta correlações relevantes com múltiplos algoritmos, em especial com *Autoencoder (AE) e K-means*. O *KDDCup99* ainda se mantém presente em diversos trabalhos, embora em menor intensidade, indicando que muitos artigos recentes sobre o tema ainda utilizam *datasets* antigos, o que pode interferir negativamente no resultado final das métricas levantadas.

Em relação aos algoritmos, destaca-se o uso expressivo de técnicas como *K-means, Isolation Forest e LOF*, que são métodos amplamente adotados em contextos de detecção de

anomalias, dada sua capacidade de trabalhar com dados não rotulados. O uso de *Autoencoders* também aparece de forma recorrente, indicando o avanço no emprego de técnicas mais modernas baseadas em redes neurais profundas para representação e reconstrução de dados. O mapa ainda revela uma diversidade crescente de datasets mais recentes e específicos, como CIC-IDS2017, IoT-23, ToN_IoT e CIC-DDoS2019, que vêm sendo utilizados em conjunto com diferentes algoritmos, demonstrando a busca por abordagens mais realistas e atualizadas para avaliação de modelos de detecção.

A figura 8 permite inferir, portanto, que a correlação entre *dataset* e algoritmo escolhido está frequentemente associada à familiaridade da comunidade científica com *benchmarks* específicos, mais do que fundamentada na adequação técnica ou na eficácia da técnica frente ao tipo de tráfego analisado. Tal constatação reforça a importância de se realizarem experimentos em múltiplos conjuntos de dados (*cross-dataset*), de modo a avaliar a capacidade de generalização dos modelos propostos e garantir maior robustez às soluções desenvolvidas.

5.6 Quais as principais limitações/desafios encontrados na literatura?

As principais limitações e desafios encontrados na literatura sobre Sistemas de Detecção de Intrusões (IDS) baseados em Aprendizagem de Máquina (ML) e Aprendizagem Profunda (DL) são multifacetados, abrangendo desde a disponibilidade e qualidade dos dados até as restrições inerentes aos modelos e os impedimentos na sua implementação prática [A4, A17, A19, A23, A37]. Uma das barreiras mais significativas é a escassez de *datasets* de tráfego de rede sistemáticos, atualizados e que representem fielmente o cenário complexo e dinâmico das ameaças cibernéticas do mundo real [A4, A17, A19, A23, A37]. Muitos estudos, por exemplo, ainda se baseiam em conjuntos de dados mais antigos, como o *KDDCup99*, que pode apresentar vieses de classe e conter tipos de ataques desatualizados, não refletindo as ameaças contemporâneas [A11, A15, A18, A19, A34, A36, A40]. Para ambientes específicos, como redes veiculares (*CAN*) ou dispositivos de Internet das Coisas (IoT), há uma notável ausência de dados de ataque de alta qualidade e uma quantidade insuficiente de dados para o treinamento eficaz de modelos de ML [A14, A22, A37, A48].

Adicionalmente, a obtenção de dados rotulados é um processo árduo, caro e, em muitos casos, inviável [A8, A18, A22, A30, A38]. Os *datasets* disponíveis podem conter imperfeições significativas, como contagens inconsistentes de flags TCP, valores negativos em features de fluxo, dados ausentes e rótulos imprecisos ou misturados (onde o tráfego

normal é erroneamente categorizado como malicioso) [A24, A27, A31]. Há também um desafio persistente na generalização do aprendizado entre diferentes ambientes de rede e *datasets* diversos, com a maioria dos trabalhos que utilizam múltiplos *datasets* falhando em explorar essa generalização, avaliando as abordagens apenas isoladamente para cada conjunto de dados [A2, A22, A26, A31]. A ausência de uma comparação sistemática e padronizada entre diferentes técnicas e modelos usando os mesmos *datasets* dificulta a avaliação da robustez e eficácia real das soluções propostas [A1, A5, A16, A17, A18, A45]. Modelos de detecção de anomalias, embora mais flexíveis para identificar ataques desconhecidos, frequentemente sofrem de elevadas taxas de falsos positivos, classificando erroneamente o tráfego benigno como anômalo [A6, A10, A11, A22, A30, A33]. Isso pode levar à "fadiga de alerta" e à eventual desativação dos sistemas pelos usuários [A22]. Classificadores baseados em DL também podem apresentar falsos negativos para ataques nunca vistos no treinamento [A10].

