

Article

A Data-Centric Network Traffic Dataset for Anomaly Detection: Construction, Reproducible Pipeline, and Technical Validation

Daniel Quirumbay Yagual ^{1,2,*}, Diego Fernández Iglesias ², Francisco J. Nóvoa ² and Daniel Garabato ²

¹ Faculty of System and Telecommunications, Santa Elena State University (UPSE), La Libertad 240204, Santa Elena, Ecuador

² Center for Information and Communications Technologies Research (CITIC), Department of Computer Science and Information Technologies, University of A Coruña, 15071 A Coruña, Spain; diego.fernandez@udc.es (D.F.I.); francisco.javier.novo@udc.es (F.J.N.); daniel.garabato@udc.es (D.G.)

* Correspondence: dquirumbay@upse.edu.ec

Abstract

The effectiveness of machine learning and deep learning methods for network anomaly detection depends strongly on the quality and representativeness of the datasets used for training and evaluation. Despite recent advances, many publicly available benchmarks rely on synthetic traffic, outdated attack scenarios, or limited representation of encrypted communications. This work presents a network traffic dataset derived from operational firewall logs collected in a heterogeneous institutional environment dominated by HTTPS/TLS traffic. A structured data-centric pipeline was implemented, including preprocessing, behavioral feature engineering, unsupervised pseudo-labeling through the EFMS–KMeans algorithm, class balancing using SMOTE, and the generation of model-oriented sequential representations for deep learning analysis. The resulting dataset contains large-scale flow-level records describing volumetric, behavioral, and temporal traffic characteristics while preserving privacy through anonymization procedures. Technical validation was conducted using statistical analysis, entropy-based measurements, clustering quality metrics, and dimensionality reduction techniques, confirming data consistency, structural diversity, and class separability. The dataset is publicly available through the Mendeley Data repository together with metadata and documentation supporting anomaly detection research, encrypted traffic analysis, and the evaluation of machine learning and deep learning approaches in realistic cybersecurity environments.

Keywords: anomaly detection; data-centric cybersecurity; encrypted traffic; network traffic dataset; pseudo-labeling; reproducible pipeline



Academic Editor: Keke Chen

Received: 9 June 2026

Revised: 31 July 2026

Accepted: 4 August 2026

Published: 6 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The effectiveness of modern Intrusion Detection Systems (IDSs) and anomaly detection approaches based on Machine Learning (ML) and Deep Learning (DL) is fundamentally conditioned by the quality, structure, and representativeness of the datasets used for training and evaluation. Although detection architectures have evolved rapidly, significant limitations persist in the design, documentation, and realism of available datasets. In particular, the increasing prevalence of encrypted communications, the heterogeneity of network environments, and the continuous evolution of cyber threats demand datasets that accurately reflect real operational conditions.

A growing body of work has highlighted the central role of dataset design in the development of reliable anomaly detection systems. Ferriyan et al. [1] introduced the HIKARI-2021

dataset, emphasizing the importance of incorporating encrypted traffic and defining transparent and reproducible data generation procedures. Similarly, Alshaibi et al. [2] conducted a comparative analysis of widely used cybersecurity datasets, demonstrating that current research is constrained not only by the architectural complexity of ML and DL models but also by the quality, representativeness, and heterogeneity of available data. In parallel, Choi et al. [3] showed that high-dimensional raw traffic representations introduce scalability challenges and increase the risk of overfitting, highlighting the need for effective preprocessing and feature engineering strategies.

The emergence of domain-specific datasets further reinforces the importance of capturing the operational context in which network traffic is generated. Ayodele and Buttigieg [4] demonstrated that traffic behavior in virtualized environments differs substantially from traditional infrastructures, indicating that datasets lacking contextual realism may lead to limited generalization capability. Likewise, recent contributions published in the *Data* journal emphasize that datasets must be treated as first-class scientific artifacts, requiring reproducibility, traceability, and methodological transparency as essential components of their design [5,6]. This perspective aligns with the broader transition toward data-centric artificial intelligence, in which data quality and engineering are recognized as primary determinants of system performance.

Recent studies on encrypted traffic analysis indicate that appropriate feature representation is essential for effective deep learning. Structured behavioral representations extracted from flow metadata enable convolutional and sequential models to capture discriminative traffic patterns even when payload inspection is unavailable due to encryption [7,8].

Despite these advances, the literature consistently reports a set of recurring limitations in existing datasets. In general, many datasets rely on synthetic traffic generation, exhibit outdated behavioral patterns, provide insufficient representation of encrypted communications, and fail to capture realistic attack distributions and operational heterogeneity. These limitations reduce the external validity of experimental results and hinder the deployment of ML/DL-based detection systems in real-world environments [1,2,9]. Furthermore, such deficiencies often lead to biased learning processes and overly optimistic performance evaluations that do not generalize to operational scenarios.

