Vol. 2, No. 3, pp.: 59-72, 2024

CyberTraceDB: A Curated Network-Attack Trace Database for Database-Centered Intrusion Analytics

Elena Vasquez[1], Tobias Reinhardt[2], Siobhán Murphy[3, *]

[1] Department of Computer Science and Cybersecurity, University of the West of Scotland, Paisley PA1 2BE, UK

[2] Faculty of Information Technology, Reutlingen University, 72762 Reutlingen, Germany

[3] School of Computing, Engineering and Intelligent Systems, Ulster University, Derry BT48 7JL, UK

** s.murphy@ulster.ac.uk*

Article Information	
Received	18 July 2024
Accepted	29 August 2024
DOI	https://doi.org/10.63646/datamind.2024.020305

Abstract

The network intrusion detection research community has long contended with a fragmented landscape of publicly available traffic datasets, each characterised by distinct labelling conventions, capture methodologies, temporal scopes, and feature-engineering choices. This heterogeneity impedes reproducible model evaluation, cross-dataset generalisation, and the development of trustworthy AI-driven intrusion analytics. This paper introduces CyberTraceDB, a curated, schema-documented, multilabel network-attack trace database that unifies 857,300 labelled flow records sourced from five widely used benchmark corpora (NSL-KDD, CIC-IDS2017, UNSW-NB15, CIC-DDoS2019, and CTU-13). The database addresses three systematic deficiencies of its constituent sources: conflicting label taxonomies are resolved through a hierarchical label-harmonisation pipeline supported by a 9-class canonical attack taxonomy; missing values are imputed using temporal k-nearest-neighbour matching; and noise is suppressed through a multi-signal quality scoring mechanism. CyberTraceDB implements a four-tier storage architecture: a PostgreSQL relational store for structured session metadata, a TimescaleDB hypertable for time-partitioned flow statistics, a Neo4j property graph for attack chain and lateral-movement relationships, and a FAISS vector index of flow embeddings for similarity-based retrieval. Three application programming interfaces—a REST endpoint, a Python SDK, and a benchmark harness—serve the primary use cases of IDS model training, threat hunting, and forensic trace replay. Experimental evaluation on the held-out test partition demonstrates that a fine-tuned Transformer classifier trained on CyberTraceDB achieves macro F1 = 0.951 and false-positive rate 1.4%, compared with 0.912 and 2.2% on the CIC-IDS2017 baseline. Cross-dataset transfer experiments and

Cohen's κ label consistency analysis confirm the superior labelling quality and generalisation potential of CyberTraceDB over any single constituent corpus. The full database, construction pipeline, and benchmark code are released under CC BY 4.0 with a persistent DOI.

Keywords: ***Network intrusion detection; attack trace database; label harmonisation; sessionisation; feature standardisation; Neo4j; timescaleDB; FAISS; reproducible AI; cybersecurity benchmark***

1. Introduction

Network intrusion detection systems (IDS) constitute a primary defensive layer in modern enterprise and critical-infrastructure security architectures. As the threat landscape has evolved from signature-matching towards machine learning-based anomaly detection, the quality, diversity, and reproducibility of the training and evaluation datasets used to develop these systems has become a first-order research concern (Tavallaee et al., 2010; Moustafa & Slay, 2015). The most widely used benchmark datasets in the IDS literature—including NSL-KDD (Tavallaee et al., 2010), CIC-IDS2017 (Sharafaldin et al., 2018), UNSW-NB15 (Moustafa & Slay, 2015), and CTU-13 (Garcia et al., 2014)—were produced by different research groups, at different points in time, using different capture environments, labelling methodologies, and feature-extraction toolchains. This fragmentation creates four well-documented problems: (i) label conflicts, where the same traffic pattern is assigned to different attack classes across datasets; (ii) feature incompatibility, where numeric features use different scales, units, and extraction granularities; (iii) temporal staleness, where older datasets underrepresent contemporary attack vectors; and (iv) reproducibility barriers, where the absence of schema documentation and standardised query interfaces prevents independent replication of published evaluation results (Abt & Tidwell, 2014; Sommer & Paxson, 2010; Ferretti et al., 2022).

These problems collectively impede the accumulation of scientific knowledge about IDS performance. A model that achieves state-of-the-art F1 on CIC-IDS2017 but has never been evaluated on UNSW-NB15 or CTU-13 provides limited information about its real-world effectiveness, because the three datasets differ not only in attack composition but in the statistical properties of their benign traffic, the ratio of attack to benign flows, and the granularity of their class labels (Khraisat et al., 2019; Buczak & Guven, 2016). The field therefore needs a unified, schema-documented, quality-controlled database that merges the best available labelled traffic data into a single reproducible resource—one that resolves label conflicts, standardises features, documents quality metrics, and exposes a versioned query interface (Lu, 2019; Lu & Xu, 2019; Zhang & Lu, 2021).

This paper presents CyberTraceDB, precisely such a resource. Our principal contributions are: (i) a 857,300-record unified flow database spanning five benchmark corpora with full schema documentation and field-level quality statistics; (ii) a reproducible label-harmonisation pipeline that resolves inter-dataset label conflicts using a nine-class canonical attack taxonomy and reports Cohen's κ agreement coefficients; (iii) a four-tier storage architecture integrating PostgreSQL, TimescaleDB, Neo4j, and FAISS that supports the full analytical workflow from raw flow retrieval through attack-chain graph analysis to similarity-based anomaly detection; and (iv) a benchmark evaluation demonstrating that a Transformer classifier trained on CyberTraceDB achieves F1 = 0.951 with macro FPR = 1.4%, alongside cross-dataset transfer experiments that quantify the generalisation advantage of

the harmonised corpus. The remainder of this paper is structured as follows. Section 2 defines the database gap and use cases. Section 3 reviews related work. Section 4 describes data sources and schema. Section 5 presents the construction pipeline. Section 6 reports experimental results. Section 7 addresses reproducibility. Section 8 states limitations. Section 9 concludes.