A complexidade e baixa interpretabilidade de muitos modelos de DL/ML são desafios notáveis. Apesar de seu alto desempenho, operam como "caixas pretas" devido à sua complexidade intrínseca, tornando difícil para especialistas humanos compreenderem e interpretarem as razões por trás de suas predições [A2, A9, A25, A30, A41]. Essa limitação compromete a confiança e a adoção dessas soluções em ambientes críticos de cibersegurança [A9].

Os requisitos computacionais elevados são outro obstáculo. A maioria das abordagens baseadas em DL/ML exige poder de processamento significativo, memória substancial e, por vezes, hardware especializado [A5, A8, A18, A19, A21, A23, A27, A28, A29, A32, A35, A41, A47]. Isso representa um desafio considerável para a implementação em dispositivos com recursos limitados, como os dispositivos IoT [A2, A8, A13, A35]. A vulnerabilidade a evasão e ofuscação, por meio de técnicas como criptografia e tunelamento (ex: *DNS-over-HTTPS* - *DoH*), também representa um grande desafio para a inspeção do payload e a detecção de incidentes, tornando as abordagens tradicionais ineficazes [A13, A30]. O escopo limitado de detecção de alguns modelos existentes, que focam em poucos tipos de ataques ou realizam apenas classificação binária, limita sua capacidade de abordar o espectro completo das ameaças [A2, A25, A35, A47]. Além disso, a adaptação a topologias de rede dinâmicas e a novas técnicas de ataque em tempo real é um desafio contínuo para esses sistemas [A23, A25, A33, A38].

A definição de limiares de detecção e a otimização de hiperparâmetros são tarefas complexas. A seleção de um limiar adequado para dados do mundo real é desafiadora, pois

um limiar mal definido pode resultar em classificações erradas [A24, A29, A33, A41, A44, A46]. A otimização de hiperparâmetros para modelos de ML/DL é frequentemente uma tarefa manual e complexa, especialmente na ausência de conhecimento prévio sobre o comportamento do modelo [A5, A15, A18, A20, A28, A41, A45].

A latência e escalabilidade em tempo real também são aspectos preocupantes. O pré-processamento de dados e a complexidade dos modelos podem introduzir latência na inferência, o que é crítico para detecção em tempo real [A2, A21, A23, A27, A28, A41, A45, A46]. A escalabilidade dos modelos em ambientes de alto tráfego com volumes massivos de dados continua sendo um desafio [A4, A18, A19, A27, A28, A31, A46].

Muitos métodos de detecção de anomalias ainda dependem de intervenções humanas ou especialistas para atualizações, ajustes ou validação, o que pode limitar a automação e a escalabilidade em ambientes de alto tráfego [A4, A9, A38, A40]. Além disso, a literatura aponta a vulnerabilidade de modelos de ML/DL a ataques adversariais (por exemplo, *data poisoning*), embora a robustez a tais ataques seja uma área de pesquisa ativa [A2, A10, A23, A41, A46]. Observa-se que os estudos frequentemente utilizam um número limitado de métricas para avaliação do desempenho, o que pode não fornecer uma visão completa da eficácia dos sistemas [A1, A5, A7, A18, A25, A29, A31, A37, A44, A45].

Ademais, propõem-se algumas direções práticas para mitigar os desafios identificados: (1) estabelecer metas de desempenho realistas, considerando *benchmarks* recentes e diversos que posicionam o F1-score médio acima de 95%; (2) reprovar abordagens que apresentam desempenho elevado em apenas um *dataset*, exigindo validação cruzada em múltiplos conjuntos de dados; (3) incorporar métricas de tempo de resposta e consumo energético, considerando restrições reais de hardware e latência; (4) ampliar a cobertura dos conjuntos de dados para incluir ataques furtivos como movimentos laterais e ameaças persistentes (APTs); (5) traçar métricas de interpretabilidade aos modelos, combinando técnicas como *Autoencoders* ou *Isolation Forest* com mecanismos interpretáveis, tais como aprendizado por protótipos e contrafactuais; e (6) validar modelos incrementados com *GANs* em tráfego real antes de reportar resultados.