Beyond these limitations, recent research highlights persistent challenges in ML security, robustness, and data dependency. Fragmented vulnerability analysis and the absence of unified evaluation frameworks underscore the need for more rigorous dataset design and validation practices [10,11]. Likewise, trade-offs among model complexity, accuracy, and resilience reinforce the central role of high-quality data in developing reliable ML systems [10].

Previous work demonstrated that hybrid anomaly detection architectures based on EFMS–KMeans clustering and CNN–GRU models can effectively exploit behavioral representations extracted from operational network traffic [12]. However, the datasets generated during these studies were not formally documented or released as reusable scientific resources. This work addresses that gap by publishing the complete dataset together with its construction pipeline and technical validation.

Building upon these observations, datasets derived from real operational traffic with behavioral feature representations and reproducible processing pipelines remain essential for advancing anomaly detection research. Since real-world traffic rarely provides explicit labels, unsupervised and semi-supervised strategies have become fundamental for constructing datasets under class imbalance and evolving threat conditions [13,14]. Consequently, data-centric approaches play a key role in improving dataset quality, reproducibility, and scientific reusability in cybersecurity research.

The main contribution of this work is the public release of a data-centric network traffic dataset derived from real operational FortiGate firewall logs collected in an environment dominated by encrypted communications. Unlike existing datasets based on synthetic traffic or limited documentation, this resource is accompanied by a fully documented generation pipeline comprising preprocessing, behavioral feature engineering, EFMS–KMeans pseudo-labeling, class balancing, sequential sliding-window representation, and technical validation. Together with comprehensive metadata and supporting documentation, the released data facilitate reproducible research and benchmarking for behavioral anomaly detection under realistic encrypted traffic conditions.

The objective of this study is to construct, technically validate, and publicly release a reproducible flow-level network traffic dataset to support methodological research and the evaluation of machine learning and deep learning methods for behavioral anomaly detection. Because manually verified attack annotations are unavailable in operational traffic, the released labels correspond to behavioral pseudo-labels generated through unsupervised clustering and should therefore be interpreted as statistically significant deviations from dominant traffic patterns rather than confirmed malicious events. Consequently, the generalizability of the dataset should be interpreted considering the characteristics and traffic composition of the target operational environment.

The remainder of this paper is organized as follows: Section 2 describes the complete dataset generation pipeline, including data acquisition, preprocessing, behavioral feature engineering, behavioral pseudo-label generation, class balancing, sequential sliding-window representation, and the reproducibility protocol. Section 3 presents the published data records and repository organization. Section 4 reports the statistical, geometric, structural, and clustering validation procedures used to assess the consistency of the generated behavioral pseudo-labels. Section 5 discusses dataset usage recommendations, limitations, and reproducibility considerations. Finally, Section 6 summarizes the main contributions and outlines future research directions.

2. Dataset Generation Pipeline

2.1. Overview and Design Principles

The objective of this section is to describe the complete and reproducible workflow followed to construct the proposed network traffic dataset. Unlike model-centric studies, this work considers the dataset itself as the primary scientific contribution. Accordingly, the proposed pipeline emphasizes data quality, traceability, reproducibility, and methodological transparency throughout all stages, including traffic acquisition, preprocessing, behavioral feature engineering, pseudo-labeling, balancing, sequential sliding-window representation, validation, and public release.

The dataset was designed to support reproducible and reusable cybersecurity research under realistic encrypted traffic conditions. In particular, the published resource supports binary anomaly detection, maintains compatibility with both conventional machine learning and deep learning architectures, preserves the statistical and behavioral characteristics of real operational traffic through anonymization, and facilitates independent validation and reproducible experimentation.

The methodology was developed under three fundamental design principles:

- **Reproducibility:** explicit definition and documentation of the complete dataset generation pipeline.
- **Traceability:** transparent description of every transformation stage from raw traffic acquisition to label generation.
- **Experimental validity:** construction based on real operational traffic and formal statistical validation of dataset properties.

The complete pipeline comprises the following stages: data acquisition, preprocessing, behavioral feature engineering, unsupervised pseudo-labeling, class balancing, temporal structuring, and final dataset generation.

The proposed dataset follows recent data-centric AI principles by emphasizing data quality, methodological transparency, and reproducibility throughout the complete dataset lifecycle. Comprehensive documentation, metadata, and validation procedures are provided to facilitate independent verification, benchmarking, and long-term reuse of the published resource [15]. Previous studies have emphasized that well-documented datasets facilitate experimental consistency, cross-study comparability, and long-term reuse, while insufficient transparency during dataset construction may introduce biases, data leakage, and evaluation inconsistencies [15,16].

