Ataques de Envenenamento de Rótulos contra a Detecção de Zero-Day em Sistemas de Detecção de Intrusão Colaborativos

Giovanni H. M. de L. Siervo[1], Maria Eduarda S. Chessio[2], Silvio E. Quincozes[1,2], Daniel Mossé[3], Vagner E. Quincozes[4], Célio Albuquerque[4] e Diego Passos[5]

[1] Universidade Federal de Uberlândia (UFU) – Uberlândia, Brasil
[2] Universidade Federal do Pampa (UNIPAMPA) – Alegrete, Brasil
[3] University of Pittsburgh – Pittsburgh, USA
[4] Universidade Federal Fluminense (UFF) – Niterói, Brasil
[5] Instituto Superior de Engenharia de Lisboa (ISEL) – Lisboa, Portugal

{gsiervo, sequincozes}@ufu.br

Abstract. *Collaborative Intrusion Detection Systems (IDSs), such as Counselors Networks (CNs), enhance zero-day attack detection through the exchange of recommendations among autonomous nodes. However, this mechanism becomes vulnerable to label poisoning attacks, which can propagate through exchanged advice. This paper formalizes the advice poisoning attack and evaluates its impact on CN-based IDSs by analyzing learning dynamics and detection performance in a cyber-physical scenario involving unmanned aerial vehicles. Experiments conducted on a three-node CN, considering poisoning rates of 0%, 50%, and 100%, reveal an increase in conflicts among classifiers under selective poisoning, as well as a severe degradation in zero-day attack detection as the poisoning rate increases, ultimately collapsing the collaborative intelligence.*

Resumo. *Sistemas de Detecção de Intrusões (IDSs) colaborativos, como as Redes de Conselhos (CNs), ampliam a detecção de ataques zero-day por meio da troca de recomendações entre nós autônomos. Entretanto, esse mecanismo torna-se vulnerável a ataques de envenenamento de rótulos, capazes de se propagar pelos conselhos trocados. Este trabalho formaliza um novo tipo de ataque denominado envenenamento de conselhos e avalia o seu impacto em IDSs baseados em CNs, analisando a dinâmica de aprendizado e o desempenho de detecção em um cenário ciberfísico com veículos aéreos não tripulados. Experimentos em uma CN de três nós, com taxas de envenenamento de 0%, 50% e 100%, mostram aumento de conflitos entre classificadores quando há envenenamento seletivo e degradação severa da detecção de zero-day quando há envenenamento total, culminando no colapso da inteligência colaborativa.*

1. Introdução

Sistemas de Detecção de Intrusões (*Intrusion Detection Systems* – IDSs) colaborativos têm emergido como uma estratégia promissora para lidar com a crescente complexidade, heterogeneidade e escala dos ataques em ambientes distribuídos [Vasilomanolakis et al. 2015]. Ao permitir que múltiplos nós que executam IDSs independentes compartilhem conhecimento ou parâmetros de modelos, essas abordagens ampliam a capacidade de detecção e mitigam limitações do aprendizado isolado, como escassez de dados locais,

baixa generalização e maior vulnerabilidade a ataques desconhecidos e não catalogados (*zero-day attacks*) [Belenguer et al. 2025].

No contexto do aprendizado distribuído, duas vertentes relevantes foram desenvolvidas: as Redes de Conselhos (*Counselors Network* - CN) [Quincozes et al. 2021] e o Aprendizado Federado (*Federated Learning* – FL) [McMahan et al. 2017]. Nas CN, cada IDS preserva a sua autonomia decisória e operacional, trocando recomendações entre pares para apoiar a atualização de modelos de classificações, possibilitando adaptação contínua. Já o FL foi inicialmente concebido com agregação coordenada, mas evoluiu para variantes descentralizadas e *peer-to-peer*, nas quais a colaboração ocorre sem um agregador central, aproximando-se de arquiteturas totalmente distribuídas como as CN.

Apesar dos benefícios, a natureza colaborativa desses sistemas amplia significativamente a superfície de ataque, tornando-os suscetíveis a ameaças específicas que exploram os próprios mecanismos de cooperação. Entre essas ameaças, destacam-se os ataques de envenenamento de rótulo (*label poisoning*), nos quais um adversário manipula rótulos durante o processo de aprendizado com o objetivo de degradar o desempenho do IDS ou induzir decisões incorretas de forma persistente [Lavaur et al. 2025, Jha et al. 2023]. Na literatura, demonstra-se que ataques como envenenamento e inversão de rótulos degradam significativamente o desempenho de sistemas de aprendizagem federados, mesmo sob simples suposições adversárias [Yang et al. 2023, Rodríguez-Barroso et al. 2023].

Já no contexto das CNs, a troca de conselhos com rótulos envenenados pode se propagar entre nós e influenciar modelos locais mesmo na ausência de compartilhamento de conjuntos de dados e de parâmetros dos modelos. Em cenários ciberfísicos, como redes de veículos aéreos não tripulados (VANTs), essa propagação tende a ser particularmente crítica devido à natureza distribuída, dinâmica e sensível a falhas desses ambientes. Apesar desse potencial de impacto, a literatura ainda carece de investigações específicas sobre os efeitos de ataques de envenenamento de rótulos em arquiteturas colaborativas baseadas em CNs. Diante dessa lacuna, este trabalho investiga a seguinte questão de pesquisa:

QP1: Como os ataques de envenenamento de conselhos afetam a dinâmica de aprendizado, a confiança e o desempenho global de IDSs organizados como CNs, considerando um cenário ciberfísico como estudo de caso?

Com base nessa questão, este trabalho tem como objetivo geral avaliar sistematicamente o impacto de ataques de envenenamento de rótulos em IDSs colaborativos organizados como CNs. Como estudo de caso, a avaliação é conduzida em um cenário ciberfísico de detecção de intrusões envolvendo veículos aéreos não tripulados. Especificamente, o estudo busca: (i) caracterizar o ataque de envenenamento de conselhos, derivado do *label poisoning* no contexto da troca de conselhos; (ii) analisar os efeitos desses ataques sobre a estabilidade dos modelos locais; e (iii) mensurar como conselhos envenenados afetam a eficácia global do sistema, considerando métricas de detecção como acurácia, *recall* e F1-Score.