6. Conclusões e Trabalhos Futuros

Este trabalho realizou uma revisão sistemática da literatura sobre a aplicação de algoritmos de aprendizado de máquina em sistemas de detecção de intrusão (IDS), seguindo o

método proposto por Kitchenham e Charters (2007). A análise de 57 artigos publicados entre 2020 e 2025 revelou tendências, desafios e lacunas significativas no campo, destacando a eficácia dessas abordagens para identificar ataques desconhecidos e padrões anômalos em ambientes dinâmicos e complexos. Os resultados demonstraram que algoritmos como *Autoencoders* (presentes em 16,3% dos estudos), *Isolation Forest* (10,4%) e *K-means* (8,9%) são os mais utilizados devido à sua capacidade de lidar com dados não rotulados e detectar desvios sutis.

Além disso, os *datasets NSL-KDD, CIC-IDS2017 e UNSW-NB15* emergiram como os mais empregados, refletindo sua relevância como *benchmarks* para avaliação de modelos. Quanto aos tipos de ataques, esses modelos de aprendizado de máquina mostraram-se particularmente eficazes contra ataques de negação de serviço (*DoS/DDoS*), varredura de portas e ameaças *zero-day*, que exigem detecção baseada em comportamento anômalo. Apesar do desempenho promissor (com médias de AUC/AUROC de 0,9715 e F1-Score de 0,9504), os desafios identificados foram significativos. Entre as principais limitações estão a escassez de conjuntos de dados atualizados e representativos de ameaças contemporâneas, além de vieses em bases antigas como o *KDDCup99*. As altas taxas de falsos positivos podem levar à "fadiga de alerta" e reduzir a confiança nos sistemas. A complexidade e custo computacional também se destacam como obstáculos, já que modelos como *autoencoders* e redes neurais demandam recursos significativos, limitando sua aplicação em dispositivos com restrições, como os dispositivos IoT. Outro desafio relevante é a dificuldade de generalização, pois muitos modelos falham em adaptar-se a diferentes ambientes de rede ou a novas técnicas de ataque sem retreinamento.

As contribuições deste trabalho incluem a sistematização do conhecimento por meio do mapeamento dos algoritmos, conjunto de dados e métricas mais relevantes para IDSs, além da identificação de lacunas importantes, como a necessidade de conjunto de dados mais diversificados e métodos para reduzir falsos positivos. Como direcionamentos para pesquisas futuras, sugere-se a exploração de técnicas híbridas (semi-supervisionadas), o aprimoramento da interpretabilidade de modelos e o desenvolvimento de *frameworks* para adaptação em tempo real.

Em suma, os modelos de aprendizado de máquina representam uma ferramenta valiosa para a segurança cibernética, especialmente em cenários onde ataques desconhecidos são uma ameaça constante. No entanto, avanços em eficiência computacional, qualidade de dados e integração com abordagens complementares são essenciais para superar as limitações atuais e ampliar sua adoção em ambientes práticos. Este estudo fornece, assim, um ponto de partida

para pesquisas futuras que busquem equilibrar precisão, escalabilidade e robustez na detecção de intrusões, contribuindo para o desenvolvimento de sistemas de segurança mais adaptáveis e eficazes frente às crescentes ameaças cibernéticas.

7. Bibliografia

[1] Yuyang, Z., et al. Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier, 2020.

[2] Chuanlong Y., et al. A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks, 2017.

[3] Joldzic, O., Djuric, Z., Vuletic, P.. A transparent and scalable anomaly-based dos detection method, 2016.

[4] Papamartzivanos, D., Mármol, F.G., Kambourakis, G. Dendron: Genetic trees driven rule induction for network intrusion detection systems, 2018.

[5] A. Nisioti, A. Mylonas, P. D. Yoo and V. Katos. From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods. IEEE Commun. Surveys & Tut., vol. 20, no. 4, 2018.

[6] Wang, K., Du, M., Sun, Y., Vinel, A., Zhang, Y. Attack detection and distributed forensics in machine-to-machine networks. IEEE Network, 2016.

[7] Wang, K., Du, M., Yang, D., Zhu, C., Shen, J., Zhang, Y., 2016b. Game-theory-based active defense for intrusion detection in cyberphysical embedded systems. ACM Transactions on Embedded Computing Systems (TECS) 16, 18. doi:10.1145/2886100.

[8] Awad, M., Khanna, R. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Apress, 1ª edição, 2015.

