Um Arcabouço de Classificação em Conjunto Aberto com Adaptação via Retreino Dinâmico para Detecção de Intrusão

Giovanna Vieira Souza[1], **Fernando Nakayama**[1], **Michele Nogueira**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

{giovanna.vieira, fernandonakayama, michele}@dcc.ufmg.br

Abstract. *The effectiveness of network attack detection is often compromised by the nature of closed-set classifiers, which are intrinsically incapable of handling unknown threats. Open-set classification allows for the identification of new samples to mitigate this problem; however, traditional models operate statically, failing to incorporate knowledge from new discoveries and, consequently, suffering from performance degradation over time. This paper proposes a semi-supervised open-set classification framework aiming to overcome this limitation, structured with a short-term module for immediate response and a long-term module for continuous learning. In the latter, unknown instances are analyzed to identify new classes and, as a central contribution, the model is incrementally retrained. The proposed framework was evaluated on two distinct datasets subjected to stability, plasticity, and robustness tests, with final post-retraining accuracies reaching 89.88%.*

Resumo. *A eficácia da detecção de ataques em redes é frequentemente comprometida pela natureza dos classificadores de conjunto fechado, intrinsecamente incapazes de lidar com ameaças desconhecidas. Para mitigar esse problema, a classificação em conjunto aberto viabiliza a identificação de amostras desconhecidas. Contudo, os modelos tradicionais operam de forma estática, falhando em incorporar conhecimento das novas descobertas e, consequentemente, sofrendo degradação de desempenho ao longo do tempo. Com o objetivo de superar tal limitação, este trabalho apresenta um arcabouço semi-supervisionado de classificação em conjunto aberto, estruturado com um módulo de curto prazo para resposta imediata e um outro de longo prazo para aprendizado contínuo. Neste último, as instâncias desconhecidas são analisadas para a identificação de novas classes e, como contribuição central, o modelo é retreinado incrementalmente. O arcabouço proposto foi avaliado em dois conjuntos de dados distintos, submetidos a testes de estabilidade, plasticidade e robustez, com acurácias finais após o retreino chegando a 89,88%.*

1. Introdução

A detecção eficaz de anomalias e ataques em redes de computadores constitui um pilar fundamental da cibersegurança moderna [Passoni 2024]. Contudo, este domínio é perpetuamente desafiado pela natureza evolutiva e dinâmica das ameaças cibernéticas. As abordagens tradicionais de classificação, tais como Regressão Logística e Árvores de Decisão, operando sob a premissa de conjunto fechado, são intrinsecamente restritas a um

conjunto finito de padrões de tráfego de redes conhecido, fator que as torna inadequadas para cenários operacionais reais, onde a emergência de novos vetores e padrões de ataque é constante [Dahanayaka et al. 2023]. Como resposta a essa lacuna, a classificação em conjunto aberto emergiu, introduzindo a capacidade crucial de identificar e rejeitar instâncias "desconhecidas", ou seja, aquelas que não pertencem a nenhuma das classes de treinamento predefinidas [Geng et al. 2020].

Apesar do avanço que o reconhecimento em conjunto aberto representa, suas implementações atuais adotam, majoritariamente, uma lógica operacional estática. Tais soluções limitam-se a rejeitar amostras desconhecidas, sem a capacidade de incorporar o conhecimento adquirido sobre elas [Geng et al. 2020]. O contínuo surgimento de ataques leva à rápida obsolescência dos modelos, visto que o crescente volume de amostras não classificadas compromete o desempenho da detecção e invalida a base de conhecimento [Agrahari and Singh 2021]. A solução convencional (o retreinamento completo) configura um ciclo operacionalmente insustentável e evidencia a necessidade de transcender o paradigma estático, urgindo o desenvolvimento de sistemas de reconhecimento em mundo aberto que possuem a capacidade de aprender autonomamente com novas classes.

Os avanços recentes na literatura têm focado em maximizar a robustez da detecção de tráfego desconhecido. No domínio de redes, o método RoNeTC [Geng et al. 2025] aumenta a confiabilidade da rejeição ao quantificar a incerteza da decisão com a distribuição de *Dirichlet*. Outras abordagens, como o TrafficGPT [Ginige et al. 2024], exploram o uso de LLMs para otimizar a extração de características sequenciais e melhorar o desempenho em conjunto aberto. Contudo, uma limitação central dessas metodologias é que elas ainda descartam o tráfego desconhecido. Começando a endereçar essa lacuna, o paradigma *Reason and Discovery* [Fu et al. 2025] propõe a clusterização das amostras desconhecidas para identificar novas categorias, porém, carece de um arcabouço autônomo para pseudo-rotulagem e retreinamento. Essa lacuna é crítica, pois transfere a responsabilidade de análise para um humano, o que se transforma em um gargalo operacional impraticável. O arcabouço autônomo que se faz necessário é, portanto, capaz não apenas de descobrir novos padrões, mas de avaliar, rotular e integrar esse conhecimento ao modelo.

Como contribuição para solucionar esse desafio, propõe-se um arcabouço semi-supervisionado de natureza cíclica. A estrutura inicia com uma etapa de classificação em conjunto aberto; subsequentemente, as instâncias rejeitadas como desconhecidas passam por processos de filtragem, clusterização e rotulação. Diferente das abordagens atuais, o arcabouço proposto utiliza Aprendizado Incremental para o retreinamento autônomo, abordagem que fundamenta-se na técnica de Repetição de Experiência, utilizando um buffer de memória balanceado que combina exemplares das classes previamente aprendidas com os novos dados detectados. A atualização do modelo é realizada através de um ajuste fino ponderado, estratégia desenhada para abrandar o esquecimento catastrófico e assegurar a preservação do conhecimento preexistente enquanto assimila as novas classes.