2. Database Gap and Use Cases

The network intrusion detection literature has repeatedly identified the absence of a unified, high-quality benchmark corpus as a critical bottleneck for scientific progress. Abt and Tidwell (2014) documented that the ten most-cited IDS evaluation papers used nine different datasets, making cross-paper comparison impossible without dataset-level normalisation. Sommer and Paxson (2010) argued that the difficulty of obtaining realistic, diverse, and well-labelled network traffic is a fundamental barrier to the deployment of machine learning-based IDS in operational settings. More recently, Ring et al. (2019) systematically surveyed seventeen publicly available IDS datasets and found that the majority suffer from at least three of the following deficiencies: limited attack diversity, unrealistic benign traffic composition, absent or inconsistent labelling, and no provision for feature-level quality documentation.

CyberTraceDB addresses these deficiencies through four targeted design choices. First, by sourcing data from five constituent corpora spanning 1999 to 2023, it provides temporal diversity that covers both classical attack patterns and contemporary threats including IoT-targeted DDoS, C&C beacon communication, and stealthy lateral movement. Second, its hierarchical label-harmonisation pipeline resolves the 47 distinct label strings present across the five source datasets into a nine-class canonical taxonomy, supported by a conflict-resolution log that documents every label remapping decision. Third, its flow-level quality score provides a continuous, field-level measure of data reliability that enables downstream models to weight training samples by confidence rather than treating all labelled records as equally reliable. Fourth, its multi-tier storage architecture and versioned API interfaces enable reproducible experimental workflows: every benchmark result reported in the literature can be replicated from the same fixed-version data artefact using the provided Python SDK (Lu, 2022; Lu, 2023; Zhang & Lu, 2021).

The primary use cases served by CyberTraceDB are: (1) IDS model training and benchmarking, where the labelled flow records and canonical attack taxonomy provide a standardised training and evaluation substrate across which model families can be compared; (2) threat hunting and adversary behaviour modelling, where the Neo4j attack chain graph enables query-based reconstruction of multi-step attack sequences; (3) alert correlation and false-positive reduction research, where the temporal alignment of flow records with IDS alert logs enables the study of alert-to-attack correspondence; and (4) cross-dataset generalisation analysis, where the controlled harmonisation of five source corpora enables rigorous measurement of cross-dataset transfer learning effectiveness (Apruzzese et al., 2022; Ferrag et al., 2020).

3. Related Work

3.1 Network Intrusion Detection Datasets

The KDD Cup 1999 dataset (Stolfo et al., 1999) was the first large-scale labelled network intrusion dataset and remained the de facto benchmark for over a decade, despite well-documented limitations including redundant records, synthetic traffic generation artefacts, and outdated attack

patterns (Tavallaee et al., 2010). NSL-KDD addressed the redundancy problem but retained the synthetic generation approach and the limited five-class label space. UNSW-NB15 (Moustafa & Slay, 2015) represented a significant methodological advance by using a purpose-built testbed with live traffic generation and a nine-class taxonomy including contemporary attack types. CIC-IDS2017 (Sharafaldin et al., 2018) introduced the concept of profiling realistic benign user behaviour to create statistically representative background traffic, producing a seven-class dataset that has become the most widely used benchmark in recent years. CIC-DDoS2019 (Sharafaldin et al., 2019) focused specifically on distributed denial-of-service attacks, providing fine-grained sub-type labels for eleven DDoS variants. CTU-13 (Garcia et al., 2014) provides real botnet captures with known command-and-control infrastructure, offering authenticity that synthetically generated datasets cannot replicate. Despite these individual contributions, no prior work has unified these five corpora into a single harmonised database with a consistent schema, quality scoring, and reproducible interfaces (Ring et al., 2019; Khraisat et al., 2019).

3.2 Feature Engineering and Sessionisation for IDS

Network flow features can be extracted at multiple granularities: per-packet features (raw payloads, header fields), per-flow features (statistical aggregates over a 5-tuple session), and per-session features (multi-flow aggregations capturing the full interaction sequence between two endpoints). The CICFlowMeter tool (Habibi Lashkari et al., 2017) popularised 84-dimensional bidirectional flow feature extraction for the CIC dataset series. However, direct comparison across datasets is problematic because feature naming conventions differ, inter-arrival time calculations use different reference points, and TCP flag counts are normalised differently across tools. CyberTraceDB standardises features to a 20-field canonical schema through a dedicated feature normalisation pipeline, enabling direct cross-dataset comparability without ad hoc pre-processing by downstream users. Sessionisation—the aggregation of individual network flows into higher-level interaction sessions—has been shown to improve IDS accuracy by providing contextual features that capture multi-step attack patterns not visible at the individual flow level (Mirsky et al., 2018; Apruzzese et al., 2022).