[9] Rao, M. V., Midhunchakkaravarthy, D., Dandu, S. Joint detection and classification of signature and NetFlow based internet worms using MBGWO-based hybrid LSTM. Journal of computer virology and hacking techniques, 2023.

[10] Vinayakumar, R., Alazab, Mamoun, Soman, K. P., Poornachandran, Prabaharan, Al-Nemrat, A. and Venkatraman, Sitalakshmi. Deep Learning Approach for Intelligent Intrusion Detection System, 2019.

[11] Mitchell, T. M. *Machine Learning.* McGraw-Hill Science/Engineering/Math. 1997.

[12] Introduction to Unsupervised Learning. Disponível em: <https://www.bombaysoftwares.com/blog/introduction-to-unsupervised-learning>. Acesso em: 22 mai. 2024.

[13] Medeiros, C. J. F., Costa, J. A. F. Uma Comparação Empírica de Métodos de Redução de Dimensionalidade Aplicados a Visualização de Dados. Revista da Sociedade Brasileira de Redes Neurais (SBRN), Vol. 6, No. 2, pp. 81-110, 2008.

[14] Congyuan, X. et al. An Intrusion Detection System Using a Deep Neural Network With Gated Recurrent Units. 2018.

[15] Yu Liu, Y. , Jin, Y., Chen, J. e Wu, H. A Novel Semi-Supervised Learning Approach for Network Intrusion Detection on Cloud-Based Robotic System, 2018.

[16] Zoppi, T., Ceccarelli, A. e Bondavalli, A. Unsupervised Algorithms to Detect Zero-Day Attacks: Strategy and Application, 2021.

[17] Nassif, A. B., Talib, M. A., Nasir, Q. e Dakalba, F. M. Machine Learning for Anomaly Detection: A Systematic Review, 2021.

[18] Halbouni, A., Gunawan, T. S., Habaebi, M. H., Halbouni, M., Kartiwi, M. e Ahmad, R. Machine Learning and Deep Learning Approaches for CyberSecurity: A Review, 2021.

[19] J. Mirkovic e P. Reiher. A taxonomy of DDoS attack and DDoS defense mechanisms. ACM SIGCOMM Computer Communication Review, 2004.

[20] M. Sharif, L. Bauer, e M. K. Reiter. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016.

[21] H. Debar, M. Dacier, e A. Wespi. Towards a taxonomy of intrusion-detection systems. Computer Networks, 1999.

[22] Kitchenham, B. e Charters, S. Guidelines for performing systematic literature reviews in software engineering. Engineering, vol. 45, no. 4, p. 1051, 2007, doi: 10.1145/1134285.1134500.

[23] Chalapathy, R. e Chawla, S. Deep Learning for Anomaly Detection: A Survey, 2019.

[24] Chandola, V.; Banerjee, A.; Kumar, V. Anomaly Detection: A Survey. ACM Computing Surveys, v. 41, n. 3, p. 1–58, 2009.

[25] Schölkopf, B.; Bauer, S.; Ruff, L. Learning with Anomalies. Journal of Machine Learning Research, v. 22, p. 1–48, 2021.

[26] Goldstein, M.; Uchida, S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLOS ONE, v. 11, n. 4, p. e0152173, 2016.

[27] Tan, P.-N.; Steinbach, M.; Kumar, V. Introduction to Data Mining. 2. ed. Pearson, 2018.

[28] Liu, F. T.; Ting, K. M.; Zhou, Z.-H. Isolation-Based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, v. 6, n. 1, p. 1–39, 2020.

[29] Agrawal, R.; Imieliński, T.; Swami, A. Mining Association Rules Between Sets of Items in Large Databases. Data Mining and Knowledge Discovery, v. 1, n. 2, p. 203–228, 2021.

[30] Talaei Khoei, T.; Kaabouch, N. A Comparative Analysis of Supervised and Unsupervised Models for Detecting Attacks on the Intrusion Detection Systems, 2023.

[31] Borgioli, N.; Aromolo, F.; Phan, L. T. X.; Buttazzo, G. A convolutional autoencoder architecture for robust network intrusion detection in embedded systems, 2024.

[32] Kasongo, S. M. A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework, 2023.

[33] Uddin, M. A.; et al. A dual-tier adaptive one-class classification IDS for emerging cyberthreats, 2025.

[34] Touré, A.; et al. A framework for detecting zero-day exploits in network flows, 2024.