O restante deste trabalho está organizado como segue. Na Seção 2, a fundamentação teórica é apresentada. Na Seção 3, são discutidos os trabalhos relacionados. Na Seção 4, a arquitetura de CNs é apresentada. Na Seção 5, o modelo de ameaças de envenenamento de conselhos é apresentado, enquanto que na Seção 6 um estudo de caso específico é apresentado. Na Seção 7, os resultados são discutidos. Na Seção **??**

se encontra o *link* para acesso ao repositório com os artefatos do trabalho e por fim, na Seção 8, as conclusões e trabalhos futuros são discutidos.

2. Fundamentação Teórica

Nesta seção, são apresentados os conceitos fundamentais para a compreensão deste trabalho, abrangendo arquiteturas de IDSs colaborativos (Seção 2.1) e uma breve taxonomia de ataques de envenenamento conhecidos (Seção 2.2).

2.1. IDSs Colaborativo

Arquiteturas colaborativas de IDSs têm sido investigadas como uma resposta prática à limitação de conhecimento local em ambientes distribuídos, nos quais diferentes nós observam contextos, padrões de tráfego e condições operacionais distintas. Nessa linha, o FL consolidou-se como um paradigma de referência ao permitir treinamento local com compartilhamento de atualizações, viabilizando modelos mais robustos sem centralizar dados brutos. Esse tipo de abordagem tem sido aplicado também em domínios ciberfísicos, como redes assistidas por VANTs, onde restrições de comunicação e mobilidade exigem soluções distribuídas [Chen et al. 2025, Xue et al. 2025].

De forma complementar, a CN representa uma arquitetura colaborativa com foco na autonomia decisória dos nós, na qual a cooperação ocorre por meio da solicitação e troca de recomendações entre pares, especialmente em situações de incerteza ou conflito entre classificadores. Enquanto o FL estrutura a colaboração em torno de atualizações de modelos, a CN opera sobre decisões e conhecimento compartilhado de forma direta entre IDSs, caracterizando uma dinâmica distribuída orientada a consenso e adaptação incremental. A arquitetura e o protocolo operacional das CNs serão detalhados na Seção 4, servindo como base conceitual para a leitura da seção de trabalhos relacionados e para o modelo avaliado neste estudo [Quincozes et al. 2021].

2.2. Ataques de Envenenamento

Entre as possíveis estratégias adversariais, os ataques de envenenamento representam uma ameaça particularmente severa. No envenenamento de modelos, os nós maliciosos alteram intencionalmente seus parâmetros locais para influenciar negativamente outros modelos por meio do compartilhamento desses parâmetros [Feng et al. 2025].

Já os ataques de envenenamento de dados consistem na adulteração dos dados de treinamento de um ou mais modelos, por meio de estratégias como amostras envenenadas, ataques fora da distribuição (*out-of-distribution*) e envenenamento de rótulos [Rodríguez-Barroso et al. 2023].

O envenenamento de amostras se baseia na modificação de instâncias do conjunto de treinamento por meio da inserção de ruído em amostras de uma classe alvo. Similarmente, os ataques fora da distribuição introduzem instâncias fora de limiares conhecidos no domínio original dos dados, em vez de alterar amostras já existentes.

Por fim, no envenenamento de rótulos o adversário não modifica as características de entrada, mas corrompe os rótulos associados às amostras. Uma variação do evenenamento de rótulos é o *label flipping* [Jha et al. 2023]. Ao inverter ou atribuir incorretamente as classes, o atacante pode enviesar de maneira sutil os classificadores locais. Em IDSs,

amostras de tráfego malicioso podem ser rotuladas intencionalmente como amostras de tráfego legítimo ou vice-versa.

Em IDSs colaborativos, os ataques de envenenamento podem ser potencializados, espalhando a desinformação entre modelos e comprometendo severamente a capacidade do IDS de distinguir corretamente entre comportamentos normais e anômalos.

3. Trabalhos Relacionados

A literatura recente tem demonstrado que arquiteturas colaborativas de detecção e aprendizado distribuído ampliam a superfície de ataque, tornando-se vulneráveis a estratégias de envenenamento conduzidas por participantes maliciosos. Nesta seção, são discutidos trabalhos que investigam ataques e mecanismos de mitigação em cenários centralizados e descentralizados, com ênfase em como o comportamento colaborativo pode ser explorado para degradar modelos e decisões.

Em arquiteturas descentralizadas, Feng et al. 2025 propuseram o *Decentralized Model Poisoning Attack* (DMPA), no qual múltiplos nós maliciosos atuam de forma coordenada para degradar modelos locais e comprometer o desempenho global do sistema. De forma complementar, Tan et al. 2023 apresentaram o *Collusive Model Poisoning Attack* (CMPA), que explora conluio entre participantes para reduzir a acurácia e retardar a convergência, ajustando dinamicamente o nível de perturbação do envenenamento para permanecer dentro de limites estatisticamente plausíveis. Em conjunto, esses estudos reforçam que a descentralização não elimina o risco adversarial, podendo inclusive favorecer ataques coordenados.

No contexto centralizado de aprendizado federado, Lavaur et al. 2025 investigaram ataques de *label-flipping* em IDSs, mostrando que os impactos variam conforme a distribuição dos dados entre participantes, sendo especialmente severos em cenários heterogêneos. Já Liu et al. 2023 propuseram um ataque baseado na manipulação da seleção de *features*, modificando apenas atributos de maior relevância para reduzir a discrepância entre modelos e evadir mecanismos de detecção. Esses trabalhos indicam que, mesmo sem alterações explícitas nos dados brutos, estratégias de envenenamento podem degradar o aprendizado colaborativo e comprometer a detecção.

Em cenários ciberfísicos com VANTs, da Silva and Branco 2025 propuseram o IDS colaborativo REMY, baseado em aprendizado federado, demonstrando a viabilidade de colaboração para detecção em enxames. Por outro lado, Chen et al. 2025 e Xue et al. 2025 exploraram *frameworks* de FL para redes assistidas por drones, propondo técnicas para reduzir comunicação e mitigar envenenamento, incluindo filtragem de atualizações e mecanismos de detecção de modelos adversariais. Esses estudos evidenciam que o domínio de VANTs é especialmente sensível a ataques, exigindo soluções colaborativas robustas e com baixo *overhead*.