Recent research further highlights the importance of structured dataset engineering practices throughout the machine learning lifecycle, particularly in domains where data quality directly affects model performance and evaluation reliability [17,18]. Accordingly, the dataset is released as a benchmark resource for anomaly detection research in encrypted traffic environments, accompanied by documentation describing its generation process, repository resources, and validation procedures.

2.2. Data Acquisition Environment

The dataset was constructed from real records generated by perimeter security devices (Next-Generation Firewalls) deployed within a heterogeneous institutional network. In particular, the dataset was derived from the experimental framework previously introduced by Quirumbay Yagual et al. [12], where real network traffic was collected using an operational FortiGate firewall integrated into an academic infrastructure. Unlike synthetic or laboratory-generated datasets, the captured traffic reflects a real-world production environment that includes academic services, administrative activity, IoT communications, and diverse web-based interactions dominated by encrypted HTTPS/TLS traffic.

This operational environment enabled the capture of operational traffic patterns, including the following:

- Encrypted TLS traffic, including HTTPS-based communications operating over TCP and QUIC protocols.
- Academic, administrative, and IoT traffic.
- TCP, UDP, and ICMP protocols.
- Real normal and anomalous behavioral patterns.

Traffic acquisition was performed at the bidirectional flow level following a flow-based monitoring approach, which is widely adopted in encrypted traffic analysis scenarios where payload inspection may be operationally restricted, computationally expensive, or limited by privacy-preserving encryption mechanisms [19]. Instead of relying on packet contents, this methodology enables the modeling of network behavior through statistical metadata and flow-level descriptors, including traffic volume, duration, communication rates, temporal information, and behavioral indicators.

In encrypted traffic environments, flow-level metadata constitute the primary source of analytical information, since payload inspection becomes limited under modern encryption protocols [14]. Consequently, the resulting dataset captures realistic behavioral dynamics associated with contemporary encrypted communications while preserving compatibility with privacy-preserving monitoring strategies.

The collected records include attributes associated with source and destination interactions, ports, traffic volume, flow duration, temporal metrics, application behavior, and security-related indicators derived from firewall monitoring processes. To preserve privacy and ensure ethical compliance, sensitive identifiers were anonymized during preprocessing

while maintaining the statistical and structural properties required for reproducible cybersecurity research. The anonymization process was completed before any public release of the dataset.

The operational firewall logs used in this study were obtained from the institutional cybersecurity infrastructure with the formal authorization of the Director of Information Technology of the institution for academic research purposes. Prior to preprocessing and public release, all identifiers capable of revealing operational infrastructure or individual users were anonymized. The published dataset therefore contains only anonymized flow-level metadata and derived behavioral features required for reproducible cybersecurity research, excluding packet payloads and personally identifiable information.

The network traffic analyzed in this study was collected from a FortiGate 2500E next-generation firewall deployed at the institutional Internet gateway. The monitored infrastructure represents a real academic production network predominantly characterized by encrypted HTTPS/TLS communications. The main characteristics of the operational data collection environment are summarized in Table 1. Native FortiOS traffic logs provide protocol information, ports, byte and packet counters, application identification, security actions, application risk levels, destination reputation indicators, and session statistics, from which the behavioral features described in Section 2.4 were subsequently derived.

Table 1. Characteristics of the operational data collection environment.

Characteristic	Description
Firewall platform	FortiGate 2500E Next-Generation Firewall (NGFW)
Operating system	FortiOS
Deployment location	Institutional Internet gateway
Collection period	February, March, and August 2025
Traffic source	Bidirectional FortiGate traffic logs
Network environment	Academic institutional campus
User profiles	Students, faculty members, administrative staff
Approximate network scale	Approximately 200 connected endpoints during the acquisition campaigns.
Communication profile	Predominantly encrypted HTTPS/TLS traffic
Published variables	Native FortiGate traffic log attributes and derived behavioral features

Table 1 summarizes the operational characteristics of the monitored environment. Collecting traffic from a production academic network provides realistic communication patterns generated by heterogeneous users and services while preserving the behavioral diversity required for anomaly detection research. At the same time, these characteristics define the scope of the published dataset, whose behavioral patterns should be interpreted within the context of the monitored infrastructure.

2.3. Preprocessing Pipeline

The raw data were subjected to a systematic cleaning and transformation process defined as

$$D_p = P(D_0) = \text{Normalize}(\text{Encode}(\text{Clean}(D_0))) \quad (1)$$

This process included the following:

- Data cleaning and schema harmonization;
- Encoding of categorical variables;
- Normalization of numerical attributes;
- Filtering of non-informative features.

In the preprocessing stage, the feature count ips was observed to present a constant value across all traffic records. The variable was intentionally preserved in the dataset to

maintain compatibility with the original firewall log schema and with previous studies that use the same feature space. Although the feature does not contribute statistical variance in the current dataset, retaining it ensures structural consistency and facilitates interoperability with related datasets and experimental pipelines.

