The Best Network Traffic Dataset: A Comparative Analysis

Luca Corsetti, Federico Mancini, Samuele Mazziotti

February 2025

Indice

1 Introduzione

I dataset di riferimento sono strumenti essenziali nella ricerca sul machine learning, poiché permettono ai ricercatori di confrontare metodi in modo standardizzato e stimolano l'innovazione. In particolare, l'interesse per il machine learning e il deep learning applicato ai NIDS (Network Intrusion Detection Systems) sta crescendo rapidamente grazie alla disponibilità di strumenti come Netflow e CICFlowMeter, che trasformano i flussi di rete in dati pronti per l'addestramento, alla presenza di dataset pubblici per il rilevamento delle intrusioni e all'evoluzione di hardware e software per il machine learning.

Nonostante questi sviluppi, il panorama dei NIDS non è così avanzato come potrebbe sembrare. Per valutare realmente l'efficacia di questi modelli, è necessario che i dataset utilizzati riflettano condizioni reali, fornendo un indicatore affidabile delle loro capacità predittive. Tuttavia, l'accesso ai dati autentici del traffico di rete è limitato, poiché le aziende sono spesso riluttanti a condividere informazioni sensibili per motivi di privacy e sicurezza. Questo ha portato alla creazione di dataset sintetici, che cercano di replicare il traffico reale in ambienti controllati. Tuttavia, le scelte fatte dai fornitori, come la selezione degli attacchi da includere o la simulazione del traffico normale, devono essere attentamente progettate, poiché decisioni sbagliate possono compromettere la qualità dei dati e influenzare negativamente i risultati della ricerca.

Nel 1998, la DARPA (Defense Advanced Research Projects Agency) ha pubblicato uno dei primi dataset per simulare attacchi informatici in un ambiente controllato. Il dataset, noto come *DARPA 1998*[1], è stato fondamentale per l'addestramento e il test dei primi sistemi di rilevamento delle intrusioni. Tuttavia, è stato criticato per la scarsa realisticità del traffico benigno e per il modo in cui venivano simulati gli attacchi.
Queste limitazioni hanno portato alla creazione del dataset *KDD99*[2] nel 1999. Tale dataset, sviluppato dal MIT Lincoln Laboratory come evoluzione del DARPA 1998, ha migliorato la struttura e la rappresentazione dei flussi di rete, rendendolo più utile per l'addestramento dei NIDS. Nonostante il suo impatto significativo sulla ricerca, KDD99 è stato oggetto di critiche per la ridondanza dei dati e la scarsa varietà degli attacchi, che rendevano i modelli di machine learning ottimizzati per quel dataset specifico, ma poco adattabili a scenari reali.
Le criticità di questi dataset storici hanno messo in evidenza come i problemi nella fase di etichettatura e le limitazioni intrinseche dei dati possano compromettere la qualità delle analisi e dei modelli di machine learning sviluppati. I dataset sviluppati successivamente però (come CIC-IDS-2017 e CSE-CIC-IDS-2018), pur essendo ampiamente utilizzati, presentano problemi simili legati alla generazione delle feature e alla documentazione, sollevando dubbi sulla loro capacità di rappresentare scenari reali.

Per affrontare queste problematiche, è stata effettuata un'analisi di sette dataset popolari per i NIDS, evidenziando sei principali problemi di progettazione (definiti "*bad design smell*") che possono influire negativamente sulle performance e sulla generalizzazione dei modelli. Un'attenzione particolare è stata rivolta proprio ai dataset CIC-IDS-2017 e CSE-CIC-IDS-2018, infatti attraverso un processo di reverse-engineering, sono stati creati dataset migliorati, con una logica di etichettatura corretta e nuove feature, migliorando così la rappresentatività dei dati.

2 Dataset di Riferimento

I dataset per la rilevazione delle intrusioni di rete (NIDS) rivestono un ruolo cruciale nella ricerca accademica e industriale, poiché consentono di addestrare, testare e valutare le prestazioni di modelli di machine learning e strumenti di sicurezza informatica. Tuttavia, la difficoltà di reperire dati reali, a causa di vincoli di privacy e sicurezza, ha portato alla creazione di dataset sintetici, ottenuti attraverso la simulazione controllata del traffico di rete. In questo contesto, sono stati sviluppati diversi dataset nel corso degli anni, ognuno con specifiche peculiarità, vantaggi e limiti. Sebbene questi dataset seguano una struttura comune, ci sono importanti differenze che complicano i confronti tra di essi. Ad esempio, nonostante i tentativi di standardizzare le caratteristiche dei dati, molti ricercatori utilizzano statistiche specifiche per ogni dataset, rendendo difficile confrontare i risultati

ottenuti. In questa sezione, viene fornita una breve panoramica di ciascun dataset esaminato e delle scelte progettuali documentate dai loro autori [8]:

- **ISCX 2012**: Nel 2012, è stato sviluppato il dataset ISCX 2012 (International Security Cybersecurity eXperiment 2012), una delle prime iniziative sistematiche per la creazione di un benchmark utile allo sviluppo di sistemi di rilevazione delle intrusioni. Sviluppato dall'Università del New Brunswick, questo set di dati si distingue per l'impiego dello strumento IBM QRadar, utilizzato per estrarre caratteristiche statistiche dai flussi di rete. Le principali peculiarità del dataset risiedono nella struttura degli attacchi, organizzati in quattro scenari multistadio sovrapposti, un aspetto che cerca di simulare la progressione reale di un'intrusione. Tuttavia, la mancanza di una chiara etichettatura e la limitata variabilità del traffico generato costituiscono criticità importanti. Nonostante queste limitazioni, ISCX 2012 ha rappresentato una base per lo sviluppo di dataset successivi, come CIC-IDS-2017 e CSE-CIC-IDS-2018. Il dataset include cinque giorni di traffico di rete, sia benigno che attacchi simulati, e si compone di 25 host.
- **CTU-13**: Nel 2014 è stato pubblicato il dataset CTU-13, sviluppato dall'Università Tecnica della Repubblica Ceca, con l'obiettivo di fornire un ambiente di test per lo studio del traffico generato da reti di computer compromessi. Ciò che lo rende particolarmente prezioso è la presenza di tredici scenari distinti, ognuno caratterizzato dalla registrazione di attività reali di botnet. A differenza di altri dataset, CTU-13 si distingue per l'alta granularità delle etichette, che consente una distinzione dettagliata tra traffico benigno, malevolo e di background. Tuttavia, la presenza di traffico misto può rendere difficile l'addestramento dei modelli, specialmente se non si adottano strategie di pre-processing adeguate. Il dataset contiene oltre 20 milioni di connessioni e coinvolge numerosi host, tra cui macchine compromesse e dispositivi legittimi.
- **UNSW-NB15**: Nel 2015, il dataset UNSW-NB15 è stato sviluppato dal Cyber Range and Security Research Group dell'Università del New South Wales (UNSW) in collaborazione con l'Australian Centre for Cyber Security (ACCS) per superare le limitazioni dei dataset precedenti, combinando traffico reale e sintetico al fine di migliorare la rappresentazione degli attacchi informatici. L'elemento distintivo di questo dataset è la sua suddivisione in dieci categorie di attacco, con un set di feature articolato in quattro macro-aree: caratteristiche di base, caratteristiche di contenuto, caratteristiche temporali e flag di protocollo. Un aspetto problematico evidenziato nella letteratura è la presenza di feature altamente dipendenti, che possono semplificare eccessivamente la classificazione, riducendo la capacità dei modelli di machine learning di generalizzare su traffico reale. Il dataset include 2,5 milioni di flow records, combina traffico benigno e malevolo generato con un mix di dati reali e simulati, e comprende 45 host.
- **CIC-IDS-2017** (CIC 17) **& CSE-CIC-IDS-2018** (CIC 18): Nel 2017 e 2018, il Canadian Institute for Cybersecurity (CIC) ha sviluppato CIC-IDS-2017 e CSE-CIC-IDS-2018, due dataset creati per ampliare la scala e la varietà degli attacchi simulati, con l'obiettivo di fornire un benchmark più moderno per i sistemi di rilevamento delle intrusioni. Questi dataset rappresentano un'evoluzione rispetto a ISCX 2012, con una maggiore complessità della rete e degli scenari di attacco. CIC 17 introduce quattordici classi di attacco e si distingue per l'uso di modelli comportamentali per generare traffico benigno realistico. Tuttavia, soffre di problemi di etichettatura e limitata diversità del traffico. CIC 18, invece, migliora la scala della rete, includendo 500 host e un dataset più esteso, sebbene rimangano problemi nella separazione tra alcune classi di attacco. CIC 17 ha una dimensione di 80 GB, mentre CIC 18 comprende 16 milioni di flow records. Entrambi contengono attacchi moderni e traffico benigno, con attacchi simulati più variegati.
- **TON IoT**: Nel 2019 è stato sviluppato il dataset TON IoT (Telemetry Operational and Network data for the Internet of Things) dal Cybersecurity Research Group presso l'Università di New York (NYU), per esaminare le vulnerabilità specifiche dei dispositivi IoT e Industrial IoT (IIoT). Questo set di dati include dati di rete e informazioni provenienti da dispositivi IoT, concentrandosi sulla loro sicurezza. Questa rete simulata comprende dodici dispositivi e il dataset analizza gli attacchi simulati, in particolare DDoS, con un focus sui comportamenti di rete di IoT e IIoT. Tuttavia, presenta una limitata eterogeneità delle feature e l'uso predominante di attacchi volumetrici. Il dataset include oltre 22 milioni di record tra i vari tipi di dati.

- **Bot-IoT**: Nel 2021 è stato sviluppato il dataset Bot-IoT dal Canadian Institute for Cybersecurity (CIC) nell'ambito di un'iniziativa di ricerca in cybersecurity, con l'obiettivo di analizzare le minacce legate alle botnet nelle reti IoT. Come TON IoT, Bot-IoT include milioni di record e simula attacchi, in particolare botnet e DDoS, per lo studio della sicurezza delle reti. La rete simulata di Bot-IoT comprende dieci dispositivi e affronta le minacce specifiche delle botnet nelle reti IoT, ma anche questo dataset soffre di una limitata eterogeneità delle feature e di un uso predominante di attacchi volumetrici.

3 Analisi dei Dataset

I NIDS basati sull'apprendimento automatico si avvalgono di statistiche di flusso anziché di dati di rete grezzi, il che può oscurare informazioni sui servizi interni e sugli attacchi. Inoltre, la documentazione che accompagna questi dataset è spesso limitata e priva di una descrizione dettagliata del traffico incluso. Di conseguenza, i ricercatori meno esperti potrebbero non comprendere pienamente la natura dei dati analizzati, affidandosi a etichette generiche senza una verifica approfondita. Nella prassi comune, i dataset vengono utilizzati senza un'analisi critica della loro qualità e affidabilità, poiché si tende a presupporre che siano accurati e rappresentativi. Questo approccio può condurre a: conclusioni basate su dati distorti o non sufficientemente rappresentativi; modelli che ottengono ottime prestazioni su dataset sintetici ma falliscono in contesti reali; bias sperimentali che favoriscono tecniche complesse a discapito di soluzioni più semplici ed efficaci.

Per valutare l'uso di questi dataset, sono stati analizzati 38 articoli pubblicati tra il 2015 e il 2023 nelle principali conferenze di sicurezza informatica, escludendo quelli che si concentrano sulla sistematizzazione o su altre tematiche. Da questa analisi emergono i cosiddetti "*data design smells*" (indicatori di potenziali problematiche nel design dei dati), che evidenziano come determinate assunzioni sui dati possano condurre a conclusioni errate.

Di seguito sono riportati i sei principali *data design smells* più comunemente riscontrati nei dataset NIDS:

1. **Wrong labels**: gli errori nell'etichettatura dei dati, come il traffico benigno erroneamente etichettato come malevolo, sono frequenti e possono generare risultati fuorvianti. Inoltre, alcune classi di dati sono definite in modo vago o mancano di una documentazione adeguata.
2. **Unclear ground truth**: quando i dataset vengono corretti o aggiornati in modo non uniforme, si creano discrepanze che rendono difficile confrontare i risultati tra studi diversi.
3. **Highly dependent features**: alcuni dataset contengono caratteristiche che influenzano eccessivamente la classificazione, ma che non sono correlate ai meccanismi effettivi degli attacchi.
4. **Artificial diversity**: alcune condizioni di rete simulate, come tassi di ritrasmissione irrealistici o connessioni malformate, introducono variazioni nei dati che non rispecchiano scenari reali.
5. **Poor data diversity**: molti dataset presentano una bassa variabilità nei tipi di attacchi e nel traffico benigno.
6. **Traffic collapse**: in alcuni casi, gli attacchi non vengono completati o sono indirizzati a servizi chiusi, generando flussi di traffico con risposte limitate o assenti da parte della rete bersaglio, facilitando così la classificazione dei dati.