Em síntese, os trabalhos revisados demonstram que ataques de envenenamento são altamente eficazes em arquiteturas colaborativas, inclusive em variantes descentralizadas e com nós em conluio. Entretanto, a maioria dessas contribuições está fundamentada no paradigma de aprendizado federado, no qual a colaboração ocorre por meio da troca de gradientes ou parâmetros de modelos, frequentemente sob mecanismos de agregação ou sincronização. Em contraste, nas Redes de Conselhos (CNs), o elemento compartilhado

é a decisão na forma de conselho rotulado, que é incorporado ao aprendizado local e pode se propagar recursivamente ao longo da rede. Assim, embora a literatura ofereça evidências consolidadas sobre o risco do envenenamento em colaboração, ainda há uma lacuna específica na investigação de como esse tipo de ataque se manifesta e se amplifica em CNs, motivando a proposta e avaliação conduzidas neste trabalho.

4. Rede de Conselheiros

Em uma rede de conselhos [Quincozes et al. 2019, Quincozes et al. 2021], referida neste trabalho como CN, cada IDS atua como um conselheiro autônomo, com suporte a múltiplos algoritmos classificadores, que é responsável por monitorar continuamente o tráfego de sua rede local e tomar decisões de forma independente (Seção 4.1).

No entanto, quando um nó possui conflitos (decisões conflitantes) entre classificadores ou quando ele experimenta insuficiência de conhecimento para a tomada de decisão, este nó pode solicitar conselhos para outros pares (Seção 4.2).

4.1. Decisão Local

As decisões locais dos IDSs pertencentes a uma rede de conselhos são fundamentadas no desempenho histórico dos classificadores de um nó em diferentes contextos, obtido a partir da segmentação prévia das amostras conhecidas por meio de técnicas de *clustering*. Essa segmentação permite identificar regiões distintas do espaço amostral que contemplam tráfego normal e diferentes tipos de ataques, associando a cada *cluster* um histórico consolidado de métricas de desempenho, como a F1-Score.

Quando uma nova amostra desconhecida é observada, o sistema identifica o *cluster* correspondente e seleciona dinamicamente o classificador com melhor desempenho esperado naquele contexto. Se apenas um classificador for selecionado, sua decisão é considerada soberana e utilizada diretamente. No entanto, quando os históricos de desempenho dos classificadores são estatisticamente equivalentes — *i.e.*, situam-se dentro de um limiar predefinido, como F1-Score $\pm$ 5% — o sistema pode combinar múltiplos classificadores para a tomada de decisão local. Nesse caso, se dois ou mais classificadores forem escolhidos, suas saídas são comparadas. Se houver divergência entre as classificações, um conselho deve ser solicitado à rede de conselheiros (Seção 4.2). Da mesma forma, se nenhum classificador apresentar desempenho histórico suficiente — *i.e.,* F1-Score abaixo de um limiar definido, como $< 75\%$ — para aquele *cluster*, ou se a amostra estiver muito distante do centróide, caracterizando um possível *outlier* ou ataque *zero-day*, o nó local também recorre à rede de conselhos, sem tomar uma decisão autônoma.

4.2. Protocolo de Troca de Conselhos

A Figura 1 demonstra o protocolo de troca de conselhos para a resolução de conflitos. Ao solicitar um conselho, o Nó Requisitante envia a amostra conflitante a um de seus pares selecionados. O IDS Conselheiro que recebe a solicitação repete internamente os mesmos passos de análise, incluindo a clusterização da amostra, a seleção dinâmica de classificadores e a verificação de conflitos.

Caso esse conselheiro não experimente conflitos internos, ele envia uma mensagem de *conselho* ao nó requisitante a fim de compartilhar a sua decisão (Figura 1a). Essa

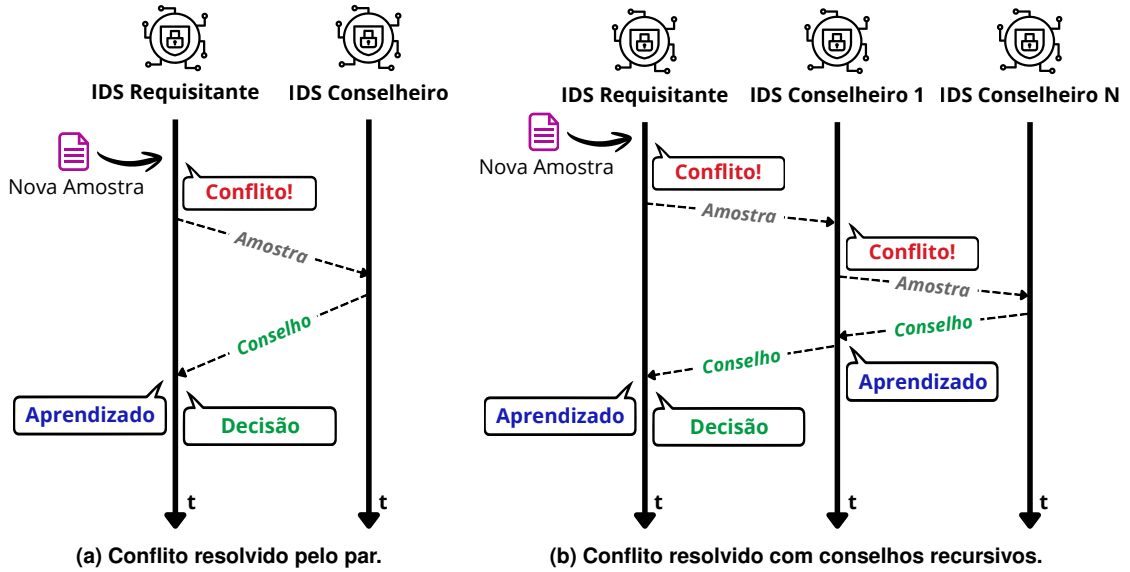


Figura 1. Protocolo de troca de conselhos em uma CN.

decisão se dá a partir do modelo local do IDS Conselheiro que foi consultado, construído por meio de seu próprio conjunto de dados de treinamento.