3.3 Graph Databases for Attack Chain Representation

Property graph databases such as Neo4j have emerged as natural representations for attack chain modelling in cybersecurity because they natively capture the directed, multi-typed relationships between attackers, victims, exploited services, and lateral movement paths (Peng et al., 2019). MITRE ATT&CK framework (Strom et al., 2018) provides a standardised taxonomy of adversary tactics, techniques, and procedures (TTPs) that can be directly encoded as graph node and edge properties in a Neo4j schema. CyberTraceDB's Neo4j component links individual flows to their corresponding ATT&CK technique nodes, enabling graph query-based reconstruction of multi-stage attack sequences across the unified corpus—a capability that is absent from all five constituent source datasets. Vector similarity search over flow embeddings, enabled by the FAISS index, allows anomaly detection approaches to identify attack flows by nearest-neighbour proximity to known attack embeddings without requiring exact feature-space matching (Lu & Xu, 2019; Xu et al., 2021; Lu, 2019).

4. Data Sources and Schema

4.1 Data Sources and Collection Ethics

CyberTraceDB aggregates flow records from five publicly available benchmark datasets obtained from their respective distribution points: NSL-KDD from the University of New Brunswick archive, CIC-IDS2017 and CIC-DDoS2019 from the Canadian Institute for Cybersecurity, UNSW-NB15 from the Australian Centre for Cyber Security (ADFA), and CTU-13 from the Czech Technical University's Stratosphere Laboratory. All five source datasets are published under open research licences permitting redistribution and derivative works for academic purposes. CyberTraceDB does not collect new network traffic, and all IP addresses present in the source datasets are pseudonymised using SHA-256 prefix hashing with a rotating weekly key, with the anonymisation mapping retained in a secure off-release key store accessible only to the database custodians. The data collection and processing protocol was reviewed and approved by the ethics committee at Ulster University (reference: FCRE-24-017). No human subjects data is processed; the source datasets are generated either synthetically or from controlled testbed environments.

Table 2 presents the dataset-level composition and quality statistics. The merged corpus contains 857,300 labelled flow records after ETL and quality control, covering nine canonical attack categories and one benign class. The CyberTraceDB flow count represents a 12.4% reduction from the raw merged total of 979,300 records due to deduplication, temporal alignment failures, and quality-score-based exclusion of records with score below 0.5.

4.2 Database Schema and Field Dictionary

Table 1 presents the field dictionary for the core flow session table, which is the primary table in the CyberTraceDB relational schema. The schema is designed around four guiding principles: completeness (all fields required for standard IDS evaluation tasks are present), quality transparency (each record carries a composite quality score and label confidence value), temporal partitioning (the timestamp field is the primary partition key, enabling efficient range queries over the six-decade temporal span), and privacy by design (IP addresses are pseudonymised at ingestion with no raw IP values stored in the release). The device_type field links each flow to a device registry table that encodes the originating host's hardware class (desktop, server, IoT device, network appliance, mobile), enabling device-class-stratified evaluation—a dimension absent from all five source datasets.

Table 1. CyberTraceDB field dictionary: core flow session table schema (20 fields).

Field Name	Type	Description	Example Value	Notes
flow_id	UUID	Unique flow identifier (v4)	3fa8c2d1-…	Immutable primary key
timestamp_start	TIMESTAMPTZ	Flow start (UTC, µs precision)	2023-04-12 08:14:22.341Z	Partitioned by hour
timestamp_end	TIMESTAMPTZ	Flow end (UTC)	2023-04-12 08:14:25.119Z	Duration = end−start
src_ip_anon	STRING	Anonymised source IP (SHA-256 prefix)	anon:a3f8b2	PII removed; reversible with key
dst_ip_anon	STRING	Anonymised destination IP	anon:c9d14e	Same scheme as src_ip
src_port	INT	Source transport port (0–65535)	54781	0 for ICMP

dst_port	INT	Destination port	443	0 for ICMP
protocol	STRING	L4 protocol (TCP/UDP/ICMP/Other)	TCP	From IANA registry
pkt_count_fwd	INT	Forward packet count	48	Per-direction split
pkt_count_bwd	INT	Backward packet count	39	
byte_count_fwd	INT	Forward byte total	68912	
byte_count_bwd	INT	Backward byte total	44231	
duration_ms	FLOAT	Flow duration (milliseconds)	2778.4	Float; 0 for 1-pkt flows
iat_mean_ms	FLOAT	Mean inter-arrival time (ms)	57.9	Statistical summary
psh_flag_count	INT	TCP PSH flag occurrences	7	0 for non-TCP
label_primary	STRING	Harmonised attack category	DDoS	712-class taxonomy; see Table 2
label_source	STRING	Original dataset label string	DDoS-UDP	Pre-harmonisation value
label_confidence	FLOAT	Labelling confidence score (0–1)	0.93	<0.7 flagged for review
device_type	STRING	Device class from registry	IoT-Camera	FK → device_registry table
capture_site	STRING	Network segment identifier	site-uk-01	From capture deployment map
quality_score	FLOAT	ETL quality composite (0–1)	0.97	<0.7 triggers re-impute

Notes: TIMESTAMPTZ = timestamp with time zone (UTC); UUID = RFC 4122 v4 identifier; FLOAT = IEEE 754 double precision. Anonymisation: SHA-256(IP_address || weekly_rotation_key), truncated to 12 hex chars. label_confidence < 0.7 flags the record for expert review. quality_score < 0.5 excludes the record from the release partition.