Preprocessing was designed as a record-preserving transformation that improved the consistency, structural coherence, and statistical stability of the feature representation without modifying the number of observed network flows, which is particularly important in dynamic encrypted traffic environments [20]. The raw firewall dataset (D_0) and the processed dataset (D_p) both contain 951,560 records, corresponding to a retention rate of 100%. Consequently, no flow records were removed during preprocessing.

The preprocessing stage focused on schema harmonization, categorical encoding, numerical normalization, feature engineering, and the anonymization of sensitive identifiers. Fields containing missing values in the original FortiGate export corresponded to optional inspection attributes that are naturally absent when the associated security module, application, or security event is not triggered. These fields were standardized during the transformation process without deleting the corresponding flow records, thereby preserving complete traceability between the raw and processed datasets while producing a consistent representation for subsequent analysis [20].

To clarify the distinction between removed duplicates and repeated behavioral profiles, an audit of the original firewall dataset confirmed that no exact duplicate records were present when all original exported attributes were considered. Consequently, duplicate elimination did not modify the number of observations. In contrast, repeated behavioral profiles were intentionally preserved because independent network flows may legitimately exhibit identical behavioral characteristics after feature engineering. The audit identified 5486 repeated behavioral profile groups involving 19,619 observations (2.06% of the processed dataset), representing recurrent operational activities generated by applications, services, and users rather than duplicated acquisition records. Preserving these repeated behavioral profiles maintains the empirical traffic distribution and improves the representativeness of the released dataset for behavioral anomaly detection.

2.4. Feature Engineering

A behavioral-oriented feature engineering process was designed to capture intrinsic traffic properties from encrypted and heterogeneous network environments. After preprocessing, each network flow instance x_i was transformed into a behavioral representation vector defined as

$$\mathbf{z}_i = F(\mathbf{x}_i) \quad (2)$$

where $F(\cdot)$ denotes the feature transformation and behavioral extraction process applied to the original flow representation.

Unlike payload-dependent inspection methods, the proposed representation relies exclusively on flow-level metadata and behavioral descriptors extracted from bidirectional communications. This approach is particularly suitable for encrypted traffic environments, where packet payload inspection becomes infeasible due to TLS/HTTPS encryption mechanisms.

The engineered feature space integrates heterogeneous variables describing network behavior across multiple dimensions, including the following:

- Volume and intensity metrics;
- Flow behavior ratios;
- Temporal variables;
- Risk and reputation indicators;
- Service and application-related attributes.

To improve behavioral representation, several derived variables were generated from the original firewall flow records. In particular, volumetric aggregation features were constructed to model communication intensity and transfer dynamics:

$$\text{bytes_total} = \text{sentbyte} + \text{rcvbyte} \quad (3)$$

$$\text{pkt_total} = \text{sentpkt} + \text{rcvpkt} \quad (4)$$

These variables represent the total communication volume exchanged during each bidirectional flow and enable the characterization of traffic burstiness and transfer intensity.

To capture the flow efficiency and communication rate, the following normalized traffic-rate feature was computed:

$$\text{bytes_per_sec} = \frac{\text{bytes_total}}{\text{duration} + 1} \quad (5)$$

where the constant term avoids division instability in short-duration flows.

In addition, relational traffic dynamics between forward and backward communications were modeled using packet asymmetry indicators:

$$\text{pkt_ratio} = \frac{\text{sentpkt} + 1}{\text{rcvpkt} + 1} \quad (6)$$

This feature captures directional imbalance patterns frequently associated with anomalous or irregular communication behaviors.

Temporal contextualization was incorporated by extracting temporal descriptors from the original event timestamps, including the following:

- Hour of communication (hour).
- Day of week (dayofweek).

These temporal variables enable the modeling of periodic behavioral patterns and temporal persistence in operational traffic environments.

The resulting dataset combines statistical, behavioral, and temporal descriptors extracted from bidirectional network flows. The released features include volumetric traffic metrics (e.g., `sentbyte`, `rcvbyte`, `bytes_total`, and `pkt_ratio`), risk-related attributes such as `dstreputation_cat` and `apprisk`, firewall policy and security action indicators, and temporal descriptors (hour and dayofweek).

The resulting feature space combines statistical, behavioral, and temporal descriptors extracted from bidirectional network flows, providing a payload-independent representation particularly suitable for encrypted traffic analysis. Previous studies have shown that such metadata-driven representations effectively support anomaly detection and deep learning models under encrypted communication scenarios [12,20,21].