Para avaliar a eficácia da estrutura apresentada, a avaliação de desempenho é conduzida através de um protocolo de avaliação de forma incremental. Este protocolo particiona o conjunto de dados em um conjunto base e uma sequência de tarefas subsequentes, simulando a emergência temporal de novas classes. A cada etapa, o desempenho do sistema é aferido em três dimensões principais: (1) Estabilidade, medindo a retenção de conhecimento; (2) Plasticidade, avaliando o aprendizado de novas classes; e (3) Robustez,

monitorada pela capacidade de rejeição contra classes ainda desconhecidas.

Este artigo procede como segue. A Seção 2 apresenta os trabalhos relacionados à classificação em conjunto aberto. A Seção 3 descreve o método proposto. A Seção 4 discute a avaliação e os experimentos realizados. Por fim, a Seção 5 reúne as conclusões e as considerações finais deste trabalho.

2. Trabalhos Relacionados

A ineficácia das abordagens tradicionais de classificação de tráfego, baseadas em portas ou carga útil dos pacotes, impulsionou a adoção de métodos de *Machine Learning* e *Deep Learning* focados em padrões estatísticos [Geng et al. 2025]. Por consequência, essa linha de pesquisa evoluiu até o uso de Modelos de Linguagem de Grande Escala (LLMs), como demonstrado por [Zhou et al. 2024]. Em seu trabalho, os autores propuseram um *pipeline* que pré-processa os dados de rede em um formato textual e utiliza *fine-tuning* do GPT-3.5-turbo para a classificação. Contudo, essa metodologia opera fundamentalmente sob a premissa de conjunto fechado, ou seja, o modelo é treinado em um número fixo de classes, sendo incapaz de gerenciar tráfego de categorias não vistas. Essa limitação é a motivação central para a Classificação em Conjunto Aberto (*Open Set Classification - OSR*), um problema formalizado por [Scheirer et al. 2013], no qual é postulado que um classificador em ambiente real opera com conhecimento incompleto do mundo, sendo capaz de rejeitar com precisão instâncias desconhecidas.

As metodologias do tipo OSR focam, portanto, na rejeição robusta. Neste contexto, a robustez é definida pela capacidade de minimizar o risco de espaço aberto, assegurando que o modelo mantenha fronteiras de decisão compactas para as classes conhecidas e maximize a incerteza para dados que caem fora dessas regiões. As soluções de ponta como o RoNeTC [Geng et al. 2025] buscam confiabilidade ao quantificar a incerteza da decisão via distribuição de *Dirichlet*, permitindo um limiar de rejeição mais estável. Em paralelo, o método *TrafficGPT* [Ginige et al. 2024] explora o uso de LLMs em bytes brutos para criar representações de características mais densas, melhorando a separabilidade e a detecção de *outliers*. Entretanto, a despeito de sua alta performance, ambas as abordagens tornam-se obsoletas ao ignorar o potencial de aprendizado com as amostras desconhecidas, optando apenas pelo descarte direto.

O problema da obsolescência estática refere-se à degradação progressiva da competência do classificador em ambientes dinâmicos: ao tratar novas ameaças recorrentes meramente como rejeições, o modelo falha em atualizar sua representação do mundo, tornando-se ineficaz diante da evolução natural do tráfego. Este desafio é central no Reconhecimento em Mundo Aberto (OWR) [Bendale and Boult 2015], que exige que os modelos aplicados do mundo real não apenas rejeitem desconhecidos, mas também os detecte como novidades e os incorpore incrementalmente. Nessa direção, embora [Liu et al. 2022] avance ao identificar a necessidade de gerenciar o tráfego desconhecido, sua solução foca na filtragem para evitar a expansão semântica. Essa estratégia de exclusão, embora proteja as classes conhecidas, trata o dado desconhecido apenas como ruído a ser descartado, ignorando o potencial de atualização incremental. Indo um pouco além do mero descarte, o paradigma *Reason and Discovery (RD)* [Fu et al. 2025] introduz um ciclo de descoberta onde as amostras rejeitadas são clusterizadas para identificar novas categorias. Embora represente um avanço conceitual, o *framework* de RD se limita à des-

coberta das classes e falha em estabelecer um ciclo autônomo de aprendizado, além de necessitar de um analista humano para lidar com a rotulação dos agrupamentos.

A literatura demonstra, portanto, um consenso sobre a necessidade de superar os classificadores estáticos. Porém, as soluções atuais não fecham o ciclo, parando na mera descoberta de novidades [Fu et al. 2025] ou dedicando-se apenas a refinar a rejeição [Geng et al. 2025, Ginige et al. 2024, Liu et al. 2022]. Assim, a lacuna crítica é a falta de um mecanismo autônomo que assimile essas novas classes ao modelo principal, lacuna esta que este trabalho propõe preencher por meio de um arcabouço cíclico e semi-supervisionado. O fluxo é iniciado pela classificação em conjunto aberto, após a qual as amostras desconhecidas são filtradas, agrupadas e rotuladas. Por fim, o Aprendizado Incremental utiliza essas amostras para retreinar e atualizar o modelo.

3. Arcabouço de classificação

Esta seção detalha o arcabouço proposto denominado de AIRA (Arcabouço Incremental de Reconhecimento em Conjunto Aberto). Diferente de métodos de classificação estáticas, o AIRA estabelece uma infraestrutura sistêmica e modular que busca reduzir a obsolescência do conhecimento, gerenciando os novos dados de intrusão através de um ecossistema de aprendizado contínuo. Conforme ilustrado na Figura 1, o fluxo de processamento tem início no Conjunto de Dados, que alimenta o ciclo com fluxos de rede. O arcabouço é estruturado em duas vertentes distintas, uma de curto prazo e outra de longo prazo. A Vertente de Curto Prazo (VCP) é focada na prontidão de resposta e classificação imediata, enquanto a Vertente de Longo Prazo (VLP) governa a etapa de descoberta e constitui o núcleo de resiliência da proposta. De modo geral, o arcabouço pode ser condensado em quatro módulos: módulo de classificação em conjunto aberto, de processamento e clusterização, de evolução incremental e de inteligência de rótulos. O módulo de classificação em conjunto aberto faz parte da Vertente de Curto Prazo e os demais módulos contemplam a Vertente de Longo Prazo.