Caso o IDS Conselheiro que recebe uma solicitação também enfrente um conflito, um novo pedido de conselho é encaminhado recursivamente à rede. Quando uma decisão é alcançada sem conflitos em algum nó, o resultado é propagado de forma recursiva até o IDS que iniciou a solicitação, permitindo que todos os nós intermediários incorporem o conhecimento adquirido no processo colaborativo (Figura 1b). Cada IDS utiliza o rótulo recebido para atualizar sua base de dados local e reconstruir seu modelo de decisão, promovendo aprendizado distribuído.

Quando nenhum IDS da rede é capaz de resolver o conflito, o último nó da CN detecta o encerramento do ciclo e retorna uma mensagem *LOOP_CLOSED* ao IDS requisitante original. Esse mecanismo é viabilizado por uma lista de rastreamento que é incrementada a cada conselheiro intermediário que apresenta conflito e encaminha a solicitação. Ao identificar que todos os nós conhecidos já constam nessa lista, o IDS Conselheiro final reconhece o fechamento do ciclo e sinaliza, de forma recursiva, a impossibilidade de resolução colaborativa.

5. Modelo de Ameaça

Dentre os ataques adversariais, existem formas já conhecidas de envenenamento que visam prejudicar as decisões de algoritmos de aprendizado de máquina. No contexto de IDSs colaborativos, esse envenenamento tem seu efeito potencializado e pode mascarar o comportamento de outros tipos de ataques, tornando-os indetectáveis pelos IDSs vítimas. Na Seção 2.2, os principais ataques de envenenamento são revisados. Em seguida, na Seção 5.1 um novo modelo de ameaças derivado dos existentes é proposto.

5.1. Novo Modelo de Ameaça Proposto: Envenenamento de Conselhos

O modelo de ameaça de envenenamento de conselhos é uma proposta original deste trabalho. Ele é definido diretamente sobre o funcionamento operacional da CN descrito na

seção anterior. Assume-se que o atacante possui a capacidade de comprometer os classificadores pertencentes a um subconjunto A de IDSs da rede, denotado por $\mathcal{A} \subset \{1, \ldots, N\}$, fazendo com que esses nós passem a emitir conselhos maliciosos, mantendo comportamento aparentemente legítimo no restante do sistema.

O atacante não interfere diretamente no processo de treinamento local, na clusterização inicial das amostras ou na execução dos classificadores locais dos nós não comprometidos. Em vez disso, sua atuação concentra-se exclusivamente na camada colaborativa, explorando os pontos do fluxo decisório nos quais a solicitação de conselhos para outros nós da CN é acionada.

Conforme o funcionamento do sistema discutido na Seção 4, um pedido de conselho é disparado quando um IDS enfrenta uma das seguintes condições: (i) conflito entre classificadores com histórico de desempenho equivalente para um cluster de amostras similares dentro de um mesmo nó, (ii) ausência de classificadores com desempenho histórico suficiente para o cluster identificado em um nó, ou (iii) detecção de uma amostra distante do centróide do cluster de um nó, caracterizando um possível *outlier*. O modelo de ameaça explora precisamente esses cenários de incerteza, nos quais a decisão local é suspensa e delegada a pares participantes da CN e depois aprendida e incorporada nos modelos locais de todos os nós que estavam na cadeia de pedidos de conselho.

Quando um nó que implementa um IDS legítimo i solicita um conselho, ele encaminha a amostra $x^{(t)}$ a um nó conselheiro j. Caso $j \in \mathcal{A}$, o conselheiro comprometido executa internamente as mesmas etapas de clusterização e seleção dinâmica de classificadores, porém manipula deliberadamente o resultado final do processo decisório. Essa manipulação pode assumir diferentes formas operacionais, incluindo: (i) inversão sistemática da classe prevista, (ii) seleção intencional de um classificador subótimo para o cluster identificado, ou (iii) retorno de uma decisão enviesada apenas em regiões específicas do espaço de amostras, de modo a preservar plausibilidade estatística. Neste trabalho, será empregado um parâmetro que define a probabilidade do conselheiro malicioso inverter o rótulo do conselho a ser prestado ao requisitante (*i.e.,* substituindo um rótulo *normal* por *ataque* ou *vice-versa*).

Uma vez recebido o conselho adulterado, o IDS de origem aceita a decisão como legítima, utiliza-a para resolver o conflito inicial e incorpora a amostra rotulada à sua base de dados local. Esse passo é crítico, pois o conselho envenenado passa a influenciar diretamente a atualização do modelo local, alterando estatísticas de desempenho por cluster, critérios de seleção de classificadores e decisões futuras. Esse efeito colateral é propagado recursivamente, sempre que IDSs conselheiros enfrentam conflitos e encaminham as requisições recebidas do IDS Requisitante a outros pares conselheiros.

6. Estudo de Caso: Rede de VANTs como Instanciação do Modelo

Para instanciar o modelo descrito, este trabalho adota como estudo de caso um cenário de Veículos Aéreos Não Tripulados (VANTs), ilustrado na Figura 2. O sistema é composto por VANTs, estações de controle e uma camada de detecção de intrusões colaborativa organizada como CN. Cada VANT possui sua estação de controle, na qual possui um IDS sendo executado correspondendo a um nó da CN, cada IDS é responsável por monitorar a comunicação entre estação de controle e VANT.

Dessa forma, esses IDSs observam sinais associados às comunicações e ao con-

trole dos veículos, tais como o posicionamento, velocidade, nível de carga da bateria, além de mensagens de comando e controle e características do enlace. Esses sinais constituem as amostras locais $x_i^{(t)}$, processadas por IDSs para identificar comportamentos anômalos ou maliciosos.