Paper Assumptions: *✓: assumption present, ✓*: assumption partially present, ×: assumption not present/relevant, -: unclear.* ***Datasets:*** *I: ISCX 2012, C: CIC 2017, C2: CIC 2018, U: UNSW NB15, CT: CTU-13, B: Bot IoT, T: Ton IoT.*

		Assumptions			
Paper	Dataset	FV	AV	HDF	W/U
[40]	I,U	✓	✓	✓	✓
[115]	I,CT	✓	✓	✓	✓
[79]	U	✓	✓	✓	✓
[96]	U	✓	✓	✓	✓
[6]	I,CT	✓	✓*	✘	✓
[61]	C	✓	✓	✓	✓
[112]	CT	✓	✓	✓	✓
[15]	C	✓	✓	✓*	✓
[62]	C	✓	✓	✓	✓
[100]	C	✓	✓	✓	✓
[114]	U	✓	✓*	✓	✓
[47]	C	✓	✓	✓	✓
[113]	C2	✓	✓	✓*	✓

		Assumptions			
Paper	Dataset	FV	AV	HDF	W/U
[18]	C2,U	✓	✓	✓	✓
[5]	I,C,C2	✓	✓	✘	✓
[49]	I,C	✓	✓	✓	✓
[108]	U	✓	✓*	✓	✓
[107]	C	✓	✓	✓	✓*
[52]	C	✘	✘	✘	✓
[63]	C,C2	✓	✓	✓	✓
[19]	U	✓	✓	✓	✓
[14]	B,T	✓	✓	✓	✓
[87]	I,C2,CT	✓	✓	✓	✓
[111]	U	✓	✓	✓	✓
[13]	T	✓	✓	✓	✓
[106]	C2	✓	✓	✓	✓

		Assumptions			
Paper	Dataset	FV	AV	HDF	W/U
[48]	C2	✓	✓	✓	✓
[59]	C	✘	✘	✘	✘
[55]	U	✓	✓	✓	✓*
[33]	C	✓	✓	✓	✓
[44]	U	✓*	✓*	✘	✓
[105]	CT,T	✓	✓	✓	✓
[86]	C	-	-	✘	✓*
[90]	U	✓	✓	✓	✓
[34]	B	-	-	-	-
[30]	C,C2	✓*	✓	✓	✓
[78]	U	✓	✓	✓	✓
[20]	I,C	✘	✘	✘	✓

Figura 1: Risultati analisi

I risultati dell'analisi dei 38 articoli sono riassunti nella Figura 1, dove le assunzioni sono indicate come "non chiare" quando non è stato possibile eseguire una valutazione uniforme. La maggior parte degli articoli ha implicitamente presupposto che i dataset non contenessero caratteristiche altamente dipendenti, con rare eccezioni. Alcuni hanno impiegato tecniche come l'oversampling con SMOTE[3], ma l'uso di tali metodi senza la dovuta attenzione non contribuisce ad aumentare la varietà delle caratteristiche, poiché molte di esse, a livello di pacchetto, mostrano una bassa variazione.

Molti studi hanno fatto affermazioni generali sui dati NIDS senza considerare l'influenza che le caratteristiche del dataset potessero avere sui risultati. Infatti, a eccezione di pochi, nessuno ha analizzato i dati grezzi in formato PCAP[4][5][6], portando all'uso di dati etichettati erroneamente o con verità di base poco chiare. In alcuni casi, questi errori erano irrilevanti per gli obiettivi della ricerca, mentre pochi articoli hanno utilizzato versioni corrette dei dataset (ad esempio CIC 17[7] e CIC 18[5]), rilasciati dopo la pubblicazione di molti studi. Tuttavia, un'analisi manuale superficiale avrebbe comunque permesso di individuare i problemi di etichettatura.

In molti casi, testbed mal configurati hanno causato attacchi falliti, come quelli diretti contro porte chiuse. Sebbene il rilevamento di tali connessioni possa essere utile per un sistema IDS (Intrusion Detection System), la loro identificazione in un contesto di machine learning risulta banale a causa dell'omogeneità del traffico e dei problemi di data leakage derivanti dalla divisione casuale dei dati tra training e testing. Nessun articolo ha affrontato questa problematica, rendendo il compito di classificazione più semplice di quanto apparisse inizialmente.

3.1 CIC-IDS-2017 & CSE-CIC-IDS-2018

La valutazione dei dataset CIC 17 e CIC 18 ha evidenziato diverse problematiche che compromettono la qualità e la validità dei dati, con un impatto diretto sulla capacità di addestrare modelli di machine learning e deep learning efficaci. Le principali difficoltà riscontrate riguardano la mancanza o l'errata rappresentazione di alcuni attacchi, l'assegnazione imprecisa delle etichette e la sovrapposizione tra classi, tutti fattori che ostacolano l'affidabilità e la performance dei modelli di rilevamento.

Di seguito vengono elencati i problemi principali identificati durante l'analisi:

- **Attacchi mancanti dalla versione pubblicata**: Uno dei problemi più critici riscontrati riguarda la carenza o l'assenza di alcune tipologie di attacchi nei dataset pubblicati, in particolare quelli rari ma essenziali per una valutazione completa delle capacità del modello. La Figura 2 mostra il numero di flussi dannosi che vengono persi nel dataset originale, evidenziando un marcato sbilanciamento a favore del traffico benigno. Questa mancanza è particolarmente rilevante per gli attacchi con una rappresentazione minoritaria, poiché limita la capacità del modello di apprendere correttamente le loro caratteristiche

distintive. In molti casi, questi attacchi rari risultano completamente assenti o addirittura etichettati erroneamente, aggravando ulteriormente il problema.

Un esempio concreto è il port scanning, che nei dataset del 2017 e del 2018 risulta non correttamente etichettato o addirittura del tutto assente. Questa lacuna introduce una distorsione nei risultati dei modelli di rilevamento, portando a una sovrastima delle loro capacità predittive. Infatti, senza la possibilità di analizzare questi attacchi, il modello non viene realmente testato sulla loro individuazione, compromettendo la validità delle valutazioni sulle sue prestazioni.

	2017		2018	
Attack	**Count***	**%Gain***	**Count***	**%Gain***
DDoS LOIC	[128,025] +31,339	24.48%	-	-
DDoS HOIC	-	-	[1,246,034] +918,543	73.72%
DoS GoldenEye	[10,293] +203	1.97%	-	-
DoS Hulk	[230,124] +3,680	1.60%	[923,824] +884,408	95.73%
DoS Slow HTTP	[5,499] +163	2.96%	-	-
DoS Slowloris	[5796] +10	0.17%	-	-
FTP Patator	[7,935] +19	0.24%	-	-
Heartbleed	[11] +1	9.09%	-	-
Infiltration	[36] +4	11.11%	[160,639] +63,381	39.46%
Port Scan	[158,804] +61,003	38.41%	-	-
SSH Patator	[5,897] +1	0.02%	[187,589] +389	0.21%
Web Brute Force	[1,507] +1	0.07%	-	-
Web SQL Injection	[21] +3	14.29%	[87] +1	1.15%
Web XSS	[652] +2	0.31%	[230] +2	0.87%

* Values with reference to label counts in the published dataset.