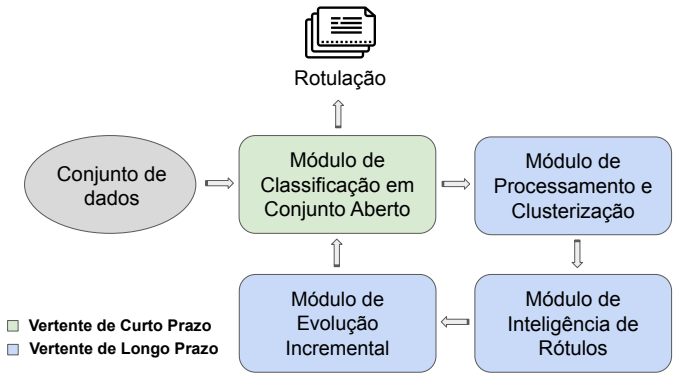

Figura 1. Arcabouço Incremental de Reconhecimento em Conjunto Aberto

O Módulo de Classificação em Conjunto Aberto é situado na VCP, sendo o primeiro ponto de contato com o conjunto de dados. Ele atua como um filtro inteligente que utiliza métricas de incerteza para distinguir entre tráfego conhecido (rotulado imediatamente) e amostras desconhecidas. Sua função é garantir que o sistema não force uma classificação errônea em ataques inéditos, encaminhando-os para o buffer de retenção. Em seguida, o Módulo de Processamento e Clusterização, já integrado à VLP, realiza a

limpeza das anomalias acumuladas. Para tal, são utilizados algoritmos de filtragem para remover ruídos e falsos positivos, seguidos de técnicas de agrupamento que identificam a estrutura geométrica das novas ameaças no espaço latente, descobrindo quantos novos tipos de ataques estão presentes na rede.

O Módulo de Inteligência de Rótulos atua na conversão do conhecimento não supervisionado em supervisionado, sendo responsável pela atribuição de pseudo-rótulos aos agrupamentos validados e pela transformação das anomalias em classes estruturadas. Este processo elimina a necessidade de intervenção humana constante, permitindo que o arcabouço gere seu próprio conjunto de dados de atualização. Por fim, o Módulo de Evolução Incremental é a última fase de atualização dos pesos da rede neural. Através de técnicas de expansão arquitetural e Repetição de Experiência, as ameaças são integradas à base de conhecimento evitando o esquecimento catastrófico. Após a atualização, o conhecimento é retroalimentado ao módulo inicial, completando o ciclo e capacitando o AIRA a identificar ameaças previamente desconhecidas.

3.1. Módulo de Classificação em Conjunto Aberto

Como porta de entrada do AIRA, o Módulo de Classificação em Conjunto Aberto tem a função de criar uma fronteira de decisão discriminativa, atuando como um filtro para distinguir se o tráfego é familiar ou se representa uma novidade de mundo aberto. Diferente de sistemas tradicionais, este módulo não apenas classifica dados, mas governa o fluxo de informações entre a resposta imediata (VCP) e o aprendizado evolutivo (VLP). A entrada do módulo consiste em tráfego de rede bruto, representado na forma de pacotes individuais. Para que a rede neural possa processar esses dados, o tráfego é convertido em um vetor de características numéricas $\vec{x}$, que captura atributos críticos como o tamanho dos pacotes e a frequência de acesso às portas de destino. Na implementação do AIRA, esses dados sofrem normalização logarítmica e são agregados estatisticamente por porta. Essa abordagem permite modelar o comportamento dinâmico do fluxo sem inspeção profunda de conteúdo, garantindo eficiência mesmo sobre tráfego criptografado.

O processamento interno ocorre através da instanciação de uma rede neural profunda (MLP). Esta rede atua como uma extratora de características que projeta o vetor de entrada $\vec{x}$ em um espaço latente, produzindo como saída intermediária um vetor de logits z(x). Os logits, definidos como as ativações brutas da última camada linear pré-normalização, são componentes críticos na arquitetura do arcabouço. Ao preservarem a magnitude absoluta das ativações, eles evitam a compressão de informação típica de funções probabilísticas, permitindo uma distinção mais eficaz entre classes conhecidas e anomalias. Diferente de abordagens tradicionais que dependem da função **Softmax**, a qual tende a gerar altas probabilidades mesmo para amostras desconhecidas, o módulo utiliza o paradigma do *Energy Score*. Esta métrica interpreta os logits z(x) como estados de energia termodinâmica, calculada pela equação:

$$E(\mathbf{x}) = -T \cdot \log \sum_{k=1}^{K} e^{z_k(\mathbf{x})/T}$$

Nesta formulação, o parâmetro de temperatura T desempenha um papel crucial na regulação da suavidade da distribuição de energia. Em termos de rede, T calibra a sensibilidade do modelo: valores adequados permitem que o sistema mapeie com precisão a

densidade dos dados conhecidos em estados de baixa energia (alta compatibilidade), enquanto dados anômalos e desconhecidos resultam naturalmente em estados de alta energia (baixa compatibilidade estatística). A saída do módulo é governada por uma lógica de decisão binária baseada em um limiar L, calibrado dinamicamente através do percentil de scores das classes conhecidas. O fluxo final de dados é definido por duas condições:

1. Se $E(x) \leq L$: A amostra possui energia compatível com o domínio conhecido. O módulo atribui o rótulo da classe correspondente e finaliza a operação na Vertente de Curto Prazo.
2. Se $E(x) > L$: A amostra é identificada como uma anomalia de mundo aberto. Em vez de ser descartada ou classificada erroneamente, ela é desviada para o buffer de retenção. Este componente funciona como uma memória temporária que acumula o tráfego desconhecido para que o arcabouço possa, posteriormente, descobrir novos padrões e evoluir através da Vertente de Longo Prazo.