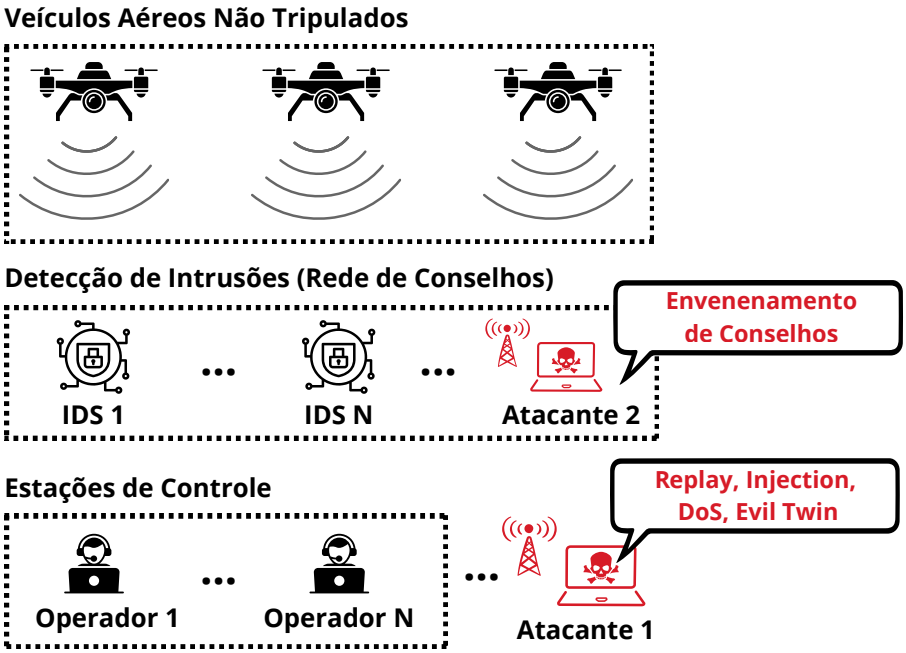


Figura 2. Modelo de ameaça em uma Rede de Conselhos: Ataques na camada de sistema e envenenamento de conselhos na camada de detecção.

Ataques clássicos na camada de sistema, incluindo *replay*, *injection*, *Denial of Service (DoS)* e *Evil Twin*, podem afetar diretamente a infraestrutura de comunicação entre VANTs e estações de controle. Esses ataques impactam a disponibilidade e a integridade do sistema monitorado e são tratados como parte do ambiente operacional. Por esse motivo, a camada colaborativa de detecção de intrusões é essencial. Entretanto, a partir do modelo de ameaças que tem como alvo essa camada, o foco deste estudo recai sobre a exploração de vulnerabilidades nessa camada colaborativa de detecção de intrusões.

Assim, de modo a observar de forma controlada os efeitos do envenenamento de conselhos sobre a convergência e estabilidade dos modelos locais, bem como mensurar como conselhos envenenados afetam a eficácia global do sistema, são realizados experimentos práticos seguindo três cenários:

- Cenário 0 (*baseline*): representa uma situação onde ataques *zero-day* são experienciados por um IDS; neste cenário, o IDS não está conectado a uma CN, portanto não tem acesso a conhecimentos externos;
- Cenário 1: IDSs de uma CN possuem diferentes contextos em seus bancos de dados de amostras e trocam conselhos a fim de otimizar as suas métricas;
- Cenário 2: Replica-se o Cenário 1, mas considerando a presença de um nó comprometido com a taxa de envenenamento de conselho de 50% – *i.e.,* a cada conselho, esse nó tem 50% de chance de responder com um rótulo incorreto.
- Cenário 3: Replica-se o Cenário 2, mas com a taxa de envenenamento de 100%.

É importante observar que, embora o estudo de caso apresentado seja específico para o domínio de VANTs, o modelo de ameaça permanece aplicável a outros domínios distribuídos, servindo como uma instância concreta para análise e validação experimental.

7. Experimentos e Resultados

Para avaliar os efeitos de ataques de envenenamento de conselhos, foram conduzidos experimentos empíricos com o objetivo de analisar o comportamento de um IDS, denominado Nó 1, pertencente a uma CN composta por três nós, na qual Nó 2 e Nó 3 atuam como conselheiros. A principal característica dos cenários de experimentações consiste na existência de ataques *zero-day* e na necessidade de colaboração entre IDSs. Os artefatos como código-fonte e conjuntos de dados utilizados estão publicamente disponíveis no repositório do GitHub[1].

7.1. Configuração Geral: Ataques Zero Day em VANTs

Os experimentos consideram o cenário de classificação de ataques conhecidos (*Replay*) e amostras benignas pelo Nó 1, bem como de ataques desconhecidos (*zero-day*). Esses ataques, por sua vez, são conhecidos pelos nós 2 e 3 e rotulados como *False Data Injection (FDI)* e *Evil Twin*. Adicionalmente, os nós 2 e 3 foram treinados com amostras benignas e do ataque *Denial of Service*, de modo a diversificar sua base de conhecimento. Para tanto, foi utilizado o conjunto de dados apresentado em [Hassler et al. 2024], composto por amostras de parâmetros cibernéticos e físicos de um veículo aéreo não tripulado (VANT) em operação normal e sob esses ataques. A distribuição das amostras entre os IDSs foi realizada conforme apresentado na Tabela 1.

Tabela 1. Distribuição de amostras por classe nos conjuntos de dados

Dataset	Total	Benigno	Replay	Evil Twin	FDI	DoS
Nó 1 (Treino)	2.325	1.300	1.025	0	0	0
Nó 1 (Teste)	1.000	500	0	250	250	0
Nó 2 (Treino)	17.871	1.159	10.981	5.183	548	0
Nó 3 (Treino)	21.062	6.466	0	250	2.675	11.671

A Figura 3 mostra o desempenho dos classificadores implantados no IDS do Nó 1 *standalone* (*i.e.,* antes de sua participação na CN). As matrizes de confusão indicam que, no cenário *zero-day*, todos os classificadores desse nó apresentam limitações significativas na identificação dos ataques *zero-day*. Essa configuração foi concebida especialmente para i) demonstrar a necessidade e efetividade de uma CN em cenários *zero-day* e o ii) impacto que ataques de envenenamento de conselhos podem causar nessas situações.