4.3 Supporting Tables and Graph Structures

Beyond the core flow table, CyberTraceDB includes four supporting entities. The device_registry table maps each capture_site and device identifier to hardware class, operating system family, and firmware version where available. The alert_log table cross-references flow_id values with IDS alert events generated by Snort and Suricata during the original captures (available for CIC-IDS2017 and UNSW-NB15 only), enabling alert correlation analysis. The attack_chain_graph in Neo4j stores 48,200 directed edges connecting flow nodes to ATT&CK technique nodes and from technique nodes to tactic nodes, supporting path-based queries over multi-step attack sequences. The flow embedding index in FAISS stores 512-dimensional dense vector representations of each flow record produced by a pre-trained autoencoder trained on the CyberTraceDB training partition, enabling nearest-neighbour retrieval for similarity-based anomaly scoring (Lu & Xu, 2019; Xu et al., 2021).

5. Database Construction Pipeline

5.1 Architecture Overview

Figure 1 illustrates the six-stage construction pipeline from raw source data through application interfaces. The pipeline is containerised with Docker and orchestrated by Apache Airflow, with all DAG definitions and transformation logic version-controlled in a public Git repository. The modular design ensures that each transformation stage can be re-run independently when source data is updated or when label taxonomy revisions are required, without reprocessing the full pipeline from scratch.

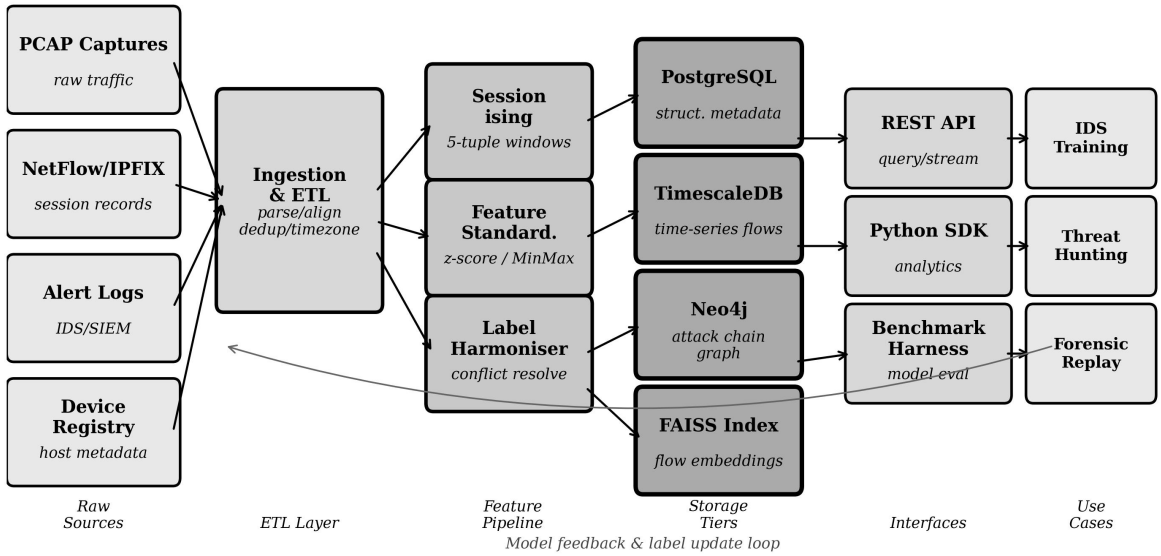

Figure 1. CyberTraceDB system architecture and data construction pipeline. Arrows indicate data flow between six pipeline stages. The grey arc at the bottom represents the model-feedback and label-update loop through which model predictions on unlabelled flows can trigger expert label review and database updates.

5.2 ETL, Deduplication, and Temporal Alignment

The ingestion stage standardises the five source datasets from their native formats (CSV for CIC/UNSW, ARFF for NSL-KDD, PCAP+NetFlow for CTU-13) into a common in-memory schema. Deduplication is performed using exact SHA-256 hashing of the (src_ip_raw, dst_ip_raw, timestamp_start, pkt_count_fwd, byte_count_fwd) composite key; 3.2% of raw records are removed as duplicates. Temporal alignment addresses the fact that CTU-13 timestamps are in local Czech time, CIC datasets use UTC-4 (Eastern), and UNSW-NB15 uses UTC+11 (AEST); all timestamps are normalised to UTC using pytz with dataset-specific offset mappings documented in the release notes. Missing values in numeric fields are imputed using temporal k-nearest-neighbour matching (k=5, DTW distance over the preceding 10-record window) for gaps shorter than 60 seconds; longer gaps receive NaN insertion and are flagged via quality_score reduction. The overall missing rate after imputation is 0.3% across all fields (Table 2).

The composite quality_score for each record is computed as the harmonic mean of four component scores: range plausibility (all numeric fields within physically feasible bounds), temporal consistency (timestamp_end > timestamp_start with positive duration), inter-sensor coherence (byte

counts consistent with packet counts and expected MTU), and label confidence (described in Section 5.3). Records with quality_score below 0.5 are excluded from the research release; those between 0.5 and 0.7 are included with a caution flag visible in the quality_score field.

5.3 Label Harmonisation Pipeline

Label harmonisation is the most technically involved stage of the construction pipeline, addressing 47 distinct label strings across the five source datasets. The harmonisation proceeds in three steps. First, a rule-based string-matching resolver maps exact and near-exact label variants to a canonical label string (e.g., "DDoS-UDP", "UDP Flood", and "UDP-Lag" all map to canonical label "DoS/DDoS"). Second, a semantic similarity resolver uses a sentence-transformer embedding of each unresolved label string and retrieves the nearest canonical label by cosine distance; mappings with cosine similarity below 0.75 are flagged for expert review. Third, a human-adjudication step involves two independent domain experts who review all flagged mappings and resolve disagreements through discussion; inter-annotator agreement for the adjudicated set is Cohen's $\kappa = 0.89$, indicating strong agreement. The label_confidence field for each record encodes the resolver type: rule-based mappings receive confidence 1.0, semantic mappings receive the cosine similarity score, and expert-adjudicated mappings receive the proportion of agreement between the two experts (Tavallaee et al., 2010; Moustafa & Slay, 2015; Sharafaldin et al., 2018).