Figura 2: Etichette di attacchi mancati nella versione pubblicata

- **Riassegnazioni errate delle etichette**: Un altro problema significativo emerso dall'analisi riguarda l'errata assegnazione delle etichette a numerosi flussi di traffico. Questo errore può derivare da diversi fattori, tra cui artefatti di simulazione, traffico non direttamente correlato agli attacchi e imprecisioni nella logica di etichettatura originaria. La Figura 3 evidenzia i casi in cui la corruzione delle etichette supera il 5% del totale dei flussi con quella specifica etichetta. In entrambi i dataset analizzati, molti attacchi presentano tassi di corruzione estremamente elevati, ma la tabella non include tutte le riassegnazioni errate, poiché molte di esse hanno un tasso inferiore al 5%.

  Le riassegnazioni errate delle etichette si verificano in diversi scenari:
  - Payload vuoti o non dannosi, erroneamente classificati come attacchi.
  - Artefatti di avvio o terminazione degli attacchi, che generano traffico simile a quello benigno, ma vengono etichettati come parte dell'attacco. Un esempio sono i tentativi di brute force, in cui alcune richieste iniziali, prive di elementi dannosi, vengono comunque classificate come malevole.
  - Flussi di traffico interrotti, che frammentano connessioni TCP prolungate in più segmenti separati. Questi flussi, privi di informazioni significative, vengono spesso classificati in modo errato, compromettendo l'accuratezza dell'analisi del traffico. Questi errori nell'assegnazione delle etichette possono influire negativamente sulle prestazioni dei modelli di rilevamento, introducendo distorsioni nei risultati.

Year	Published Label	Revised Label	% Corruption*	Remarks*
2017	Bot	Botnet - Attempted	25.05%	Port/System Closed. Continued C&C connection attempts by victim after attack terminated at the published time, and C&C is no longer reachable (presumably shut down).
	DoS GoldenEye	DoS GoldenEye - Attempted	25.06%	Empty Payload
	DoS Hulk	DoS Hulk - Attempted	32.70%	No malicious payload (4.47%) Empty Payload (28.23%) Attack artifacts (0.001%)
	DoS Slow HTTP	DoS Slow HTTP - Attempted	56.34%	Empty Payload (5.13%) Target Unresponsive (51.19%) Attack startup/ tear down artifact (0.02%)
	DoS Slowloris	DoS Slowloris - Attempted	29.65%	Empty Payload (29.55%) Attack startup/ tear down artifact (0.10%)
	FTP-Patator	FTP-Patator - Attempted	49.93%	No malicious payload (49.69%) Empty Payload (0.20%) Attack startup/ tear down artifact (0.04%)
	SSH-Patator	SSH-Patator - Attempted	49.80%	No malicious payload
	Web - Brute Force	Web - Brute Force - Attempted	95.16%	Empty Payload (89.85%) Attack startup/ teardown artifact (0.60%) Attack artifact (4.71%)
	Web Attack - XSS	Web Attack - XSS - Attempted	94.48%	Attack startup/tear down artifact (0.61%) Empty Payload (93.87%)
2018	Bot	Botnet Ares - Attempted	50.06%	Empty Payload. No evidence of Botnet Zeus found during traffic analysis. Affected by TSO issue.
	Web - Brute Force	Web - Brute Force - Attempted	53.85%	Empty Payload (30.77%) Attack startup/ tear down artifact (2.46%) Attack Implemented Incorrectly (20.62%) - Flows with only single login requests do not qualify as 'brute force' in the context of a single flow
	Web - Brute Force	Benign	24.71%	Mixture of empty flows comprising TCP handshake only, and browsing traffic not specifically related to the attack.
	Web Attack - XSS	Web Attack - XSS - Attempted	50.43%	Attack startup/tear down artifact (1.30%) Empty Payload (48.26%) No malicious payload (0.87%)
	DDoS HOIC	DDoS HOIC - Attempted	86.86%	Empty Payload
	DDoS LOIC-HTTP	DDoS LOIC-UDP	0.14% (LOIC) 46.07% (UDP)	All DDoS LOIC-UDP flows on 20-02-2018 are mislabeled as DDoS LOIC-HTTP
	DDoS LOIC-HTTP	DDoS-LOIC-HTTP - Attempted	49.65%	Empty Payload
	DoS GoldenEye	DoS GoldenEye - Attempted	33.22%	Empty Payload (33.09%) Target Unresponsive (0.13%)
	DoS Hulk	DoS Hulk - Attempted	94.47%	Empty Payload
	DoS Slow HTTP	FTP-Patator - Attempted	100%	Attack mislabeled. Traffic shows TCP connection attempts to port 21 which appears to be closed
	DoS Slowloris	DoS Slowloris - Attempted	21.88%	Empty Payload
	FTP-Patator	FTP-Patator - Attempted	100%	Port/System Closed. Unsure if this was an accidental oversight. All TCP 'Syn' connection attempts to victim host/port are responded to with 'RST'.
	Infiltration	Benign	76.84%	Hosts not involved in the attack are classified as infiltration. Suspect severe contamination of background traffic due to time-based labelling
	SQL Injection	SQL Injection - Attempted	54.02%	Empty Payload (36.78%). Some flows in this category occur in the timeframe of another attack. Attack startup/tear down artifact (17.24%)
	SSH-Patator	SSH-Patator - Attempted	49.97%	Empty Payload

* Values accurate at time of publication. For the full table (including corruptions $< 5\%$), refer to `https://intrusion-detection.distrinet-research.be/CNS2022/`. There may be minor residual corrections post-publication.

Figura 3: Riassegnazioni di etichette: corruzione $> 5\%$

- **Ambiguità tra classi**: L'analisi ha anche rivelato un elevato numero di flussi che presentano caratteristiche simili a quelle di altre classi, generando una significativa sovrapposizione tra le categorie, come mostrato nella Figura 4. Tale ambiguità tra le classi è attribuibile a diversi fattori, tra cui:
  - Artefatti di simulazione, come la presenza di appendici TCP che appaiono in quasi tutte le classi, creando somiglianze tra traffico benigno e maligno.
  - Errori di etichettatura, che conducono a una classificazione errata, facendo sì che alcuni attacchi vengano identificati come traffico benigno o viceversa. Come evidenziato nella Figura 4, l'attacco DoS SlowHTTPtest è completamente etichettato in modo errato. L'analisi del tracciato dei pacchetti grezzi dai file PCAP durante questo periodo rivela tentativi di connessione FTP a un host/porta inattivo.
  - Mancanza di caratteristiche distintive, poiché alcune delle caratteristiche generate dallo strumento CICFlowMeter (es: bassa discriminazione tra classi di attacco, perdita di informazioni temporali, mancanza di feature a livello di pacchetto, similitudini tra traffico benigno e malevolo) non sono sufficientemente dettagliate o distintive da consentire una differenziazione accurata tra le classi di attacco, limitando l'efficacia dei modelli di classificazione.