A Árvore de Decisão demonstra comportamento parcialmente discriminativo, porém com elevada confusão entre amostras benignas e ataques, especialmente *FDI* e *Evil Twin*, evidenciando incapacidade de generalização adequada. Em contraste, os classificadores K-vizinhos mais próximos, Naive Bayes e Máquina de Vetores de Suporte colapsam, classificando todas as amostras como benignas, inclusive amostras de ataques. Esse comportamento revela uma falha crítica na detecção de padrões não observados durante o treinamento, resultando em taxas de falso negativo elevadas e caracterizando uma

[1] `https://github.com/sequincozes/CounselorNode`

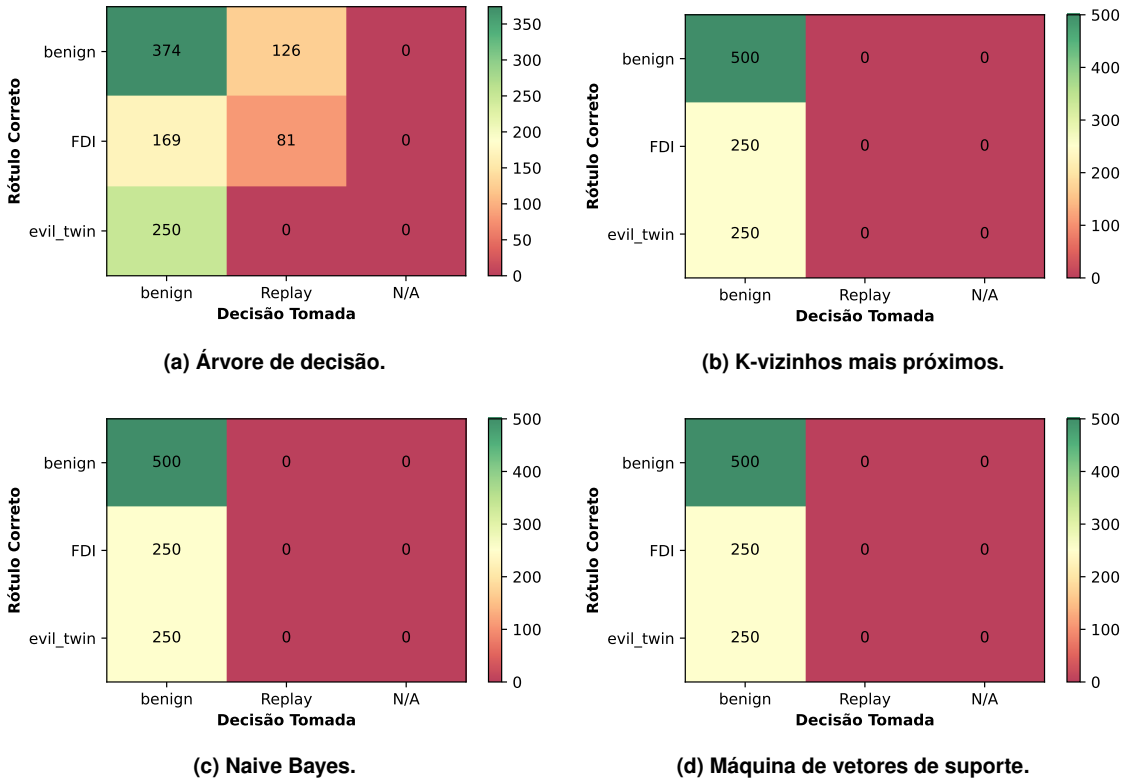


Figura 3. Matrizes de confusão no cenário zero-day no Nó 1 (standalone).

situação de alto risco operacional para o Nó 1, que passa a depender fortemente do mecanismo colaborativo da rede de conselhos para mitigar os efeitos dos ataques zero-day.

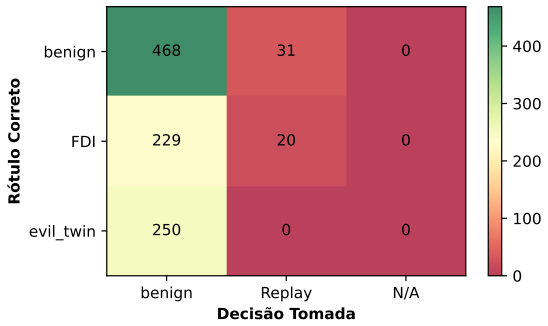
7.2. Rede de Conselhos e Ataques de Envenenamento

A Figura 4 mostra a comparação dos quatro experimentos realizados: o cenário *baseline*, sem rede de conselhos e os outros três cenários implementados para medir o desempenho no Nó 1 ao ingressar na CN e seu colapso ao se tornar vítima do envenenamento dos conselheiros representados pelos nós 2 e 3, ao serem comprometidos.

Sem conselhos, o Nó 1 acertou menos da metade do total de 1000 amostras, atribuindo corretamente 468 das 500 amostras da classe *Benign*. Sem conhecimento prévio das classes *Evil Twin* e *FDI*, do total de 500 amostras de ambos os ataques, 480 não foram detectadas; outras 20 foram detectadas como comportamentos maliciosos, porém classificadas em classes de ataques distintas das classes reais. Foram aferidas métricas de acurácia com 46,8%, precisão com 16,5%, *recall* com 31,3% e F1-Score com 21,6%.

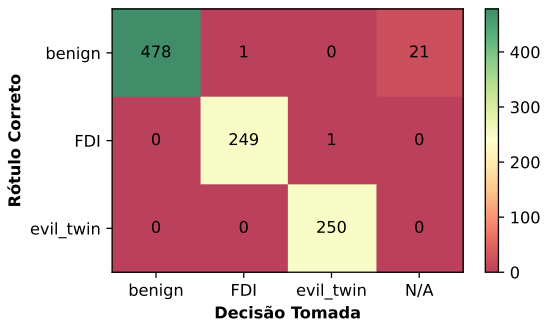
No Cenário 1, ao conectar o Nó 1 à CN, conflitos são imediatamente detectados e o mecanismo colaborativo é acionado. Com os conselhos, o Nó 1 aprende o padrão do ataque *Evil Twin* e classifica corretamente suas 250 amostras. Para o ataque *FDI*, são classificadas corretamente 249 de 250 amostras. Na análise das amostras benignas são 478 classificações corretas entre 500 amostras. Ao todo, foram classificadas corretamente 977 das 1000 amostras, uma acurácia de 97,7%, precisão de 99,7%, *recall* de 98,4% e F1-Score de 99,0%.

Em seguida, no Cenário 2, com os nós 2 e 3 comprometidos com um algoritmo



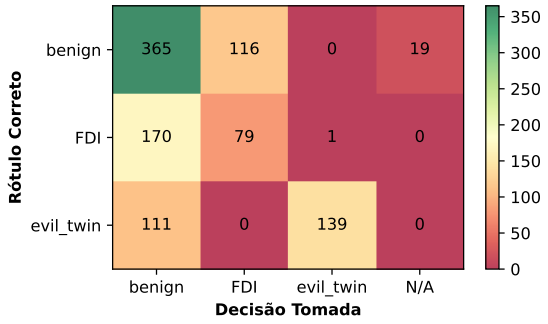
(a) Zero-day sem a rede de conselhos.[a]

[a]Esta matriz é a média dos 4 classificadores da Figura 3. As decisões locais limitam-se a: benign e replay (eixo X). Evil_twin e FDI são desconhecidos.