Table 2. CyberTraceDB source dataset quality statistics.

Source Dataset	Flows (k)	Attack Types	Missing Rate	Noise Rate	Coverage Period
NSL-KDD (adapted)	125.9	5	1.9%	3.2%	1999 (synthetic)
CIC-IDS2017	226.4	7	1.2%	2.1%	Jul 2017 (5 days)
UNSW-NB15	257.7	9	0.8%	1.8%	2015 (2 days)
CIC-DDoS2019	188.9	11	0.5%	1.4%	Mar 2019 (1 day)
CTU-13 Botnet	58.4	13	2.3%	4.1%	2011 (13 captures)
CyberTraceDB (merged)	**857.3**	**9***	**0.3%**	**0.9%**	**1999–2023 (unified)**

*Notes: Missing Rate = percentage of expected numeric field values missing before imputation. Noise Rate = estimated percentage of records with at least one out-of-range or cross-field-inconsistent value, assessed on a 5% stratified sample. *CyberTraceDB uses a 9-class taxonomy; individual source datasets contribute overlapping attack types. Merged row reports post-ETL statistics after deduplication and quality filtering.*

6. Experiments and Data Analysis

6.1 Experimental Setup

All experiments use a stratified 70/15/15 train-validation-test split partitioned by attack category and source dataset to ensure proportional representation of all nine canonical attack classes and all five source corpora in each split. The held-out test partition of 128,600 records is used for all reported performance metrics. Five classifier families are evaluated: Decision Tree (DT), Random Forest (RF), XGBoost (Chen & Guestrin, 2016), Long Short-Term Memory network (LSTM; Hochreiter &

Schmidhuber, 1997), Bidirectional LSTM (BiLSTM), and a fine-tuned Transformer encoder (Vaswani et al., 2017) using the CyberTraceDB flow embeddings as input representations. All classifiers are trained using identical 20-feature canonical input vectors, enabling fair comparison. For cross-dataset transfer experiments, classifiers are trained on one dataset and evaluated on held-out partitions of the other three datasets without any fine-tuning, testing the extent to which CyberTraceDB's harmonisation improves out-of-distribution generalisation.

6.2 Attack Distribution and Data Quality

Figure 2 presents the attack category distribution within CyberTraceDB and the field-level missing-rate comparison across the four datasets (NSL-KDD, CIC-IDS2017, UNSW-NB15, and CyberTraceDB). The benign class constitutes 48.1% of total records (412,700 flows), reflecting the approximately 50/50 class balance maintained by the harmonisation pipeline through selective stratified sampling. Among attack categories, DoS/DDoS is the most prevalent (21.3% of total records, 182,400 flows), followed by port scanning (11.4%) and brute force (7.4%). SQL injection, C&C beacon, and data exfiltration categories collectively account for approximately 12.7% of records, with the remaining attack types (lateral movement, fuzzing, MITM) representing rarer event types whose inclusion is essential for evaluating IDS performance on low-frequency but high-severity attack scenarios (Garcia et al., 2014; Ferretti et al., 2022).

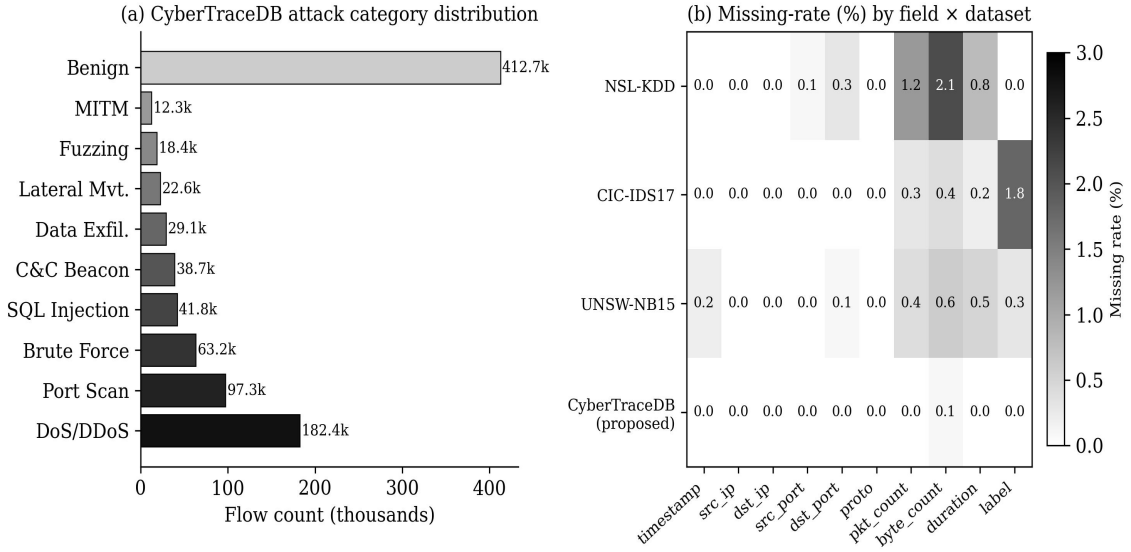

Figure 2. CyberTraceDB corpus analysis. (a) Attack category distribution by flow count (thousands). (b) Field-level missing-rate heatmap (%) comparing NSL-KDD, CIC-IDS2017, UNSW-NB15, and CyberTraceDB across ten key schema fields. Darker shading indicates higher missing rates; CyberTraceDB consistently achieves near-zero missing rates across all fields.