2.5. Formal Pipeline Representation

The dataset generation process can be formally represented as a sequence of transformations applied to the raw dataset D_0 . First, preprocessing and behavioral feature engineering are performed, followed by unsupervised pseudo-labeling using the EFMS–KMeans algorithm. This stage produces the pseudo-labeled dataset defined as

$$D_I = L_{\text{EFMS}}(F(P(D_0))) \quad (7)$$

where $P(\cdot)$ represents the preprocessing stage, $F(\cdot)$ denotes behavioral feature engineering, and $L_{\text{efms}}(\cdot)$ corresponds to the pseudo-labeling process performed by the EFMS–KMeans clustering algorithm.

To address the natural imbalance of anomalous traffic in operational environments, a balanced dataset version is generated using the Synthetic Minority Over-sampling Technique (SMOTE):

$$D_b = B_{\text{SMOTE}}(D_l) \quad (8)$$

where $B_{\text{smote}}(\cdot)$ represents the class balancing operation applied to the pseudo-labeled dataset. This transformation generates synthetic minority-class instances while preserving the local structure of the feature space, improving class representation for supervised learning experiments [22].

Finally, to support sequential modeling, the balanced dataset is transformed into ordered sliding-window representations. The windows are generated from the deterministic ordering of the released tabular representation to satisfy the input requirements of the CNN-GRU architecture. Because the original operational traffic is not stored in globally chronological order, these windows should be interpreted as computational sequential representations rather than chronological reconstructions of network events:

$$D_f = T_{\text{temp}}(D_b) \quad (9)$$

where D_f denotes the Sequential Training Representation generated for deep learning experiments.

The Sequential Training Representation is distributed in the public repository to facilitate reproducible experimentation. However, it is a model-oriented computational representation created specifically for sequential deep learning experiments rather than an independent scientific dataset. Throughout this work, the EFMS-KMeans behavioral pseudo-labeled dataset (D_l) is regarded as the primary published dataset, whereas D_f is provided exclusively to support reproducible CNN-GRU training. Unlike the behavioral pseudo-labeled dataset D_l , which preserves the observed operational traffic distribution, the sequential training representation D_f is a model-oriented representation produced to satisfy the input requirements of sequential deep learning models. Accordingly, D_f should be interpreted as an intermediate training representation rather than as an additional released dataset or a chronological reconstruction of the operational network traffic.

2.6. Unsupervised Labeling via EFMS-KMeans

In this dataset, an anomaly denotes a statistically significant deviation from the dominant behavioral patterns observed in the multidimensional feature space. Because the collected traffic originates from an operational network without expert-verified ground-truth attack annotations, these labels should be interpreted as behavioral pseudo-labels rather than confirmed malicious events.

Accordingly, the EFMS-KMeans algorithm was adopted to generate behavioral pseudo-labels from statistically normalized traffic representations, enabling the construction of labeled datasets without requiring manual annotation.

Due to the absence of manually annotated labels in real-world operational traffic, an unsupervised pseudo-labeling strategy based on the EFMS-KMeans clustering algorithm was applied to identify normal and anomalous traffic behaviors. This strategy allows behavioral pseudo-labels to be inferred directly from the statistical structure of operational traffic without requiring expert annotations.

The pseudo-labeling process was performed after preprocessing and behavioral feature engineering, allowing the clustering algorithm to operate on statistically normalized traffic representations. Each network flow was grouped according to its structural similarity within the multidimensional feature space, enabling the identification of latent behavioral deviations independently of packet payload inspection.

The anomaly label assignment was defined according to the Euclidean distance between each flow instance and its corresponding cluster centroid, computed in the normalized multidimensional feature space obtained after preprocessing and behavioral feature engineering:

$$\text{label}(\mathbf{x}_i) = \begin{cases} 1, & d(\mathbf{x}_i, \mu_k) > P_{95}, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

where $d(x_i, \mu_k)$ represents the Euclidean distance between instance x_i and centroid μ_k , while P_{95} denotes the 95th percentile of the distance distribution within each cluster. The 95th percentile provides a cluster-specific statistical criterion for identifying observations located in the upper tail of the distance distribution. Consequently, observations exceeding this threshold are assigned behavioral pseudo-labels without requiring manual annotations [21].

Therefore, the labels generated through Equation (10) constitute behavioral pseudo-labels inferred from the statistical structure of the feature space. Consequently, the anomalous class identifies observations whose behavioral characteristics differ significantly from the dominant traffic patterns and should not be interpreted as expert-validated attack annotations.

The variable `distance_to_centroid` was introduced exclusively as an auxiliary feature during the EFMS–KMeans pseudo-label generation stage to quantify the Euclidean distance between each traffic flow and its assigned cluster centroid. Because this variable directly contributes to the construction of the behavioral pseudo-labels defined in Equation (10), it is retained in the released `clustered_data_efms_kmeans.csv` dataset solely to ensure the reproducibility and transparency of the pseudo-labeling procedure.