3.2. Módulo de Processamento e Clusterização

O buffer de retenção, alimentado pelas instâncias rejeitadas no módulo anterior, apresenta uma natureza intrinsecamente ruidosa, consistindo em uma mistura heterogênea entre ameaças inéditas genuínas e falsos positivos. Para assegurar a integridade da evolução do arcabouço, é implementado um mecanismo de depuração estatística fundamentado no algoritmo *Isolation Forest*. Este componente atua no estágio inicial do fluxo de processamento da Vertente de Longo Prazo (VLP), tratando os dados acumulados que foram desviados do fluxo de resposta imediata por apresentarem alta incerteza. A escolha deste algoritmo fundamenta-se na premissa de que novos padrões de tráfego, sejam eles fluxos maliciosos ou novas aplicações legítimas, manifestam-se como agrupamentos densos no espaço de características, devido à natureza repetitiva e estruturada dos protocolos de rede [Rocha and Silva 2020]. Em contraste, erros de classificação e ruídos tendem a ser eventos dispersos e esparsos. Ao explorar o isolamento recursivo de observações, o arcabouço descarta os *outliers* e preserva os *inliers*, garantindo que apenas sinais com massa estatística relevante sejam encaminhados para a fase de descoberta.

Após a depuração, o arcabouço aciona o seu componente de particionamento latente, focado na estruturação e categorização das amostras filtradas. O objetivo central é a desagregação da massa de dados em agrupamentos que representem tipos de ataques funcionalmente distintos, utilizando para isso uma implementação do algoritmo *K-Means*. O funcionamento baseia-se na minimização da variância intra-cluster: ele define centros de gravidade (centroides) no espaço de características e associa cada amostra ao centroide mais próximo, decisão feita por meio do cálculo da distância euclidiana. Este processo é repetido iterativamente até que a posição dos centros se estabilize, particionando a massa de dados em grupos que representam tipos de ataques funcionalmente distintos.

Diante da incerteza inerente sobre a quantidade exata de novas ameaças presentes no fluxo de rede, o sistema implementa uma lógica de otimização dinâmica para a definição da cardinalidade dos grupos. Diferente de aplicações estáticas, a governança do arcabouço executa o agrupamento de forma iterativa, avaliando o hiperparâmetro K dentro do intervalo de busca ($K \in [1, 11]$), conforme ilustrado no Algoritmo 1. Para cada iteração, o sistema mensura o *Silhouette Score*, selecionando automaticamente o valor de K que maximiza a relação entre coesão interna e separação externa, definindo assim a granularidade ideal para a expansão do conhecimento do modelo.

Algoritmo 1: Busca de K-Ótimo

Input: Buffer de amostras filtradas ($Inliers$), Intervalo $K_{max} = 11$
Output: Modelo de clusterização otimizado e valor de K-ótimo

$S_{best} \leftarrow -1$;
$K_{best} \leftarrow 1$;
for $K \leftarrow 2$ **to** K_{max} **do**
    Instanciar K-Means com K centros;
    Ajustar modelo às amostras filtradas ($Inliers$);
    $S_{current} \leftarrow$ Calcular Silhouette_Score(K);
    if $S_{current} > S_{best}$ **then**
        $S_{best} \leftarrow S_{current}$;
        $K_{best} \leftarrow K$;
        $Modelo_{Final} \leftarrow Modelo_{Atual}$;
    end
end
return $Modelo_{Final}, K_{best}$;

3.3. Módulo de Inteligência de Rótulos

Este módulo atua como o elo de conversão entre a descoberta não supervisionada e a evolução supervisionada do modelo, funcionando como o núcleo de gestão semântica do arcabouço. Diferente de arquiteturas convencionais que dependem de intervenção humana e introduzem latência operacional para a nomeação de ataques, o AIRA integra uma estratégia de automação de pseudo-rótulos autônoma, cuja lógica de execução é detalhada no Algoritmo 2. As entradas consistem nas amostras filtradas e nos identificadores de agrupamentos gerados na etapa anterior, iniciando um processamento que dispara a validação de cada agrupamento C_i com base em critérios de densidade e separabilidade, tratando os que satisfazem tais critérios como uma nova classe semântica legítima.

Algoritmo 2: Expansão de Espaço Semântico e Pseudo-rotulação

Input: Grupos_Validados $\{C_1, \ldots, C_n\}$, Y_{known}, Amostras_Historico
Output: Conjunto_Final rotulado, Y_{known} atualizado

$K \leftarrow$ tamanho(Y_{known});
$Conjunto_{Novidades} \leftarrow \emptyset$;
foreach *agrupamento* $C_i \in Grupos\_Validados$ **do**
    $ID_{novo} \leftarrow y_{\text{new}} =$ nova_classe($K + 1$);
    $Y_{known} \leftarrow Y_{known} \cup \{\text{Atq\_Emerg\_} + ID_{novo}\}$;
    foreach *amostra* $x \in C_i$ **do**
        Vincular x ao rótulo ID_{novo};
    end
    $Conjunto_{Novidades} \leftarrow Conjunto_{Novidades} \cup C_i$;
    $K \leftarrow K + 1$;
end
$Conjunto_{Final} \leftarrow Conjunto_{Novidades} \cup Amostras_{Hist}$;
return $Conjunto_{Final}, Y_{known}$;

Uma vez que o componente de clusterização valida um grupo de amostras, o módulo executa a expansão incremental do espaço de classes (Y_{known}) através de uma operação de registro. Tecnicamente, a governança mantém um dicionário de mapeamento global, logo, se o modelo foi inicialmente treinado com K classes, o arcabouço detecta

o valor de K e reserva os próximos inteiros disponíveis para as novas ameaças. Assim, o primeiro grupo validado recebe o rótulo $y_{\text{new}} = \text{nova\_classe}(K+1)$, o segundo $y_{\text{new}} = \text{nova\_classe}(K+2)$, e assim por diante. Essa atribuição sequencial é crucial para evitar colisões de rótulos e garantir que cada nicho de ataque detectado possua uma identidade única dentro do arcabouço.