2017	DDoS	Benign	6
	DoS Hulk	Benign	4,872
	DoS Slowloris	Benign	1
	Portscan	Benign	1,053
2018	Brute Force - Web	Benign	1
	Web - Brute Force	Infiltration	7
	Web Attack - XSS	Benign	1
	Web Attack - XSS	SQL Injection	1
	DoS Slow HTTP	FTP-Patator	100,760
	DoS Slow HTTP	SSH-Patator	89,438
	FTP-Patator	SSH-Patator	169,745
	Infiltration	Benign	36,889

Figura 4: Conteggio di etichette ambigue

4 Qualità ed applicabilità dei dataset

Per valutare la qualità ed applicabilità dei dataset per i NIDS, è stata utilizzata una metodologia in due fasi: una manuale, per valutare qualitativamente i problemi, e una automatizzata, per misurare la prevalenza e la gravità dei "*bad smells*" tramite misure euristiche. L'obiettivo è documentare tutti i flussi all'interno di ciascuna classe di attacco. Poiché i dataset non sono fissi, ma eterogenei e corretti in modo unico, è necessaria una revisione completa dei dati PCAP sottostanti.

L'analisi manuale ricorre al raggruppamento dei flussi CSV in cluster, dove per ogni cluster si identificano i flussi rilevanti nei PCAP originali e se ne analizzano le caratteristiche distintive, alla ricerca dei "*bad smells*":

- **Wrong labels**: L'etichetta non descrive correttamente il comportamento del flusso.
- **Unclear ground truth**: La provenienza del flusso è etichettata erroneamente o in modo ambiguo.
- **Highly dependent features**: Cluster distinti condividono proprietà simili, rendendo difficile la distinzione tra essi.
- **Artificial diversity**: Le differenze tra i cluster sono dovute ad artefatti di rete piuttosto che a variazioni naturali del traffico.
- **Poor data diversity**: Se un cluster contiene oltre il 25% dei flussi, potrebbe esserci scarsa diversità tra i flussi, indicando una rappresentazione non equilibrata dei dati.
- **Traffic collapse**: Se i cluster sono troppo simili tra loro, il modello potrebbe non riuscire a distinguere efficacemente tra traffico benigno e malevolo.

Siccome un'analisi manuale qualitativa è necessaria, ma richiede molto tempo, sarebbe utile formulare un processo automatizzato che rilevi rapidamente i problemi durante la generazione del dataset. Tuttavia, i dati PCAP sono complessi e le proprietà di ciascun dataset variano notevolmente, rendendo difficile verificare tutti gli attacchi in tutte le condizioni di rete. Perciò, sono state formulate delle euristiche sui dati CSV, con minimo riferimento ai PCAP originali. L'obiettivo è simulare l'uso dei dataset da parte di un ricercatore che non ha esaminato i dati sottostanti.

4.1 Euristiche

4.1.1 Dati Errati

Per individuare etichette errate, si usano due test: **UGT** (Unclear Ground Truth) e **WL** (Wrong Labels). Ad esempio, gli algoritmi basati sugli IP spesso classificano erroneamente il traffico di background come malevolo (es. autenticazione, discovery). Per evitarlo, si utilizza un elenco di porte note per servizi di background ($BG\ Ports$): se un flusso ha una porta di destinazione ($F_{Dst\ Port}$) in questo set, viene considerato "poco

chiaro". Poiché le porte dipendono dal dataset, possono verificarsi falsi positivi su altri set di dati. L'euristica si calcola come il rapporto tra i flussi diretti a queste porte e il totale dei flussi nel dataset per ogni classe C.

$$UGT_c = \frac{F_{Dst\ Port} \in BG\ Ports}{|C|} \quad UGT_c \in [0,1]$$

Per stimare il numero di flussi con etichette errate, si utilizza un metodo che confronta ogni campione x (singolo flusso di rete) con i suoi vicini più simili (ENN – Edited Nearest Neighbor). Se la sua etichetta è diversa dalla maggioranza ($x \in \hat{C}$ if $ENN(x) \neq C$, dove $ENN(x) \neq C$), la si considera potenzialmente errata.

$$WL_c = \frac{|\hat{C}|}{|C|} \text{ where } x \in \hat{C} \text{ if } ENN(x) \neq C, \quad WL_c \in [0,1]$$

4.1.2 Artefatti della simulazione

Per rilevare caratteristiche fortemente dipendenti (**HDF** - Highly Dependent Features), si utilizza un classificatore Random Forest. Per ogni classe di attacco C, il modello viene addestrato su una singola caratteristica F_i e il punteggio $F1$ viene misurato rispetto al background. Un valore $F1$ insolitamente alto indica una forte dipendenza della caratteristica; si riporta solo il caso più critico. Poiché questo "*bad smell*" è tipico di reti instabili con elevata perdita o ritrasmissione, è l'unico test che richiede accesso ai file PCAP originali.

$$HDF_c = \max(F1(F_i))_i \quad HDF_c \in [0,1]$$

4.1.3 Traffico altamente ripetitivo

L'obiettivo è misurare la diversità dei dati, non la complessità della classificazione. Le metriche **TCC** (Traffic Collapse Coefficient) e **PDDC** (Poor Data Diversity Coefficient) si basano su un processo in due fasi con clustering. Prima, il numero di cluster N in una classe C viene stimato con l'Elbow Method; poi, KMeans assegna i dati ai cluster C_i. Infine, la somiglianza tra coppie di punti campionati casualmente viene misurata con la Cosine Similarity (CS).

$$CS_{C_i} = \frac{A \cdot B}{||A|| \cdot ||B||} \quad where\ A \sim C_i, B \sim C_i, \quad CS_{C_i} \in [0,1]$$

Per coppie identiche (o quasi), il punteggio sarà circa 1. A questo punto l'euristica $PDDC$ può essere calcolata come misura approssimativa dell'omogeneità dei dati, dove M rappresenta il numero di campioni casuali che vengono prelevati da ogni cluster:

$$PDD_c = \sum_{i<N} \sum_{j<M} \frac{|C_i|}{|C|} \frac{CS_{C_i}}{MN}, \quad PDD_c \in [0,1]$$

4.1.4 Collasso del traffico

L'obiettivo è rilevare un'eccessiva mancanza di diversità nei dati, spesso dovuta a problemi di configurazione nella generazione del traffico. Si applica lo stesso processo di clustering usato per $PDDC$, ma considerando solo i cluster con alta somiglianza. Il valore massimo tra questi cluster viene riportato come indicatore di Traffic Collapse.

$$TC_c = \max_i \left(\sum_{j<M} \frac{[CS_{C_i} > 0.95]}{M} \right) \quad TC_c \in [0,1]$$

4.2 Valutazione dei dataset

Usando le euristiche appena definite [8] si sono analizzati i dataset introdotti nei capitoli precedenti e i risultati sono inseriti nella Figura 5.