Prior to supervised learning, the auxiliary variable `distance_to_centroid` was removed from the feature set used to construct the CNN–GRU training representation. Consequently, the supervised model was trained exclusively on behavioral traffic features, thereby preventing information leakage between the pseudo-label generation stage and the subsequent classification model.

To justify the percentile threshold adopted in Equation (10), a sensitivity analysis was conducted using three candidate values: P90, P95, and P99. For each candidate threshold, behavioral pseudo-labels were regenerated from the same centroid-distance distribution while modifying only the percentile criterion. This procedure allowed the effect of threshold selectivity on the number and proportion of flows assigned to the anomalous class to be quantified. Table 2 and Figure 1 summarize the results.

Table 2. Sensitivity analysis of the percentile-based behavioral pseudo-labeling threshold.

Percentile	Distance Threshold	Anomalous Flows	Percentage (%)
P90	4.8282	95,156	10.00
P95	6.7009	47,578	5.00
P99	8.5072	9516	1.00

As expected, the proportion of flows assigned to the anomalous class decreased as the percentile threshold became more restrictive. P90 classified 95,156 flows as behavioral anomalies, corresponding to 10.00% of the dataset, whereas P99 retained only 9516 highly extreme observations, representing approximately 1.00%. The P95 criterion identified 47,578 flows, equivalent to 5.00% of the analyzed records, thereby providing an intermediate operating point between the broader P90 criterion and the more restrictive P99 criterion.

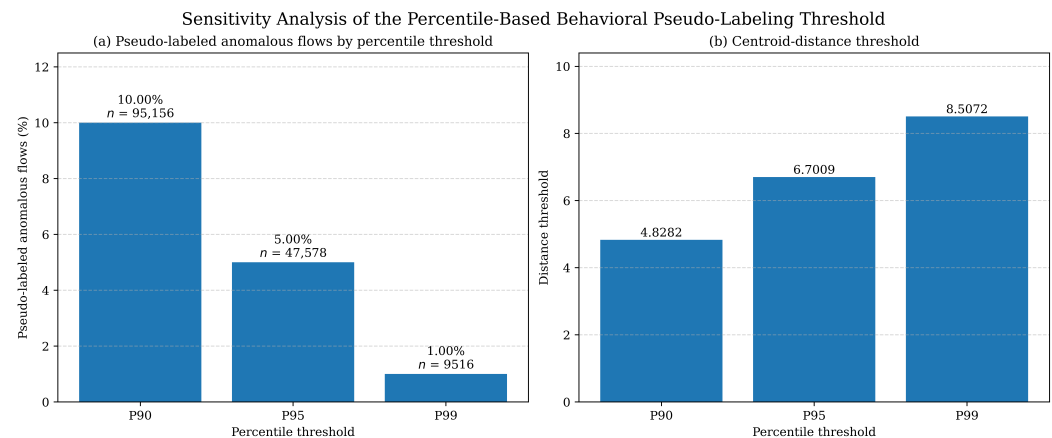


Figure 1. Sensitivity analysis of the percentile-based behavioral pseudo-labeling threshold. Panel (a) illustrates the proportion of flows assigned to the anomalous class under the P90, P95, and P99 criteria, whereas panel (b) presents the corresponding centroid-distance thresholds. The analysis supports the selection of P95 as the operational threshold adopted in this work.

Based on this sensitivity analysis, P95 was selected as the operational threshold because it provides a balanced compromise between behavioral selectivity and minority-class representation. Compared with P90, it reduces the inclusion of less pronounced deviations, while, compared with P99, it preserves a sufficiently large set of uncommon observations for subsequent methodological evaluation and benchmarking. This selection should therefore be interpreted as a data-driven operational criterion rather than as evidence that P95 is universally optimal across all network environments.

The EFMS–KMeans released behavioral dataset preserves the original class distribution obtained directly from the operational traffic after preprocessing, without applying sampling or balancing procedures. Consequently, the observed anomaly rate of 5.00% reflects the behavioral pseudo-labeling criterion defined in Equation (10) rather than the prevalence of confirmed cyberattacks in the monitored network.

The clustering process generated the pseudo-labeled dataset stored as `clustered_data_efms_kmeans.csv`, preserving the original number of records obtained after preprocessing. The published dataset retains the auxiliary variable `distance_to_centroid` to ensure the reproducibility and transparency of the EFMS–KMeans pseudo-label generation process. Since this variable is exclusively associated with pseudo-label construction, it is excluded from the feature set used to train the subsequent CNN–GRU classifier, thereby preventing information leakage during supervised learning.

The resulting dataset contains a total of 951,560 bidirectional network flows distributed as follows:

- Normal traffic: 903,982 flows;
- Anomalous traffic: 47,578 flows.