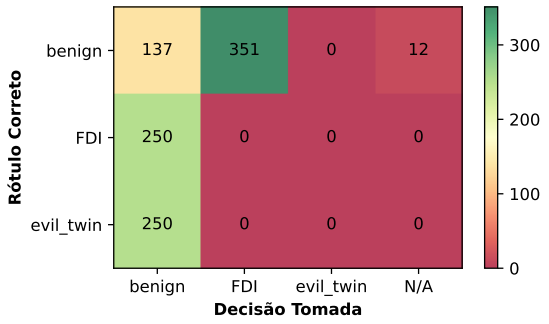
(b) Zero-day com rede de conselhos.[a]

[a]Ao solicitar conselhos, FDI e evil_twin são conhecidos. Ataques de replay (rótulo já conhecido) não fizeram parte das amostras avaliadas.



(c) Conselhos envenenados (50%).[a]

[a]Mesmo cenário de (b), mas com 50% de probabilidade de um conselho recebido estar envenenado (benign ao invés de ataques, FDI ao invés de benign).



(d) Conselhos envenenados (100%).[a]

[a]Mesmo cenário de (b), mas com 100% de probabilidade de um conselho recebido estar envenenado (benign ao invés de ataques, FDI ao invés de benign).

Figura 4. Comparação das matrizes de confusão para diferentes cenários. N/A representa conflitos não solucionados.

malicioso que emprega uma taxa de 50% de envenenamento nos conselhos enviados, nota-se que o sistema não apenas falha em detectar ataques, mas apresenta confusão cruzada entre as ameaças, como o caso de *Evil Twin* classificado como *FDI*. A acurácia global declinou para 58,3%, a precisão para 65,4%, o *recall* para 53,4% e o F1-Score para 58,8%.

Por fim, no Cenário 3, o cenário de envenenamento total, o sistema sofreu um colapso total na identificação de ameaças desconhecidas, sendo possível observar o fenômeno da camuflagem total, onde ataques reais são absorvidos pela classe benigna. As métricas globais sofreram uma queda abrupta, diminuindo para 13,7% de acurácia, 7,2% de precisão, 9,1% de recall e 8,0% de F1-Score.

7.3. Tempo de Processamento e Desempenho Geral

Para garantir que as exatas condições estavam acessíveis para todos os nós, a CN como um todo foi executada virtualmente em um único ambiente físico constituído por um servidor com processador *AMD Ryzen 5 4600H* (3.00 GHz) e 24 GB de memória RAM DDR4. Portanto, de forma intencional para este trabalho, o atraso de rede foi desconsiderado.

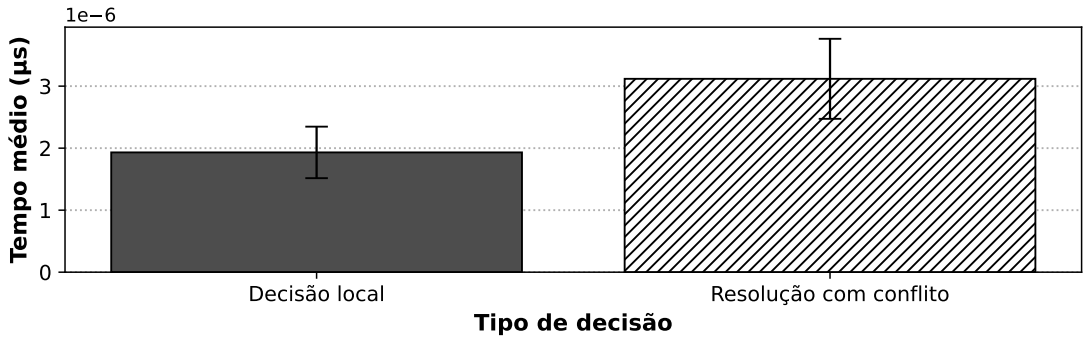


Figura 5. Tempo médio de decisão local e resolução de conflito via CN (I.C. 99%).

Assim, conforme ilustrado na Figura 5, foram computados somente o tempo médio e o desvio padrão para um nó executar uma decisão local (sem conflitos) e o tempo que os nós levam para resolver um conflito entre classificadores através da troca de conselhos. Como esperado, a troca de conselhos introduz um atraso médio adicional, mas esse atraso tem impacto marginal, situado na ordem de microssegundos, o que indica que o overhead introduzido pela troca de conselhos é computacionalmente pouco significativo.

A Figura 6 sumariza de forma comparativa o desempenho do IDS nos 4 experimentos: na situação de *baseline* com o Nó 1 atuando sem conselhos e nos 3 cenários de conselhos com as diferentes taxas de envenenamento.

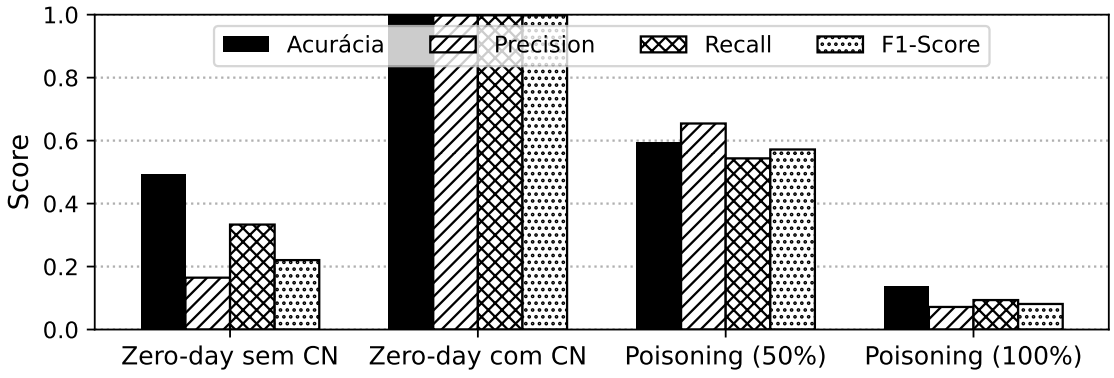


Figura 6. Comparação das métricas de desempenho.