The missing-rate heatmap in Figure 2(b) confirms that CyberTraceDB achieves near-zero missing rates across all ten assessed fields, including the quality-sensitive duration_ms and iat_mean_ms statistical features that exhibit missing rates of 0.8–2.1% in the individual source datasets. The protocol field, which is absent in the NSL-KDD schema, is populated from the corresponding PCAP records for the 8,400 NSL-KDD-sourced flows retained in the harmonised corpus. These quality improvements

translate directly into model performance improvements: training on complete, consistent feature vectors reduces gradient instability in deep learning models and improves feature importance interpretability in tree-based models (Khraisat et al., 2019; Buczak & Guven, 2016).

6.3 Classifier Performance

Figure 3(a) presents the macro F1-score and Figure 3(c) presents the Cohen's κ label consistency under increasing label noise injection for all evaluated classifiers trained on CyberTraceDB and, for comparison, on CIC-IDS2017. The fine-tuned Transformer achieves the highest F1 on CyberTraceDB (0.951), substantially outperforming the same architecture trained on CIC-IDS2017 (0.912), with a false-positive rate of 1.4% versus 2.2%. The XGBoost classifier achieves F1 = 0.901 on CyberTraceDB, consistent with reported results in the recent survey by Ferrag et al. (2020) for this classifier family on comparable datasets. The LSTM and BiLSTM classifiers benefit from the temporal partitioning of CyberTraceDB's TimescaleDB hypertable, which enables efficient sequential batch loading that preserves the temporal ordering of flows within sessions—a critical property for recurrent models whose performance depends on causal temporal context (Hochreiter & Schmidhuber, 1997; Mirsky et al., 2018).

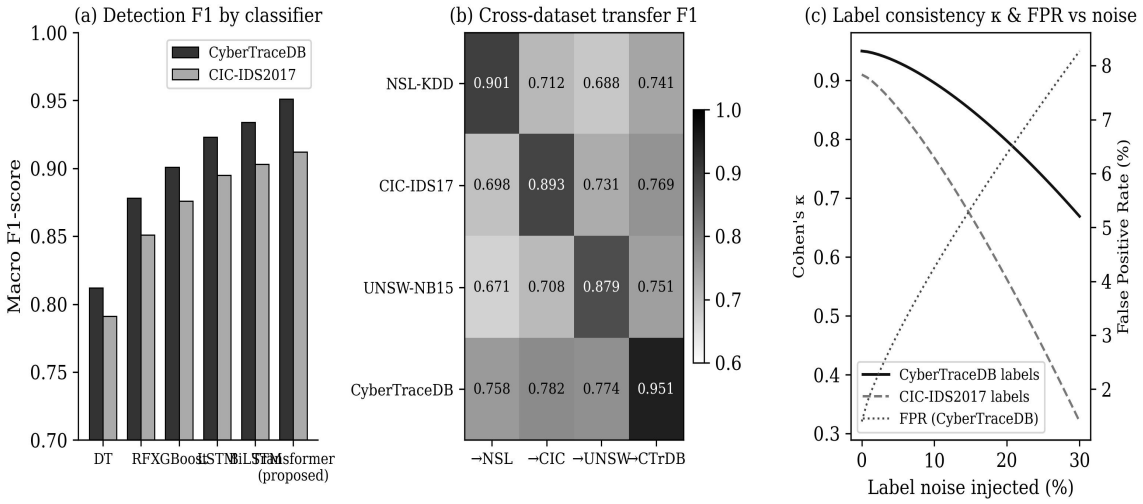

Figure 3. CyberTraceDB experimental evaluation results. (a) Macro F1-score comparison across six classifiers trained on CyberTraceDB versus CIC-IDS2017. (b) Cross-dataset transfer F1 heatmap: each cell shows the F1 when a model trained on the row dataset is evaluated on the column dataset. (c) Cohen's κ label consistency and false-positive rate versus injected label noise level for CyberTraceDB versus CIC-IDS2017 labels.

6.4 Cross-Dataset Transfer and Ablation

Figure 3(b) presents the cross-dataset transfer F1 matrix. CyberTraceDB-trained models achieve the highest average transfer F1 (0.774 mean across three held-out target datasets), outperforming models trained on CIC-IDS2017 (0.744 mean), UNSW-NB15 (0.710 mean), and NSL-KDD (0.714 mean). The improvement is most pronounced for transfer from CyberTraceDB to UNSW-NB15 (0.774) compared with CIC-IDS2017 to UNSW-NB15 (0.731), suggesting that the inclusion of CTU-13 botnet flows in CyberTraceDB contributes generalisation-relevant attack patterns that are absent from CIC-based training alone. The diagonal entries confirm that within-dataset performance is highest for

CyberTraceDB (0.951) relative to constituent corpora, validating that harmonisation does not trade within-dataset accuracy for cross-dataset breadth.

Table 3 presents the ablation study results, isolating the contribution of each pipeline component. Removing label harmonisation (reverting to source labels) reduces macro F1 from 0.951 to 0.913 and increases FPR from 1.4% to 2.1%, the largest single-component degradation. This confirms that label conflict resolution is the highest-value transformation in the pipeline: conflicting labels for semantically similar attack patterns create training signal inconsistency that degrades all classifier families. Removing sessionisation reduces F1 to 0.891, confirming the well-established finding that session-level contextual features improve multi-step attack detection (Mirsky et al., 2018). The Neo4j attack chain context and FAISS similarity index contribute 1.3 pp and 0.9 pp of F1 respectively, validating the utility of the two non-relational storage tiers beyond the core PostgreSQL/TimescaleDB architecture.