This anomaly proportion corresponds to the behavioral pseudo-labeling criterion defined in Equation (10) and reflects the original class distribution preserved in the published EFMS–KMeans behavioral pseudo-labeled dataset prior to any model-oriented preprocessing. Class imbalance constitutes a fundamental challenge in anomaly detection tasks, where anomalous events typically represent only a small fraction of the observed traffic. Previous studies have shown that clustering-based pseudo-labeling strategies combined with balancing mechanisms improve the representation of minority behavioral patterns while preserving the structural properties of the feature space [23].

For downstream deep learning experiments, an additional derived dataset compatible with the CNN–GRU architecture was generated through feature filtering and sub-

set selection applied after the pseudo-labeling stage. The resulting dataset, stored as `labeled_dataset_for_GRU-CNN.csv`, contains 554,647 records and 29 selected features, distributed as follows:

- Normal traffic: 353,557 records
- Anomalous traffic: 201,090 records

Unlike the EFMS–KMeans behavioral pseudo-labeled dataset, which preserves the original behavioral distribution obtained from the operational traffic, the sequential training representation is a model-oriented representation produced for model optimization. Consequently, any differences in class proportions or statistical properties reflect preprocessing operations required for supervised learning rather than modifications of the behavioral pseudo-labeling criterion or changes in the underlying operational traffic distribution.

This derived representation was specifically prepared to support sequential deep learning experiments while retaining the statistical and behavioral information extracted from the original operational traffic environment. The resulting representation facilitates the learning of discriminative temporal and behavioral patterns associated with anomalous network activity under encrypted traffic conditions [12,22].

2.7. Preparation of the Sequential Training Representation

Although the EFMS–KMeans behavioral pseudo-labeled dataset constitutes the primary scientific artifact released in this work, sequential deep learning architectures require additional preprocessing before model training. These transformations generate reproducible intermediate representations solely to satisfy the input requirements of sequential models and are not part of the published dataset. Therefore, the released dataset remains unchanged, preserving its statistical properties, traceability, and reproducibility. This distinction is consistent with the principles of data-centric artificial intelligence, where the dataset is the primary scientific asset and model-specific preprocessing supports downstream learning without modifying the published resource [24].

Starting from the behavioral pseudo-labeled dataset (D_l), a separate tabular representation was prepared for the downstream CNN–GRU experiment. This transformation did not modify the behavioral pseudo-labeling criterion or the primary released dataset. Instead, it generated an intermediate computational representation specifically intended to support supervised CNN–GRU training.

The preparation workflow was intentionally designed as a deterministic, model-specific transformation that preserved the behavioral labels assigned during the EFMS–KMeans stage while adapting the data to the requirements of supervised sequential learning. The workflow comprised three principal transformations applied to construct the derived representation used in the subsequent supervised learning pipeline.

The resulting representation was obtained through a deterministic preprocessing procedure applied after behavioral pseudo-label generation. This procedure comprised (i) the removal of auxiliary variables used exclusively during the EFMS–KMeans pseudo-labeling stage (e.g., `distance_to_centroid`); (ii) the retention of the 29 behavioral, volumetric, security-related, and temporal variables included in the derived representation; and (iii) the construction and export of the intermediate tabular representation used in the supervised CNN–GRU workflow. The resulting file contains 554,647 observations and preserves the behavioral pseudo-labels associated with this derived representation.

First, auxiliary variables used exclusively during behavioral pseudo-label generation, particularly `distance_to_centroid`, were removed from the supervised input space because they directly participated in the computation of the behavioral labels and could therefore introduce information leakage.

Second, only the behavioral, volumetric, security-related, and temporal features required by the CNN–GRU classifier were retained, producing a consistent feature representation while preserving the original behavioral labels assigned to every observation.

Finally, the resulting tabular representation was exported as `labeled_dataset_for_GRU-CNN.csv`, thereby creating the computational representation adopted as the input to the subsequent supervised learning pipeline.

The resulting file, `labeled_dataset_for_GRU-CNN.csv`, contains 554,647 tabular observations, comprising 353,557 normal observations (63.75%) and 201,090 behaviorally anomalous observations (36.25%). This class composition corresponds to the derived representation before the application of SMOTE. It should not be interpreted as the post-SMOTE training distribution or as a modification of the P95 pseudo-labeling criterion. Likewise, it should not be considered a new estimate of anomaly prevalence in the operational network.

This distinction is important for reproducibility because the released CSV file represents the complete tabular representation, whereas the sequential tensors required by the CNN–GRU architecture are generated deterministically during the supervised training pipeline after the data have been partitioned into training and testing subsets.