De forma geral, os resultados evidenciam duas lições centrais. Primeiro, a CN é altamente eficaz para mitigar limitações de generalização em cenários *zero-day*, permitindo que um nó com conhecimento restrito (Nó 1) aprenda rapidamente classes desconhecidas por meio da troca de conselhos, alcançando desempenho próximo ao ideal. Segundo, o mesmo mecanismo que viabiliza essa adaptação contínua também constitui um ponto crítico de dependência: uma vez que o aprendizado local incorpora conselhos como rótulos válidos, a qualidade da colaboração passa a ser determinante para a estabilidade do processo de decisão e para a manutenção do desempenho ao longo do tempo.

8. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma análise dos impactos de ataques de envenenamento de rótulos em IDSs baseados em CN. Os experimentos conduzidos em um conjunto de dados

contendo amostras de uma rede de VANTs, a partir de três cenários distintos, demonstraram que o cenário de operação legítima alcançou valores de acurácia, precisão, recall e F1-Score de 97,7%, 99,7%, 98,4% e 99,0%, respectivamente, sendo considerados satisfatórios, enquanto nos cenários de envenenamento tais métricas foram afetadas de forma inversamente proporcional ao aumento da taxa de envenenamento, chegando a valores de 13,7%, 7,2%, 9,1% e 8,0%, respectivamente, no cenário de envenenamento total dos conselhos. Dessa forma, é possível concluir que o envenenamento dos conselhos degrada a confiança e a eficácia da detecção colaborativa de maneira severa, validando a principal hipótese deste estudo.

Como contramedidas e trabalhos futuros, pretende-se investigar mecanismos de robustez específicos para CNs, com foco em reduzir a influência de conselhos inconsistentes ou potencialmente maliciosos. Entre as direções previstas estão: (i) modelagem de confiança e reputação entre nós, com base no histórico de acertos e consistência dos conselhos fornecidos; (ii) estratégias de validação cruzada do conselho, consultando múltiplos pares e aplicando regras de consenso antes da incorporação do rótulo ao aprendizado local; e (iii) mecanismos de quarentena para amostras rotuladas via conselho, evitando sua incorporação imediata ao conjunto de treinamento até que evidências adicionais sejam acumuladas. Essas contramedidas são particularmente promissoras, pois podem ser integradas ao protocolo de troca de conselhos sem introduzir overhead computacional significativo, conforme sugerido pelos resultados de tempo de processamento obtidos.

Agradecimentos

Este trabalho foi desenvolvido no âmbito do projeto IoTEdu, conduzido pelo Grupo de Trabalho IoTEdu da Rede Nacional de Ensino e Pesquisa (RNP), com apoio das agências CAPES, CNPq, FAPERJ e do United States Department of Energy sob o número de concessão DE-CR0000039.

Referências

Belenguer, A., Garcia-Teodoro, P., Diaz-Verdejo, J., and Maciá-Fernández, G. (2025). Federated learning for intrusion detection systems: A comprehensive review. *Computer Networks*, 241:110204.

Chen, L., Zhai, W., Bu, X., Sun, M., and Zhu, C. (2025). A lightweight robust training method for defending model poisoning attacks in federated learning assisted uav networks. *Drones*, 9(8):528.

da Silva, L. M. and Branco, K. R. (2025). Collaborative intrusion detection system for unmanned aerial vehicles swarm security. In *Concurso de Teses e Dissertações (CTD)*, pages 134–143. SBC.

Feng, C., Li, Y., Gao, Y., Celdrán, A. H., von der Assen, J., Bovet, G., and Stiller, B. (2025). Dmpa: Model poisoning attacks on decentralized federated learning for model differences. *arXiv preprint arXiv:2502.04771*.

Hassler, S. C., Mughal, U. A., and Ismail, M. (2024). Cyber-physical intrusion detection system for unmanned aerial vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 25(6):6106–6117.

Jha, R. D., Hayase, J., and Oh, S. (2023). Label poisoning is all you need. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lavaur, L., Busnel, Y., and Autrel, F. (2025). Investigating the impact of label-flipping attacks against federated learning for collaborative intrusion detection. *Computers & Security*, 156:104462.

Liu, Z., Liu, Z., and Yang, X. (2023). Poisoning attack based on data feature selection in federated learning. In *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 106–110. IEEE.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, Fort Lauderdale, FL, USA. PMLR.

Quincozes, S. E., dos Santos, C. R. P., Nunes, R. C., de Albuquerque, C. V. N., Passos, D. G., and Mossé, D. (2019). A counselors-based intrusion detection architecture. In Ziviani, A., de Albuquerque, C. V. N., and Moraes, I. M., editors, *9th Latin American Network Operations and Management Symposium (LANOMS 2019), Niterói, Rio de Janeiro, Brazil, September 25–27, 2019*, pages 1–8. IFIP.

Quincozes, S. E., Raniery, C., Ceretta Nunes, R., Albuquerque, C., Passos, D., and Mossé, D. (2021). Counselors network for intrusion detection. *International Journal of Network Management*, 31(3):e2111.

Rodríguez-Barroso, N., Stipcich, G., Vicent, J., Binos, E., Cazorla, M., Marín, L., Serrano, J., and Lobo, J. L. (2023). Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 89:1–38.

Tan, S., Hao, F., Gu, T., Li, L., and Liu, M. (2023). Collusive model poisoning attack in decentralized federated learning. *IEEE Transactions on Industrial Informatics*, 20(4):5989–5999.

Vasilomanolakis, E., Karuppayah, S., Mühlhäuser, M., and Fischer, M. (2015). A taxonomy and survey of collaborative intrusion detection systems. *ACM Computing Surveys*, 47(4):1–33.

Xue, L., Zhong, L., Zhang, J., Chen, Z., Cheng, H., and Li, J. (2025). Decentralized federated learning for adversarial anomaly detection in consumer-grade uav-assisted mec systems. *IEEE Transactions on Consumer Electronics*, 71(4):10797–10811.

Yang, R., He, H., Wang, Y., Qu, Y., and Zhang, W. (2023). Dependable federated learning for IoT intrusion detection against poisoning attacks. *Computers & Security*, 132:103381.