Gli **errori di etichettatura** nei dataset NIDS derivano spesso da scelte poco chiare durante la loro progettazione. La mancanza di definizioni precise su cosa viene simulato porta a classificazioni ambigue e imprecise. Nel

dataset UNSW-NB15, l'uso di "strikes" predefiniti dello strumento IXIA PerfectStorm ha reso la classificazione poco chiara. La classe Fuzzing include attacchi con minime variazioni ai protocolli, ma senza CVE, risultando indistinguibili dal traffico benigno. La classe Generic, invece, è descritta vagamente e contiene vulnerabilità nei cifrari a blocchi che non trovano riscontro nei dati raccolti. Nel dataset TON IoT, una porzione significativa del traffico DNS è etichettata erroneamente come malevola, con il 55% dei flussi nella classe DoS e il 28% in XSS classificati in modo errato. Anche dataset con etichettatura più dettagliata, come CTU-13, presentano problemi: aggiornamenti di sistema vengono classificati come minacce hardware, e alcuni flussi vengono accidentalmente esclusi durante la cattura del traffico.

Le **etichette ambigue** e la documentazione scarsa complicano ulteriormente la ricerca. Ad esempio, nel dataset ISCX 2012, la documentazione indica che l'attacco HTTP DoS utilizza Slowloris, ma i PCAP rivelano che consiste solo in richieste GET generiche. Gli artefatti di simulazione influenzano interi dataset, introducendo caratteristiche non realistiche che distorcono le analisi. Nel dataset UNSW-NB15, i valori time-to-live sembrano derivare dal fingerprinting dei sistemi operativi, causando una distribuzione anomala tra macchine attaccate e traffico benigno, rendendo alcune caratteristiche eccessivamente discriminatorie. Errori di elaborazione sono evidenti anche nel dataset CTU-13, dove centinaia di flussi UDP inesistenti, con caratteristiche impossibili, sono stati generati per errore. Questi, insieme a traffico NetBIOS etichettato erroneamente, costituiscono il 99,9% del traffico maligno UDP registrato.

La **ripetitività dei dati** compromette la costruzione di dataset significativi, limitando la capacità di generalizzazione dei modelli di machine learning. Nel dataset UNSW-NB15, l'80,5% della classe Reconnaissance è composta da flussi quasi identici generati dal protocollo Portmap. Nel dataset CTU-13, il 97,8% dei flussi maligni proviene da una semplice scansione nmap.
Un caso estremo si osserva nel dataset Bot-IoT, dove solo lo 0,00013% del traffico è benigno e la quasi totalità di quello malevolo è volumetrico, generando milioni di flussi quasi identici. Ciò favorisce l'overfitting dei modelli, impedendo di valutare realmente la loro capacità di generalizzazione, come evidenziato nella Figura 5.

Dataset/Class	PDD_C	TC_C	WL_C	HDF_C	UGT_C
UNSW NB15					
Generic	**0.98**	0.92	0.0	**1.0**	**0.98**
Exploits	0.89	0.41	**0.10**	0.88	**0.37**
Fuzzers	0.93	0.6	**0.52**	0.66	**0.51**
DoS	0.91	0.48	**0.12**	0.90	**0.78**
Recon.	**0.95**	0.76	**0.38**	**0.95**	**0.84**
Analysis	0.93	0.54	**0.21**	0.89	**0.77**
Shellcode	**0.97**	**0.95**	**0.57**	0.69	0.0
Backdoor	0.91	0.48	0.0	0.91	**0.82**
Worms	0.94	0.46	**0.60**	0.78	0.0
ToN_IoT					
scanning	**0.97**	**0.95**	0.0	**0.99**	0.01
dos	**0.99**	**0.97**	0.0	**0.98**	0.03
ddos	**0.99**	**0.98**	0.0	**0.97**	**0.12**
mitm	0.83	0.85	**0.27**	0.73	**0.56**
xss	0.84	0.86	0.0	**0.97**	**0.27**
backdoor	**1.0**	**1.0**	**0.31**	**1.0**	0.0
injection	**0.95**	0.92	0.0	**0.98**	0.03
passwords	0.89	**1.0**	0.0	**0.99**	0.0
ransomware	0.83	0.91	0.05	0.84	0.0
Bot IoT					
DDoS	0.86	0.5	0.0	**0.98**	0.0
DoS	0.87	0.37	0.0	**0.99**	0.0
Recon.	**0.93**	0.53	0.01	**0.98**	0.0
Theft	0.89	0.67	0.06	**1.0**	0.0
CTU-13					
Neris 1	0.84	0.36	0.0	0.83	0.0
Neris 2	0.85	0.36	0.0	**0.96**	0.0
Rbot 1	**0.98**	**0.99**	**0.18**‡	**1.0**	0.0
Rbot 2	0.87	0.7	0.0	0.0	**0.98**
Virut 1	0.94	0.67	0.0	0.8	0.0
Donbot	**0.97**	**1.0**	0.0	**1.0**	0.0
Sogou	-	-	-	-	-
Murlo	**0.96**	0.84	**1.0**	0.87	0.0
Neris 3	0.9	**0.97**	0.0	0.87	0.0
Rbot 3	**0.99**	**0.99**	0.0	**1.0**	0.0
Rbot 4	**0.98**	**0.96**	0.01	**1.0**	0.0
NSIS	0.78	0.68	0.02	0.88	0.0
Virut 2	0.88	**0.95**	0.0	0.9	0.0
CIC-IDS 2017					
Portscan	**0.99**	**0.99**	0.0	**0.98**	0.0
DoS Hulk	**0.98**	**0.98**	0.0	**1.0**	0.0
FTP-Patator	**0.98**	**0.98**	0.0	**0.99**	0.0
SSH-Patator	**1.0**	**1.0**	0.0	**0.98**	0.0
DDoS	**0.98**	0.94	0.0	**0.99**	0.0
Bot	**0.98**	**1.0**	0.01	**1.0**	0.0
Slowloris	**0.97**	**1.0**	0.0	**0.98**	0.0
Slowhttptest	0.88	0.53	0.0	**0.96**	0.0
GoldenEye	**0.95**	0.68	0.0	**1.0**	0.0
Infil.	0.92	0.65	**0.81**	0.75	0.0
Brute Force	**0.99**	0.93	0.06	0.91	0.0
XSS	0.78	0.48	**0.35**	0.93	0.0
SQL	-	-	**0.63**	-	0.0
Heartbleed	-	-	**0.18**	-	0.0
CIC-IDS 2018					
Infil.	0.67	0.32	**0.65**	0.63	**0.32**
Bot	**0.99**	**0.99**	0.0	**0.99**	0.0
Hulk	**0.98**	**0.99**	0.0	**1.0**	0.0
Slowloris	0.83	0.89	0.0	**0.99**	0.0
SSH-Bruteforce	**0.99**	**0.99**	0.0	**1.0**	0.0
FTP-BruteForce	**0.99**	**1.0**	0.0	**0.99**	0.0
LOIC	**0.96**	**0.99**	0.0	**1.0**	0.0
LOIC-UDP	**0.96**	0.82	**0.16**†	**0.99**	0.0
HOIC	**0.98**	0.88	0.0	**1.0**	0.0
GoldenEye	0.93	**0.99**	0.0	**1.0**	0.0
SlowHTTPTest	**0.99**	**1.0**	**0.56***	**1.0**	0.0
XSS	0.90	**1.0**	0.05	0.83	0.0
Web	0.77	**1.0**	**0.21**	0.79	0.04
SQL	0.86	0.85	**0.21**	0.77	0.0
ICSX 2012					
BruteForce	**0.99**	**0.99**	0.02	**0.96**	0.0
SSH	**0.98**	**1.0**	0.0	0.93	0.0
nmap	**0.93**	**1.0**	0.04	0.78	0.02
IRC	**0.96**	**1.0**	0.0	0.70	0
Other	0.75	0.37	**0.35**	0.56	**0.13**