Após a definição desses novos identificadores, ocorre a transmutação dos dados. Neste estágio, o módulo percorre as amostras brutas que estavam armazenadas no buffer de retenção e substitui a sua identificação de desconhecido pelo novo pseudo-rótulo. Esse processo não altera as características intrínsecas do pacote, mas vincula o comportamento observado a uma categoria formal e consolida o que o arcabouço define como o conjunto de novidades. Este procedimento transforma o problema de detecção, convertendo anomalias desconhecidas em um conjunto de dados rotulado e pronto para o aprendizado supervisionado. Portanto, essa transformação permite que, no próximo módulo, o classificador não apenas rejeite o dado desconhecido, mas assimile formalmente suas características comportamentais, integrando-o à base de conhecimento permanente.

3.4. Módulo de Evolução Incremental

Após a caracterização e rotulação das novas classes, o AIRA aciona o Módulo de Evolução Incremental, responsável por integrar o conhecimento recém-adquirido à rede neural principal. Este processo é governado pela resolução do Dilema da Plasticidade-Estabilidade, uma tensão inerente entre a capacidade de assimilar novos padrões e a de reter informações antigas. O funcionamento deste equilíbrio é vital para prevenir o Esquecimento Catastrófico, fenômeno no qual o ajuste de pesos para novas tarefas sobrescreve as sinapses responsáveis pelo reconhecimento das classes veteranas.

Logo, a modificação ocorre exclusivamente na Camada de Saída Expandida. A Figura 2 demonstra que as conexões neuronais associadas às classes previamente consolidadas (Classe 1 e Classe 2) são tratadas como Pesos Fixos, sendo efetivamente congeladas. Na prática, isto significa que, durante a retropropagação, o Fluxo do Gradiente é direcionado apenas para os neurônios da Nova Classe, impedindo que a atualização dos pesos interfira nas fronteiras de decisão das classes fixas. Simultaneamente, a camada linear é expandida dimensionalmente, inicializando novas conexões específicas para cada novo agrupamento validado, o que assegura que as novas regiões do espaço latente sejam mapeadas sem invadir os domínios estáveis.

Com o intuito de complementar a preservação estrutural, o arcabouço implementa o paradigma de Repetição de Experiência para gerenciar o fluxo de dados. Para evitar o enviesamento do gradiente por dados recentes, o sistema constrói um conjunto de dados híbrido de ajuste fino, fundindo as novas amostras com uma reserva estratégica de dados históricos recuperados do buffer de memória. Ao reintroduzir deliberadamente exemplos das classes conhecidas durante o processo de atualização, o arcabouço força a rede neural a reafirmar as suas fronteiras de decisão originais, impedindo que a densidade das novas ameaças degrade o desempenho nas classes veteranas.

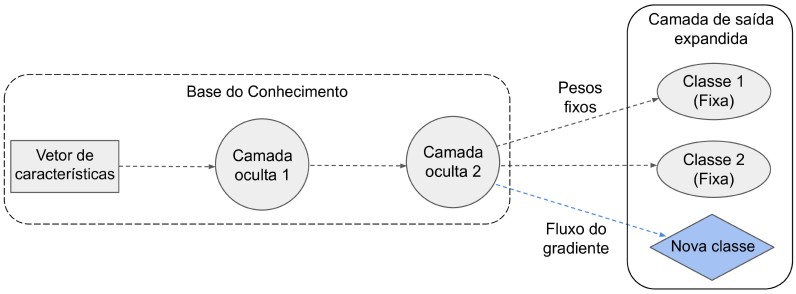

Figura 2. Adaptação Estrutural da Rede Neural

Por fim, o ciclo de atualização encerra-se com um regime de retreino de baixa intensidade, caracterizado por um número reduzido de épocas e uma taxa de aprendizado controlada. Esta configuração de governança assegura uma integração suave, permitindo que a representação latente se ajuste para acomodar múltiplas distribuições de dados sem desestabilizar a consistência dos pesos congelados. Desta forma, o arcabouço AIRA completa o seu ciclo operacional, evoluindo de forma autônoma e mantendo a resiliência frente ao dinamismo das ameaças cibernéticas.

4. Avaliação

A avaliação do arcabouço foi estruturada através de dois conjuntos de dados de referência, o *ToN_IoT* e o *UNSW-NB15*, cujas distribuições de dados estão detalhadas na Tabela 1. O conjunto de dados *ToN_IoT* foi selecionado por representar o tráfego heterogêneo de ambientes de Internet das Coisas (IoT) e Industrial (IIoT), provendo uma telemetria rica que permite ao modelo distinguir entre comportamentos legítimos de dispositivos e ataques multicamadas, enquanto o conjunto de dados *UNSW-NB15* foi integrado para avaliar a robustez do AIRA em tráfego de rede corporativa moderna, oferecendo uma diversidade de ataques que buscam burlar sistemas de detecção convencionais. Ambos os conjuntos foram processados no formato *Parquet* para otimização de memória, sendo os experimentos executados em ambiente de alto desempenho com aceleradores GPU T4.