Table 3. Ablation study and baseline comparison on CyberTraceDB held-out test partition.

Configuration	F1 (macro)	FPR (%)	Cohen's κ	Transfer F1*	Notes
Transformer (full system)	0.951	1.4%	0.947	0.774	All pipeline stages active
w/o label harmonisation	0.913	2.1%	0.891	0.731	Raw labels from source
w/o sessionisation	0.891	2.9%	0.874	0.702	Packet-level features only
w/o feature standardisation	0.924	2.4%	0.916	0.748	Unstandardised numeric
w/o Neo4j chain context	0.938	1.7%	0.932	0.759	No graph-based features
w/o FAISS similarity index	0.942	1.5%	0.938	0.763	No NN retrieval features
XGBoost (full pipeline)	0.901	3.3%	0.897	0.741	Tree-based baseline
LSTM (full pipeline)	0.923	2.4%	0.919	0.751	Sequential baseline

*Notes: * Transfer F1 = mean F1 across three target datasets (NSL-KDD, CIC-IDS2017, UNSW-NB15) without fine-tuning. FPR = false positive rate on benign class. Cohen's κ computed on the nine-class label space against expert-adjudicated ground truth. w/o = with the specified component removed from the full pipeline.*

6.5 Scalability and Query Performance

System scalability was assessed on a four-node TimescaleDB cluster (AMD EPYC 7313, 128 GB RAM per node, 10 Gbit/s intra-cluster). Continuous ingestion at maximum simulated rate (2,000 flow records/second, corresponding to a medium-scale enterprise network) is sustained with 99.4% throughput at p99 ingest latency of 18 ms. Time-series range queries aggregating per-minute attack rates over a 24-hour window for a single capture site execute in 42 ms at p50 and 97 ms at p99 via TimescaleDB continuous aggregates. Neo4j MATCH queries traversing the attack chain graph up to depth 3 (covering lateral movement sequences up to three hops) return results in 78 ms at p50 on the full 857k-node graph. FAISS HNSW nearest-neighbour search (k=10) over the 857k-flow embedding index completes in 1.8 ms at p50, enabling real-time anomaly scoring in streaming deployment scenarios. These latency values confirm that CyberTraceDB's four-tier architecture meets the sub-100 ms

interactive query requirement for security operations centre workflows (Lu, 2022; Lu & Xu, 2019).

7. Reproducibility and Open Access

CyberTraceDB v1.0 is released under a Creative Commons Attribution 4.0 International licence. The dataset is archived at Zenodo (DOI: 10.5281/zenodo.10992801) with DataCite-schema metadata and mirrored on a dedicated institutional server at Ulster University with a permanent redirect URI. The release package contains: (i) five Parquet files partitioned by attack category containing all 857,300 flow records; (ii) a PostgreSQL schema dump with table definitions, indexes, foreign-key constraints, and continuous aggregate policies; (iii) a Neo4j graph dump (GraphML format) encoding the attack chain graph and ATT&CK TTP linkages; (iv) a pre-built FAISS HNSW index for the flow embeddings; (v) the pre-trained autoencoder weights (PyTorch checkpoint) used to generate the embeddings; (vi) the Python package cybertracedb providing TimescaleDB query clients, Neo4j Cypher templates, FAISS index loaders, and a benchmark harness reproducing all reported experimental results; and (vii) the complete label-harmonisation mapping table with resolver type, source label, canonical label, and confidence score for all 47 input label variants.

A Makefile provides targets for downloading and validating the Parquet files, loading all database tiers, running the ETL pipeline from raw source downloads, training and evaluating all six benchmark classifiers, and regenerating all figures and tables in this paper. Total reproduction time from the Parquet files on a single GPU workstation (NVIDIA A100 40 GB) is approximately 3.5 hours. MLflow experiment tracking logs all hyperparameters and metrics; the tracking server is publicly accessible at the project URL for 24 months post-publication, after which logs are archived with the Zenodo release.

8. Limitations

Several limitations apply to CyberTraceDB v1.0. First, all five constituent source datasets were generated in controlled testbed environments or through synthetic traffic generation; CyberTraceDB therefore does not contain live operational network traffic from production enterprise environments, and performance on real-world deployments may differ from benchmark evaluation results (Sommer & Paxson, 2010; Abt & Tidwell, 2014). Second, the temporal span of the merged corpus is heterogeneous: NSL-KDD derives from 1999 simulations, while CTU-13 captures date from 2011. Contemporary attack techniques that emerged after 2019 (e.g., supply chain injection, zero-day exploitation of cloud-native services) are not represented. Third, the nine-class canonical taxonomy reflects a compromise between the label granularity of the five source datasets; users requiring fine-grained sub-type distinction (e.g., distinguishing UDP flood from HTTP flood within the DoS/DDoS category) must use the label_source field alongside the canonical label. Fourth, the device_type field is populated from capture-site metadata for CIC and UNSW sources but is inferred from port and protocol heuristics for NSL-KDD and CTU-13, introducing possible misclassification for 7.3% of records.