[35] Pu, G.; Wang, L.; Shen, J.; Dong, F. A hybrid unsupervised clustering-based anomaly detection method, 2021.

[36] Alem, S.; et al. A novel bi-anomaly-based intrusion detection system approach for industry 4.0, 2023.

[37] Hossain, S.; et al. A privacy-preserving Self-Supervised Learning-based intrusion detection system for 5G-V2X networks, 2025.

[38] Gu, X.; Howells, G.; Yuan, H. A semi-supervised soft prototype-based autonomous fuzzy ensemble system for network intrusion detection, 2025.

[39] Hore, S.; et al. A sequential deep learning framework for a robust and resilient network intrusion detection system, 2024.

[40] Lazzarini, R.; Tianfield, H.; Charissis, V. A stacking ensemble of deep learning models for IoT intrusion detection, 2023.

[41] Zhu, N.; et al. AEC_GAN: Unbalanced Data Processing Decision-Making in Network Attacks Based on ACGAN and Machine Learning, 2023.

[42] Shirley, J. J.; Priya, M. An Adaptive Intrusion Detection System for Evolving IoT Threats: An Autoencoder-FNN Fusion, 2025.

[43] Almohri, H. M. J.; Watson, L. T.; Evans, D. An Attack-Resilient Architecture for the Internet of Things, 2020.

[44] Kim, K. An Effective Anomaly Detection Approach based on Hybrid Unsupervised Learning Technologies in NIDS, 2024.

[45] Mohy-Eddine, M.; et al. An Ensemble Learning Based Intrusion Detection Model for Industrial IoT Security, 2023.

[46] Jang, Y.; et al. An Investigation of Learning Model Technologies for Network Traffic Classification Design in Cyber Security Exercises, 2023.

[47] Roopak, M.; et al. An unsupervised approach for the detection of zero-day distributed denial of service attacks in Internet of Things networks, 2024.

[48] Ness, S.; et al. Anomaly Detection in Network Traffic Using Advanced Machine Learning Techniques, 2025.

[49] Coscia, A.; et al. Automatic decision tree-based NIDPS ruleset generation for DoS/DDoS attacks, 2024.

[50] Rangsikunpum, A.; Amiri, S.; Ost, L. BIDS: An efficient Intrusion Detection System for in-vehicle networks using a two-stage Binarised Neural Network on low-cost FPGA, 2024.

[51] Lampe, B.; Meng, W. can-train-and-test: A curated CAN dataset for automotive intrusion detection, 2024.

[52] Azam, Z.; Islam, M. M.; Huda, M. N. Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree, 2023.

[53] Catillo, M.; Pecchia, A.; Villano, U. CPS-GUARD: Intrusion detection for cyber-physical systems and IoT devices using outlier-aware deep autoencoders, 2023.

[54] Mamun, A. A.; et al. Detection of advanced persistent threat: A genetic programming approach, 2024.

[55] Sarantos, P.; Violos, J.; Leivadeas, A. Enabling semi-supervised learning in intrusion detection systems, 2025.

[56] Pekar, A.; Jozsa, R. Evaluating ML-based anomaly detection across datasets of varied integrity: A case study, 2024.

[57] Meira, J.; et al. Fast anomaly detection with locality-sensitive hashing and hyperparameter autotuning, 2022.

[58] Li, Z.; Wang, P.; Wang, Z. FlowGANAnomaly: Flow-Based Anomaly Network Intrusion Detection with Adversarial Learning, 2024.

[59] Mungwarakarama, I.; Wang, Y.C.; Hei, X.H.; Song, X.; Nyesheja, E.M.; Turiho, J.C. FSDC: Flow Samples and Dimensions Compression for Efficient Detection of DNS-over-HTTPS Tunnels, 2024.

[60] Bertoli, G.D.; Pereira, L.A. Jr.; Saotome, O.; dos Santos, A.L. Generalizing intrusion detection for heterogeneous networks: A stacked-unsupervised federated learning approach, 2023.

[61] Sohrab, F.; Iosifidis, A.; Gabbouj, M.; Raitoharju, J. Graph-embedded subspace support vector data description, 2023.

[62] Baldoni, S.; Battisti, F. Histogram-based network traffic representation for anomaly detection through PCA, 2025.