Dada a volumetria dos dados, com o *ToN_IoT* superando 160 milhões de amostras no conjunto original, a engenharia de atributos concentrou-se na extração de metadados de fluxo de rede, selecionando variáveis estatísticas fundamentais como comprimento dos pacotes, tamanho do *payload*, TTL e informações de endereçamento lógico por meio de portas de origem e destino. Para garantir que a rede neural não fosse influenciada pela disparidade de escalas entre as variáveis, aplicou-se a transformação logarítmica em variáveis de comprimento para reduzir o impacto de *outliers* e assimetria, seguida da padronização global via *StandardScaler*, assegurando que o modelo trate características de diferentes magnitudes com igual peso estatístico e convirja de forma eficiente.

Para a classificação e detecção de anomalias, implementou-se uma arquitetura de rede neural do tipo *Perceptron Multicamadas* (MLP) estruturada com camadas densas de 256 e 128 neurônios, utilizando Batch Normalização e *Dropout* de 0,3 para prevenir o sobreajuste e garantir a estabilidade das ativações, contando ainda com uma camada de saída expansível para a inclusão dinâmica de novos neurônios conforme novas ameaças são identificadas. O protocolo experimental foi operacionalizado através de um *Parquet-DataLoader* customizado, capaz de realizar leituras em blocos e mapear *labels* textuais

Tabela 1. Distribuição do Tráfego nos Conjuntos de Dados

Tipo de Tráfego	UNSW-NB15		TON_IoT	
	Presença	**Porcentagem**	**Presença**	**Porcentagem**
Normal Traffic	✓	87,351%	✓	64,593%
DoS	✓	0,644%	✓	7,564%
DDoS	–	–	✓	12,238%
Password attacks	–	–	✓	2,154%
Injection attack	–	–	✓	1,463%
Vulnerability scanner	–	–	✓	5,522%
XSS attacks	–	–	✓	5,994%
Backdoor	✓	0,092%	✓	0,407%
MITM	–	–	✓	0,004%
Ransomware	–	–	✓	0,062%
Generic	✓	8,483%	–	–
Exploits	✓	1,753%	–	–
Fuzzers	✓	0,955%	–	–
Reconnaissance	✓	0,551%	–	–
Analysis	✓	0,105%	–	–
Shellcode	✓	0,059%	–	–
Worms	✓	0,007%	–	–

para identificadores numéricos em um cenário de mundo aberto.

O fluxo de avaliação foi segmentado em três frentes principais, iniciando pela análise de *Robustez Out-of-Distribution (OOD)*, que utiliza o *Energy Score* para avaliar a capacidade do modelo em rejeitar ataques desconhecidos via métrica AUROC. Em seguida, a plasticidade do sistema foi testada para medir a acurácia na classificação de novas ameaças após a expansão da rede, enquanto a estabilidade foi monitorada através da retenção do conhecimento prévio. Para mitigar o esquecimento catastrófico durante o retreino incremental, utilizou-se a técnica de Repetição de Experiência, em que um buffer de memória armazena amostras críticas de tarefas passadas, permitindo que o modelo evolua de forma contínua sem comprometer a integridade do aprendizado anterior.

4.1. Experimentos

Os experimentos foram estruturados em três fases cíclicas integradas, compreendendo o treinamento em ambiente fechado, a descoberta de anomalias e a adaptação incremental. Essa sequência foi aplicada sobre sobre os conjuntos de dados *ToN_IoT* (Experimentos 1 e 2) e *UNSW-NB15* (Experimento 3). No primeiro cenário, focado no *ToN_IoT* buscou-se garantir o rigor estatístico através de uma rede MLP de três camadas configurada com uma taxa de aprendizado inicial de 0,001, decaimento automático via *scheduler* e a aplicação de *Label Smoothing* em 0,1 para a calibração das predições. O alicerce deste experimento residiu em um pré-processamento minucioso, que utilizou a normalização logarítmica de comprimentos de pacotes e a engenharia de atributos estatísticos para capturar o comportamento dinâmico e heterogêneo da rede. Na fase inicial, o modelo foi treinado com as classes conhecidas, onde o tráfego legítimo detinha a maioria absoluta perante ameaças já familiares ao AIRA.

A transição para o cenário de mundo aberto testou a eficácia do fluxo ao introduzir os ataques desconhecidos DDoS e XSS. Durante à VLP, as amostras foram submetidas

à clusterização, resultando na formação de um único agrupamento devido à alta similaridade estatística das assinaturas de rede desses ataques. A validação final do fluxo foi consolidada pelo equilíbrio entre plasticidade e estabilidade, utilizando a técnica de Repetição de Experiência para mitigar o esquecimento catastrófico e permitir que a arquitetura evoluísse de forma autônoma sem reprocessar o conjunto de dados completos.

Dando continuidade à investigação no *ToN_IoT*, o segundo experimento foi projetado para testar a sensibilidade do modelo em um cenário de maior complexidade diagnóstica, mantendo a estrutura cíclica de treinamento e adaptação incremental, mas restringindo a base de conhecimento inicial para desafiar a capacidade de discriminação do sistema. Nesta configuração, as classes conhecidas estabeleceram um patamar de reconhecimento que, ao ser submetido ao regime de mundo aberto, enfrentou a introdução simultânea dos ataques DDoS e Scanning. O sistema utilizou novamente o monitoramento de incerteza via *Energy Score* para identificar anomalias, porém, o diferencial residiu na fase de descoberta e análise geométrica do espaço latente. O objetivo central foi verificar se o modelo seria capaz de distinguir diferentes tipos de ataques que ocorrem ao mesmo tempo, em vez de tratá-los como um bloco único de tráfego malicioso.

