# Estruturação dos Conjuntos de Dados

A preparação dos conjuntos de dados CIC-IDS2018, CIC-IDS2017 e UNSW-NB15, escolhidos para os experimentos, foi uma etapa essencial para garantir a comparabilidade entre os experimentos de generalização dos modelos de detecção de intrusão. Esses três conjuntos de dados, amplamente utilizados na literatura, possuem atributos distintas, refletindo a diversidade nos cenários de captura de tráfego de rede. Entretanto, para viabilizar os testes de generalização, foi necessário padronizar os atributos em cada conjunto, de forma a garantir que os modelos fossem treinados e avaliados nos diferentes cenários: *Intraset* (único domínio de conjunto de dados) e *Interset* (múltiplos domínios de conjuntos de dados).

Os três conjuntos de dados são disponibilizados originalmente em arquivos no formato *PCAP (Packet Capture)*, que armazenam o tráfego de rede em nível de pacotes. A partir desses arquivos, tornou-se possível processar e extrair informações mais detalhadas sobre os fluxos de rede, por ser uma abstração comum para análises de modelos de detecção de intrusão. Para essa tarefa, a ferramenta *NFStream* foi empregada. Trata-se de uma biblioteca de análise de tráfego que permite processar arquivos *PCAP*, extrair fluxos de rede e organizar os dados com atributos consistentes, o que foi essencial para alinhar os três conjuntos de dados de acordo com um conjunto comum de atributos.

Através do uso do *NFStream*, os fluxos de rede foram extraídos de maneira eficiente e estruturados com os mesmos atributos, incluindo informações como endereço *IP* de origem e destino, protocolo, duração do fluxo, número de pacotes transmitidos, volume de dados trocados, entre outros. Essa padronização foi um passo crítico para garantir que as intervenções aplicadas nos experimentos pudessem ser comparadas e avaliadas em cenários distintos. Ao alinhar os três conjuntos de dados, foi possível testar a robustez e a capacidade de generalização dos modelos de detecção de intrusão, mesmo quando expostos a tráfegos de rede capturados em ambientes distintos.

Em seguida, os conjuntos de dados passaram por um pré-processamento que incluiu a remoção de valores nulos e duplicados, a codificação de valores categóricos e a rotulagem manual das instâncias de classe. No caso específico de cada conjunto, os fluxos maliciosos foram organizados em categorias de ataque, facilitando a análise das instâncias e a comparação entre diferentes tipos de ameaças. Para simplificar a identificação, os conjuntos de dados originais CIC-IDS2018, CIC-IDS2017 e UNSW-NB15 que passaram pelo processo de preparação e adequação no contexto desta pesquisa, foram renomeados como GenIDS-CIC18, GenIDS-CIC17 e GenIDS-NB15, respectivamente. A Tabela 1 a seguir apresenta os detalhes dos conjuntos de dados ajustados.

Table 1: Descrição dos Conjuntos de Dados.

| Conjunto de dados | Número de fluxos extraídos | Número de fluxos após o pré-processamento | Percentual de fluxos normais após o pré-processamento | Percentual de fluxos maliciosos após o pré-processamento | Número de atributos* | Categorias de ataques | Categorias de ataques removidos |
|---|---|---|---|---|---|---|---|
| GenIDS-CIC18 | 4.461.855 | 4.461.855 | 83,5% | 16,5% | 70 | Bruteforce, Botnet, DDoS, DoS, Infiltration | WebAttack |
| GenIDS-CIC17 | 1.845.604 | 1.845.604 | 85,9% | 14,1% | 70 | Bruteforce, Botnet, DDoS, DoS, PortScan, WebAttack | Infiltration |
| GenIDS-NB15 | 2.023.965 | 576.965 | 86,7% | 13,3% | 70 | Exploits, Fuzzers, Reconnaissance, Generic, DoS, Analysis, Backdoor | Shellcode, Worms |

*Rótulo de classe não incluído. Fonte: Elaborado pelo autor.

Na Tabela 1, observa-se que, no conjunto de dados GenIDS-NB15, foi necessário reduzir os fluxos para garantir a consistência, sem comprometer a robustez do conjunto. Essa medida foi adotada devido ao grande desequilíbrio entre a quantidade de fluxos normais e maliciosos extraídos. Para corrigir essa disparidade, alguns fluxos normais foram removidos, de modo a alinhar as proporções mais próximas àquelas observadas no conjunto de dados original UNSW-NB15.

Além disso, no conjunto de dados GenIDS-NB15, devido à consistência e à baixa extração de fluxos maliciosos, as instâncias das categorias de ataque *Shellcode* e *Worms* foram excluídas. De maneira similar, nos conjuntos GenIDS-CIC18 e GenIDS-CIC17, as categorias *WebAttack* e *Infiltration* foram removidas, respectivamente. Essas exclusões foram necessárias para garantir que os experimentos focassem em categorias de ataque com uma quantidade de dados suficientemente representativa, permitindo uma

análise mais confiável dos modelos de detecção de intrusão.

Em seguida, cada conjunto de dados foi dividido em cinco subconjuntos, preservando a consistência entre os fluxos normais e maliciosos. Essa divisão foi realizada com base na quantidade de instâncias de cada categoria de ataque, assegurando que a proporção de ataques fosse mantida próxima àquela observada nos dados originais. Os detalhes dessa divisão e a distribuição das instâncias são descritas na Tabela 2.

Table 2: Descrição dos Subconjuntos de Dados.

| Conjunto de dados | Número de subconjuntos | Amostras de cada subconjunto | Percentual de fluxos normais de cada subconjunto | Percentual de fluxos maliciosos de cada subconjunto | Número de atributos* | Categorias de ataques |
|---|---|---|---|---|---|---|
| GenIDS-CIC18 | 5 | 25.000 | 80,0% | 20,0% | 70 | Bruteforce, Botnet, DDoS, DoS e Infiltration. |
| GenIDS-CIC17 | 5 | 25.000 | 80,0% | 20,0% | 70 | Bruteforce, Botnet, DDoS, DoS, PortScan e WebAttack. |
| GenIDS-NB15 | 5 | 25.000 | 80,0% | 20,0% | 70 | Exploits, Fuzzers, Reconnaissance, Generic, DoS, Analysis e Backdoor. |

*Rótulo de classe não incluído. Fonte: Elaborado pelo autor.

O percentual de cada subconjunto consistiu em 80% de fluxos normais e 20% de fluxos maliciosos, garantindo a representação de todas as categorias de ataque, definidas anteriormente, nos respectivos conjuntos de dados. Esta composição permitiu uma análise mais confiável dos modelos de detecção de intrusão, abrangendo uma ampla gama de categorias de ataque.