Figura 5: Esito calcolo delle euristiche sui dataset in esame

4.2.1 CIC-IDS-2017 & CSE-CIC-IDS-2018

Il risultato del lavoro sui dataset CIC 17 e CIC 18 da parte di [9] ha permesso di ottenere una versione migliorata di questi, utilizzabile dalla comunità scientifica.

Uno dei problemi più significativi dei dataset riguardava la correttezza delle etichette (*Wrong Label bad smell*) assegnate ai flussi di rete. Molti attacchi erano etichettati in modo errato, spesso confusi con traffico benigno o assegnati a categorie non corrispondenti alla realtà. Questo ha portato a distorsioni nei modelli, che imparavano a riconoscere schemi fuorvianti invece di distinguere con precisione tra traffico normale e malevolo.

Nella nuova versione del dataset, la logica di etichettatura è stata completamente rivista, con correzioni che hanno riguardato oltre il 6,67% delle etichette in CIC-IDS-2017 e 7,53% in CIC-IDS-2018. In particolare, attacchi come *DoS Hulk* e *SSH Patator* presentavano errori di classificazione superiori al 50%, con molti flussi catalogati come attacchi non riusciti o privi di payload dannoso. Un altro aspetto critico era la presenza di artefatti di simulazione (*Highly Dependent Features bad smell*), che compromettevano la qualità dei dati e la capacità dei modelli di apprendere in modo corretto. Alcune caratteristiche, come *Destination Port* e *Forward Initial Window Bytes*, risultavano essere indizi artificiali che semplificavano troppo il compito di classificazione, di conseguenza, i modelli finivano per basarsi su questi segnali piuttosto che sulle reali differenze tra traffico benigno e malevolo. La nuova versione ha eliminato questi artefatti, riducendo la dipendenza da feature superficiali e non correlate all'attacco, migliorando così la robustezza delle analisi. Un problema altrettanto rilevante era la ripetitività e lo squilibrio nei dati (*Traffic Collapse e Poor Data Diversity dad smell*). Alcune classi di attacco, come *Reconnaissance* in CIC 17, erano dominate da un unico tipo di traffico (Portmap), riducendo la variabilità nei dati e favorendo il sovradattamento dei modelli. Le correzioni introdotte hanno migliorato la distribuzione del traffico, riducendo la ripetitività e fornendo una varietà di esempi più rappresentativa della realtà.

Per verificare la versione migliorata, abbiamo calcolato le euristiche proposte da [8] sul dataset migliorato CIC 17, proposto da [9], analizzando in particolare $PDDC$ e TCC[10].