Diferente do teste anterior, onde os ataques foram unificados, a separação geométrica entre as novas ameaças foi bem-sucedida devido ao aproveitamento de assinaturas de rede contrastantes. Considerando a divergência estatística entre os padrões de DDoS e Scanning, o algoritmo de clusterização foi capaz não apenas de detectar as novidades, mas de separá-las em grupos específicos. O equilíbrio entre os pilares do aprendizado contínuo foi evidenciado pela manutenção da integridade do conhecimento original enquanto a rede assimilava rapidamente as novas classes, reafirmando a resiliência da arquitetura frente a variantes de intrusão que exigem separação clara de assinaturas comportamentais.

Por fim, para a validação do AIRA no cenário corporativo, o experimento sobre o conjunto de dados *UNSW-NB15* foi projetado para testar a resiliência do arcabouço em um ambiente de rede moderna, marcado pela alta heterogeneidade e por um desbalanceamento extremo de classes. O protocolo experimental foi estruturado com uma base de conhecimento inicial composta pelas classes Generic e Backdoor, enquanto a transição para o regime de mundo aberto introduziu as classes desconhecidas Exploits e Fuzzers. Esta configuração testou a sensibilidade do sistema frente a intrusões que buscam mimetizar o tráfego legítimo ou explorar vulnerabilidades de forma sequencial. Durante a fase de descoberta, o arcabouço enfrentou o desafio de processar amostras minoritárias imersas em uma volumetria colossal de dados, o que reforçou a importância da aplicação do algoritmo *Isolation Forest* para garantir a integridade do aprendizado.

4.2. Resultados

Conforme apresentados na Tabela 2, os resultados do Experimento 1 demonstram a eficácia da arquitetura MLP na assimilação de ameaças em ambientes dinâmicos. Na fase inicial, o arcabouço atingiu uma acurácia global de 89,66%, validando a precisão da rede na classificação das cinco classes conhecidas (Normal, Scanning, Password, Injection e MITM). Ao transitar para o cenário de mundo aberto, o mecanismo de *Energy Score* permitiu uma separação clara do tráfego inédito, registrando uma métrica AUROC de robustez de 77,71% e uma taxa de rejeição de 52,70%. Por fim, a validação final do fluxo incremental revelou um equilíbrio entre plasticidade e estabilidade: a capacidade de assimilação da nova classe atingiu 98%, enquanto a acurácia nas classes antigas manteve-se

estável em 84,42% via Repetição de Experiência, resultando em uma acurácia final do retreino consolidada de 89,88%.

Tabela 2. Resultados dos Experimentos

	Exp. 1	**Exp. 2**	**Exp. 3**
Classificação em Conjunto Aberto			
Acurácia Global (Classes Conhecidas)	89,66%	83,49%	97,02%
Taxa de Rejeição	52,70%	60,08%	47,10%
AUROC	77,71%	67,51%	60,46%
Retreino dinâmico			
Plasticidade (Novas Classes)	98,00%	98,09%	92,44%
Estabilidade (Classes Antigas)	84,42%	87,15%	98,73%
Acurácia Final	**89,88%**	**75,03%**	**84,17%**

Os resultados do Experimento 2 detalham o desempenho do sistema em um cenário de maior complexidade diagnóstica. Nesta configuração, o uso do *Energy Score* para monitorar a incerteza do modelo alcançou uma taxa de rejeição de 60,08% e uma acurácia de classificação em conjunto aberto de 83,49%, embora a métrica AUROC tenha se fixado em 67,51%. Ao contrário do observado anteriormente, a distinção geométrica entre as ameaças foi alcançada, permitindo que o algoritmo de clusterização segmentasse as classes DDoS e Scanning em grupos bem definidos. Na fase de adaptação incremental, o arcabouço demonstrou alta resiliência, apresentando uma estabilidade de 87,15% na preservação do conhecimento original e uma plasticidade de 98,09% na integração das novas classes. Apesar da maior carga computacional de gerenciar múltiplas categorias simultâneas, que situou a acurácia global de retreino em 75,03%, o experimento confirmou a capacidade do arcabouço de evoluir de forma autônoma e granular.

Os resultados obtidos no Experimento 3 consolidam a eficácia do AIRA sob condições de severo desbalanceamento. Na fase inicial de monitoramento em conjunto fechado, o modelo atingiu uma acurácia de 97,02% na classificação das categorias conhecidas Ao transitar para o cenário de mundo aberto, o mecanismo de *Energy Score* registrou uma Taxa de Rejeição de 47,10% para as classes desconhecidas, amostras que resultaram na formação de dois agrupamentos. A validação final revelou que a estabilidade no conhecimento acumulado se manteve em 98,73%, enquanto a plasticidade atingiu 92,44% na assimilação das novas classes, culminando em uma acurácia final consolidada de 84,17%.

4.3. Discussão

A análise comparativa dos experimentos revela comportamentos distintos do espaço latente frente às diferentes naturezas das ameaças. No Experimento 1, observou-se que a clusterização resultou na formação de um único agrupamento para os ataques DDoS e XSS. Essa unificação ocorreu devido à alta similaridade estatística das assinaturas de rede desses ataques; embora sejam logicamente distintos na camada de aplicação, seus comportamentos em nível de transporte e fluxo apresentaram características de entropia e frequência anômalas muito próximas. Para o sistema de detecção, essa convergência representou uma simplificação eficiente, permitindo o tratamento da novidade como uma categoria única sem a necessidade de fragmentar o aprendizado.

Em contrapartida, no Experimento 2, a separação geométrica entre as novas ameaças foi bem-sucedida devido ao aproveitamento de assinaturas de rede contrastantes entre DDoS e Scanning. Enquanto o primeiro se caracteriza por uma inundação volumétrica de pacotes, o segundo manifesta-se através da exploração sequencial de portas e serviços, criando padrões distintos de dispersão de dados [Fernandes 2008]. Essa distinção permitiu que o algoritmo de clusterização segmentasse as novidades em agrupamentos independentes com alta eficácia, provando que o fluxo incremental não apenas detecta que algo é novo, como também consegue categorizar múltiplas ameaças simultâneas desde que estas possuam comportamentos estatísticos suficientemente divergentes.