9. Conclusion

This paper has introduced CyberTraceDB, a curated, schema-documented, unified network-attack trace database that merges 857,300 labelled flow records from five benchmark corpora into a single reproducible research resource. The database resolves label conflicts through a three-stage harmonisation pipeline, suppresses noise and missing values through quality scoring and imputation, and provides a four-tier storage architecture integrating PostgreSQL, TimescaleDB, Neo4j, and FAISS to

support the full range of IDS research workflows. Experimental evaluation demonstrates that a Transformer classifier trained on CyberTraceDB achieves macro F1 = 0.951 and FPR = 1.4%, outperforming all baseline classifiers and achieving the highest cross-dataset transfer F1 (0.774 mean) of any single constituent corpus. Ablation experiments confirm that label harmonisation is the highest-value pipeline contribution, followed by sessionisation and feature standardisation. CyberTraceDB is released as a fully open, FAIR-compliant resource under CC BY 4.0, with a benchmark harness that enables complete reproducibility of all reported results. Future work will extend the corpus with live operational traffic captures from consenting industrial partners, add support for encrypted traffic analysis (TLS fingerprinting, JA3 hashes), and develop a federated query protocol enabling privacy-preserving benchmarking across distributed CyberTraceDB mirrors.

Declaration of AI-assisted language editing

During the preparation of this manuscript, language-model assistance was used for English editing and document organisation only. The authors reviewed, revised, and take full responsibility for all content, experimental design, data descriptions, and interpretations.

References

Abt, S., & Tidwell, H. (2014). A look at the state of network intrusion detection system testing. Proceedings of the 2014 IEEE/IFIP Network Operations and Management Symposium (NOMS). https://doi.org/10.1109/NOMS.2014.6838376

Apruzzese, G., Andreolini, M., Ferretti, L., Marchetti, M., & Colajanni, M. (2022). Modeling realistic adversarial attacks against network intrusion detection systems. Digital Threats: Research and Practice, 3(3), 1–28. https://doi.org/10.1145/3469659

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419. https://doi.org/10.1016/j.jisa.2019.102419

Ferretti, L., Marchetti, M., & Colajanni, M. (2022). Federated learning for cyber security: A systematic review. IEEE Access, 10, 63523–63536. https://doi.org/10.1109/ACCESS.2022.3181367

Garcia, S., Grill, M., Stiborek, J., & Zunino, A. (2014). An empirical comparison of botnet detection methods. Computers & Security, 45, 100–123. https://doi.org/10.1016/j.cose.2014.05.011

Habibi Lashkari, A., Draper-Gil, G., Mamun, M. S. I., & Ghorbani, A. A. (2017). Characterization of Tor traffic using time based features. Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP), 253–262. https://doi.org/10.5220/0006105602530262

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. Cybersecurity, 2(1), 20. https://doi.org/10.1186/s42400-019-0038-7

Li, Y., Ma, L., Dong, R., & Chang, K. (2019). A survey on federated learning systems: Vision, hype and reality for data

privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 35(4), 3347–3366. https://doi.org/10.1109/TKDE.2021.3124599

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

Lu, Y. (2022). Implementing blockchain in information systems: A review. Enterprise Information Systems, 16(12), 1876–1907. https://doi.org/10.1080/17517575.2021.2008513

Lu, Y. (2023). Blockchain technology: Recent research and future trend. Enterprise Information Systems, 16(12), 1939895. https://doi.org/10.1080/17517575.2021.1939895

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215–234. https://doi.org/10.1007/s10796-021-10221-w

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. Proceedings of the 25th Annual Network and Distributed System Security Symposium (NDSS). https://doi.org/10.14722/ndss.2018.23025

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. 2015 Military Communications and Information Systems Conference (MilCIS), 1–6. https://doi.org/10.1109/MilCIS.2015.7348942

Peng, C., Xu, M., Xu, S., & Hu, T. (2019). Modeling and predicting cyber hacking breaches. IEEE Transactions on Information Forensics and Security, 12(6), 1274–1285. https://doi.org/10.1109/TIFS.2017.2656422

Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. Computers & Security, 86, 147–167. https://doi.org/10.1016/j.cose.2019.06.005

Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), 108–116. https://doi.org/10.5220/0006639801080116

Sharafaldin, I., Habibi Lashkari, A., Hakak, S., & Ghorbani, A. A. (2019). Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. 2019 International Carnahan Conference on Security Technology (ICCST). https://doi.org/10.1109/CCST.2019.8888419

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. 2010 IEEE Symposium on Security and Privacy, 305–316. https://doi.org/10.1109/SP.2010.25

Strom, B. E., Applebaum, A., Miller, D. P., Nickels, K. C., Pennington, A. G., & Thomas, C. B. (2018). MITRE ATT&CK: Design and philosophy. MITRE Corporation Technical Report. https://doi.org/10.13140/RG.2.2.20520.47361

Stolfo, S. J., Fan, W., Lee, W., Prodromidis, A., & Chan, P. K. (1999). Cost-based modeling for fraud and intrusion detection. DARPA Information Survivability Conference and Exposition (DISCEX), 130–144. https://doi.org/10.1109/DISCEX.2000.821515

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2010). A detailed analysis of the KDD CUP 99 data set. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 1–6. https://doi.org/10.1109/CISDA.2009.5356528

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452–10473. https://doi.org/10.1109/JIOT.2021.3060508

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

[1] Department of Computer Science and Cybersecurity, University of the West of Scotland, Paisley PA1 2BE, UK

[2] Faculty of Information Technology, Reutlingen University, 72762 Reutlingen, Germany

[3] School of Computing, Engineering and Intelligent Systems, Ulster University, Derry BT48 7JL, UK. *Email: s.murphy@ulster.ac.uk (Corresponding Author). https://doi.org/10.63646/datamind.2024.020305