[63] Xu, W.; Jang-Jaccard, J.; Singh, A.; Wei, Y.; Sabrina, F. Improving Performance of Autoencoder-Based Network Anomaly Detection on NSL-KDD Dataset, 2021.

[64] Sadaf, K.; Sultana, J. Intrusion Detection Based on Autoencoder and Isolation Forest in Fog Computing, 2020.
[65] Kim, C.; Jang, M.; Seo, S.; Park, K.; Kang, P. Intrusion Detection Based on Sequential Information Preserving Log Embedding Methods and Anomaly Detection Algorithms, 2021.
[66] Alrayes, F.S.; Zakariah, M.; Amin, S.U.; Khan, Z.I.; Helal, M. Intrusion Detection in IoT Systems Using Denoising Autoencoder, 2024.
[67] Mahboubi, A.; et al. Lurking in the shadows: Unsupervised decoding of beaconing communication for enhanced cyber threat hunting, 2025.
[68] Gao, X.; Hu, C.; Shan, C.; Liu, B.; Niu, Z.; Xie, H. Malware classification for the cloud via semi-supervised transfer learning, 2020.
[69] Christopher, V.; et al. Minority Resampling Boosted Unsupervised Learning With Hyperdimensional Computing for Threat Detection at the Edge of Internet of Things, 2021.
[70] Bovenzi, G.; et al. Network anomaly detection methods in IoT environments via deep learning: A Fair comparison of performance and robustness, 2023.
[71] Wang, Z.; Li, Z.; Fu, M.; Ye, Y.; Wang, P. Network traffic classification based on federated semi-supervised learning, 2024.
[72] Sewak, M.; Sahay, S.K.; Rathore, H. Neural AutoForensics: Comparing Neural Sample Search and Neural Architecture Search for malware detection and forensics, 2022.
[73] Wang, C.; Erfani, S.; Alpcan, T.; Leckie, C. OIL-AD: An anomaly detection framework for decision-making sequences, 2025.
[74] Heigl, M.; et al. On the improvement of the isolation forest algorithm for outlier detection with streaming data, 2021.
[75] Bhatia, S.; Liu, R.; Hooi, B.; Yoon, M.; Shin, K.; Faloutsos, C. Real-Time Anomaly Detection in Edge Streams, 2022.
[76] Wang, S.; Xu, W.; Liu, Y. Res-TranBiLSTM: An intelligent approach for intrusion detection in the Internet of Things, 2023.
[77] Zhang, H.; Upadhyay, D.; Zaman, M.; Jain, A.; Sampalli, S. SC-MLIDS: Fusion-based Machine Learning Framework for Intrusion Detection in Wireless Sensor Networks, 2025.
[78] Sen, Ö.; et al. Simulation of multi-stage attack and defense mechanisms in smart grids, 2025.
[79] Tu, F.F.; et al. STFT-TCAN: A TCN-attention based multivariate time series anomaly detection architecture with time-frequency analysis for cyber-industrial systems, 2024.
[80] Rahman, S.; Pal, S.; Mittal, S.; Chawla, T.; Karmakar, C. SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security, 2024.

[81] Coppolino, L.; D'Antonio, S.; Mazzeo, G.; Uccello, F. The good, the bad, and the algorithm: The impact of generative AI on cybersecurity, 2025.
[82] Gibert, D.; Mateu, C.; Planes, J. The rise of machine learning for detection and classification of malware: Research developments, trends and challenges, 2020.
[83] Kamaldeep; Dutta, M.; Granjal, J. Towards a Secure Internet of Things: A Comprehensive Study of Second Line Defense Mechanisms, 2020.
[84] Saha, S.; Priyoti, A.T.; Sharma, A.; Haque, A. Towards an Optimized Ensemble Feature Selection for DDoS Detection Using Both Supervised and Unsupervised Method, 2022.
[85] Zoppi, T.; Ceccarelli, A.; Bondavalli, A. Unsupervised Algorithms to Detect Zero-Day Attacks: Strategy and Application, 2021.
[86] Zoppi, T.; Ceccarelli, A.; Puccetti, T.; Bondavalli, A. Which algorithm can detect unknown attacks? Comparison of supervised, unsupervised and meta-learning algorithms for intrusion detection, 2023.
[87] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. Intriguing properties of neural networks, 2014.
[88] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. Learning under concept drift: A review, 2019.