No cenário do Experimento 3, a dificuldade reside na sutil diferenciação entre as assinaturas de ataque e o tráfego de fundo, tornando a calibração do *Energy Score* uma tarefa crítica. Diferente de conjuntos de dados mais datados e simplistas, o *UNSW-NB15* reflete uma rede moderna onde o tráfego legítimo é altamente heterogêneo, criando um ruído de fundo que frequentemente sobrepõe as métricas de incerteza do modelo. Apesar disso, diferente da convergência observada em cenários menos ruidosos, onde ataques diferentes tendem a se agrupar em um único bloco, a análise geométrica revelou uma distinção clara entre as ameaças Exploits e Fuzzers, resultando na formação de dois agrupamentos independentes e bem definidos.

Essa separação é um indicativo da robustez do AIRA em extrair características fundamentais sob pressão. Enquanto o ataque Exploits busca comprometer falhas específicas de software através de sequências lógicas e direcionadas, o ataque Fuzzers opera sob o princípio do bombardeio de dados malformados e aleatórios [Jochem et al. 2018]. No espaço latente, essa divergência comportamental traduz-se em perfis de entropia contrastantes: os Exploits formam estruturas mais densas e previsíveis, enquanto os Fuzzers dispersam-se em padrões de alta incerteza. A capacidade do arcabouço em não apenas detectar, mas categorizar essas novidades como entidades distintas, mesmo imersas em um oceano de dados desbalanceados, reafirma a resiliência do arcabouço.

Portanto, em todos os cenários avaliados, a habilidade de segmentar as novidades em agrupamentos (sejam eles distintos ou unificados) comprovou que o ciclo de retreino consegue gerenciar múltiplas intrusões de forma autônoma. Esse processo sustenta-se no equilíbrio fundamental entre os pilares do aprendizado contínuo: a plasticidade necessária para absorver novos comportamentos e a estabilidade requerida para preservar o conhecimento prévio. Essa harmonia consolida a resiliência do modelo perante variantes de intrusão, considerando que exigem tanto uma separação clara de assinaturas quanto uma adaptação sustentável em ambientes de alta densidade de dados.

5. Conclusão

Este trabalho aborda o desafio crítico da detecção de ameaças em cenários de mundo aberto, onde o surgimento constante de novos ataques cibernéticos supera a capacidade de atualização dos classificadores estáticos tradicionais. Nesse sentido, para avançar a literatura, este artigo propõe o AIRA, um arcabouço de aprendizado incremental baseado em monitoramento de incerteza via *Energy Score* e descoberta de anomalias por clusterização. A eficácia da solução foi validada por experimentos nos conjuntos de dados distintos, demonstrando que o modelo é capaz de isolar e categorizar múltiplas intrusões simultâneas, mantendo um equilíbrio resiliente entre plasticidade e estabilidade. Para pes-

quisas futuras, vislumbra-se o aprimoramento da robustez do sistema frente a conjuntos de dados com desbalanceamento extremo, além da expansão da escalabilidade do modelo para gerenciar um volume maior de classes simultaneamente desconhecidas, ampliando assim a aplicabilidade prática do ciclo.

Referências

Agrahari, S. and Singh, A. K. (2021). Concept drift detection in data stream mining: A literature review. *Journal of King Saud University - Computer and Information Sciences*, 34:9523–9540.

Bendale, A. and Boult, T. (2015). Towards open world recognition. *Conference on Computer Vision and Pattern Reocgnition*, páginas 1893–1902.

Dahanayaka, T., Ginige, Y., Huang, Y., Jourjon, G., and Seneviratne, S. (2023). Robust open-set classification for encrypted traffic fingerprinting. *Comp. Networks*, 236:1–15.

Fernandes, G. (2008). Detecção e classificação de anomalias no tráfego de redes de computadores. Trabalho de conclusão de curso, Universidade Federal de Santa Catarina.

Fu, Y., Liu, Z., and Lyu, J. (2025). Reason and discovery: A new paradigm for open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(7):5586–5599.

Geng, C., Huang, S.-j., and Chen, S. (2020). Recent advances in open set recognition: A survey. *IEEE Trans. Netw. Service Manag.*, 43(10):3614–3631.

Geng, C., Huang, S.-j., and Chen, S. (2025). Reliable open-set network traffic classification. *IEEE Trans. Inf. Forensics Security*, 20(10):2313–2328.

Ginige, Y., Dahanayaka, T., and Seneviratne, S. (2024). TrafficGPT: An llm approach for open-set encrypted traffic classification. In *Proceedings of the 19th Asian Internet Engineering Conference*, p. 26–35. Association for Computing Machinery.

Jochem, I., Andreoni, M., Antonio, G., and Carlos, O. (2018). Um sistema de detecção de ameaças distribuídas de rede baseado em aprendizagem por grafos. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, páginas 1187–1200.

Liu, Y.-C., Ma, C.-Y., Dai, X., Tian, J., Vajda, P., He, Z., and Kira, Z. (2022). Open-set semi-supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Passoni, L. M. (2024). Detecção de anomalias em redes de computadores e equipamentos iot. In *Engenharia de Computação: Inovação Digital e Desenvolvimento Tecnológico*, páginas 338–356. Editora Científica Digital.

Rocha, M. and Silva, D. (2020). Detecção de tráfego anômalo de rede utilizando clusterização em big data. *Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, páginas 1–5.

Scheirer, W. J., Rocha, A., Sapkota, A., and Boult, T. E. (2013). Towards open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1757–1772.

Zhou, H., Huang, X., and Deng, L. (2024). Enhancing network traffic classification with large language models. *IEEE International Conf. on Big Data*, páginas 7282–7281.