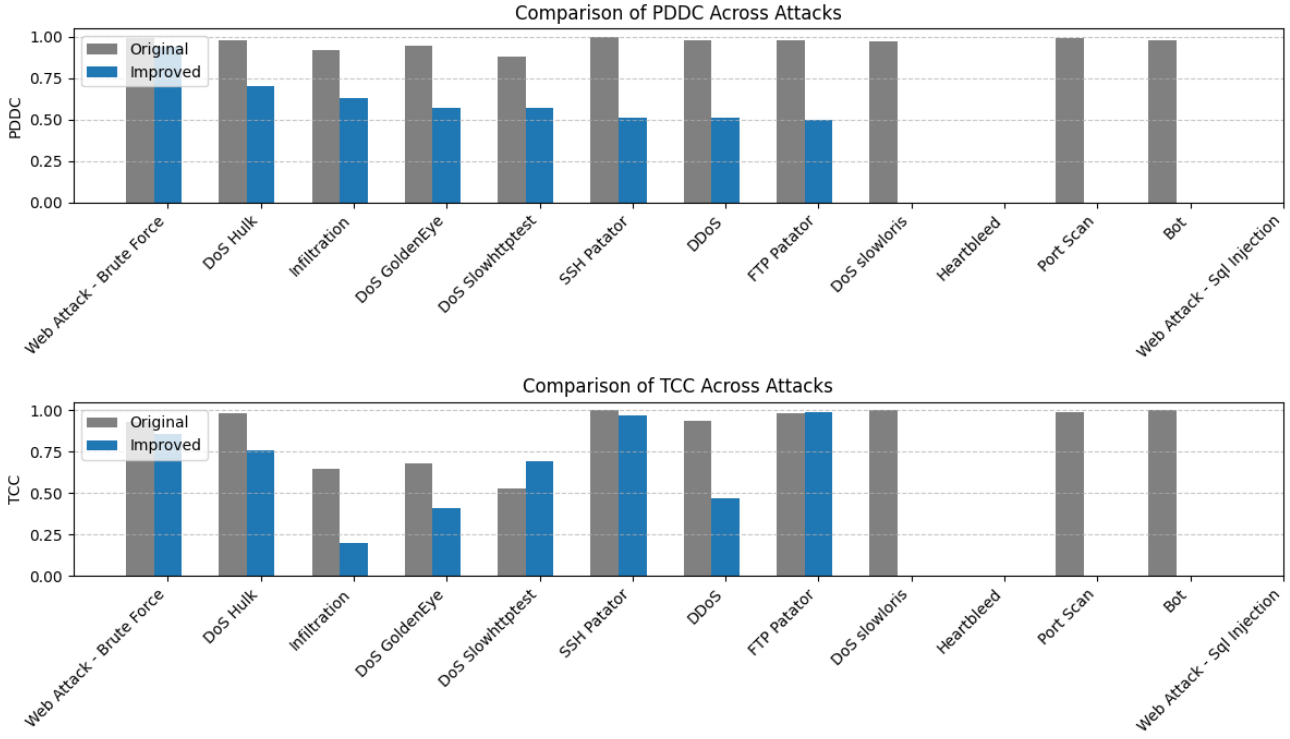

Figura 6: Esito calcolo delle euristiche

Le correzioni introdotte hanno portato a un netto miglioramento della diversità dei dati ($PDDC$) e alla riduzione del traffico ridondante (TCC) come mostrato nella figura 6. In diversi casi, le euristiche mostrano una riduzione fino al 50% rispetto al dataset originale, evidenziando una maggiore varietà nei dati e una minore presenza

di flussi duplicati. Gli attacchi come *DDoS*, *DoS GoldenEye* e *SSH Patator* hanno beneficiato maggiormente di queste correzioni, mentre alcuni attacchi, come *DoS Slowloris* e *Port Scan*, non presentano miglioramenti misurabili, probabilmente a causa di una minore attenzione nella loro rielaborazione. Un caso interessante è *FTP Patator*, che mostra un miglioramento nella diversità dei dati ($PDDC$ ridotto), ma un TCC ancora elevato, suggerendo che, nonostante la rimozione di dati ripetitivi, il traffico di attacco rimane altamente simile tra le istanze.

5 Conclusioni

L'analisi condotta in questo studio ha evidenziato diverse problematiche legate ai dataset utilizzati per il rilevamento delle intrusioni di rete (NIDS). In particolare, sono state riscontrate criticità nelle etichette assegnate ai flussi di rete, nella rappresentatività dei dati e nella presenza di artefatti di simulazione che possono distorcere i risultati dei modelli di machine learning.

L'uso di dataset sintetici è una necessità imposta dalle restrizioni legate alla privacy e alla sicurezza delle reti reali, ma le scelte di progettazione possono influenzare significativamente l'affidabilità degli studi condotti. I problemi di etichettatura errata e di scarsa diversità dei dati possono portare a una sopravvalutazione delle performance dei modelli, compromettendone l'efficacia in scenari reali. Tuttavia, non esiste un dataset intrinsecamente migliore degli altri: ogni base di dati presenta specifici punti di forza e limitazioni, che la rendono più o meno adatta a seconda del contesto applicativo. La scelta del dataset più idoneo dipende quindi dallo scenario di utilizzo e dalla tipologia di minaccia da rilevare.

La revisione critica dei dataset CIC 17 e CIC 18 ha permesso di evidenziare errori strutturali e proporre correzioni mirate, migliorando la qualità dei dati e la loro applicabilità nella ricerca. Tuttavia, l'utilizzo di dataset aggiornati è essenziale per garantire che i modelli di rilevamento siano efficaci anche contro attacchi più recenti.

Per questo motivo, è necessario che la comunità scientifica adotti pratiche più rigorose nella selezione e validazione dei dataset, evitando l'uso indiscriminato di basi di dati obsolete o inadeguate per scenari emergenti. Le euristiche proposte hanno dimostrato di essere strumenti efficaci per identificare le principali criticità dei dataset attuali, suggerendo la necessità di un approccio più metodico nella loro valutazione.

Per il futuro, è fondamentale che la ricerca si orienti verso dataset più rappresentativi e privi di artefatti che possano distorcere l'addestramento dei modelli. Inoltre, lo sviluppo di strumenti automatizzati per la verifica della qualità dei dati potrebbe contribuire a una selezione più consapevole dei dataset, migliorando la generalizzazione dei modelli e riducendo il rischio di bias sperimentali. Solo attraverso un miglioramento continuo della qualità delle basi di dati sarà possibile garantire progressi reali nel campo dei sistemi di rilevamento delle intrusioni.

Riferimenti bibliografici

[1] "Lincoln lab datasets." Accessed: 2022-2-6. https://www.ll.mit.edu/r-d/datasets.

[2] "KDD cup 99." https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[4] A. S. Jacobs, R. Beltiukov, W. Willinger, R. A. Ferreira, A. Gupta, and L. Z. Granville, "AI/ML for Network Security: The Emperor has no Clothes," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (CCS '22), Los Angeles, CA, USA, Oct. 2022, pp. 1537–1551. DOI: 10.1145/3548606.3560609.

[5] M. Lanvin, P.-F. Gimenez, Y. Han, F. Majorczyk, L. Mé, and E. Totel, "Towards Understanding Alerts raised by Unsupervised Network Intrusion Detection Systems," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses* (RAID '23), Hong Kong, China, Sep. 2023, pp. 135–150. DOI: 10.1145/3607199.3607247.

[6] A. Torralba and A. Efros, "Unbiased look at dataset bias," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR), Jul. 2011, pp. 1521–1528. DOI: 10.1109/CVPR.2011.5995347.

[7] H. Seo and M. Yoon, "Generative intrusion detection and prevention on data stream," in *Proceedings of the 32nd USENIX Conference on Security Symposium* (SEC '23), Anaheim, CA, USA, 2023, Article 242, pp. 1–17.

[8] Flood, R., Engelen, G., Aspinall, D., & Desmet, L. (2024). *Bad Design Smells in Benchmark NIDS Datasets*. In *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)* (pp. 658–675). doi: 10.1109/EuroSP60621.2024.00042.

[9] Liu, L., Engelen, G., Lynar, T., Essam, D., & Joosen, W. (2022). *Error Prevalence in NIDS Datasets: A Case Study on CIC-IDS-2017 and CSE-CIC-IDS-2018*. In *2022 IEEE Conference on Communications and Network Security (CNS)* (pp. 254–262). doi: 10.1109/CNS56114.2022.9947235.

[10] https://github.com/DataBadSmells