2.8. Class Balancing

Following the preparation of the model-oriented tabular representation described in Section 2.7, the dataset was partitioned into mutually exclusive training and testing subsets for supervised CNN–GRU development. Class balancing was subsequently performed exclusively on the training partition using the Synthetic Minority Over-sampling Technique (SMOTE), whereas the testing partition preserved its original class distribution without synthetic observations. This strategy avoids information leakage and provides an unbiased evaluation of the trained classifier [23].

The SMOTE configuration adopted in this work is summarized in Table 3. Synthetic minority-class observations were generated only within the training subset until both classes contained the same number of training instances.

The generation of a synthetic observation is defined as.

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda (\mathbf{x}_j - \mathbf{x}_i), \quad \lambda \in [0, 1]. \quad (11)$$

where $\mathbf{x}_i$ denotes a minority-class observation, $\mathbf{x}_j$ corresponds to one of its nearest minority-class neighbors, and λ is a random interpolation coefficient used to interpolate between neighboring observations [23].

Accordingly, SMOTE modifies only the training partition used for supervised optimization. Neither the original 5.00% anomaly proportion of the published EFMS–KMeans behavioral pseudo-labeled dataset (D_l) nor the pre-balancing 63.75/36.25 class distribution of the model-oriented representation described in Section 2.7 are altered by this procedure.

The parameter configuration reported in Table 3 enables the balancing stage to be reproduced deterministically from the released model-oriented representation, thereby supporting the reproducibility objectives of the proposed data-centric pipeline.

Table 3. SMOTE configuration adopted during supervised CNN–GRU training.

Parameter	Configuration
Input representation	Sequential Training Representation (Section 2.7)
Sampling stage	Training partition only
Testing partition	No synthetic observations
Sampling strategy	auto
Nearest neighbors (k)	5
Random seed	42
Balancing objective	Equal minority and majority classes in the training partition

2.9. Sequential Sliding-Window Representation

To organize the model-oriented training representation into a format compatible with sequential deep learning architectures, fixed-length overlapping sliding windows were generated from the deterministic ordering of the released tabular representation. This operation creates ordered computational samples suitable for CNN–GRU training without modifying the published behavioral pseudo-labeled dataset.

Although each window contains consecutive observations within the ordered representation, the ordering should not be interpreted as the chronological evolution of the operational network traffic. An audit of the original FortiGate logs showed that the exported records are not globally sorted by event time; therefore, the sliding-window representation preserves local computational neighborhood rather than real temporal continuity. Consequently, the generated windows constitute the Sequential Training Representation designed exclusively to satisfy the input requirements of sequential deep learning architectures.

The temporal structuring process is formally defined as

$$S_t = \{\mathbf{x}_t, \mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+w}\}. \quad (12)$$

where S_t denotes an ordered sliding window composed of consecutive observations in the Sequential Training Representation, and w denotes the adopted window length. The index t identifies the position within the ordered dataset rather than an absolute chronological timestamp.

Each generated window provides an ordered computational context that allows convolutional–recurrent architectures to learn dependencies among neighboring behavioral observations. Accordingly, the resulting representation is intended to satisfy the sequential input requirements of CNN–GRU models rather than to reconstruct the chronological evolution of operational traffic.

Unlike conventional tabular representations, the sliding-window organization enables sequential models to jointly process local behavioral relationships and contextual information extracted from adjacent observations. This representation is particularly useful in encrypted traffic analysis, where behavioral flow metadata constitute the principal source of discriminative information.

The resulting Sequential Training Representation serves as the input representation adopted for CNN–GRU experimentation. Consequently, it should be interpreted as an intermediate computational representation designed for supervised sequential learning rather than as a chronologically ordered reconstruction of the operational network traffic.

Recent studies have shown that sequential input representations can improve the capability of deep learning architectures to capture complex behavioral relationships in encrypted traffic analysis when appropriate behavioral features are available [22].

2.10. Final Dataset Generation

The final dataset integrates all transformation stages defined throughout the proposed data-centric pipeline, including preprocessing, behavioral feature engineering, EFMS–KMeans pseudo-labeling, class balancing, and sequential sliding-window representation. Each stage progressively refines the original firewall traffic records while preserving the statistical and behavioral characteristics required for anomaly detection research in encrypted network environments.

The resulting data ecosystem was designed to support multiple research scenarios while preserving complete traceability across the proposed pipeline. The repository therefore contains the primary published dataset together with the intermediate representations generated for supervised learning, including the Sequential Training Representation used for CNN–GRU experiments.

To provide a quantitative overview of the proposed data-centric pipeline, Table 4 summarizes the principal data resources generated throughout the workflow, including their record counts, class distributions, and anomaly rates. These statistics document the progression from the raw operational traffic to the published behavioral pseudo-labeled dataset and the derived computational representations created for supervised learning and sequential modeling experiments.

Table 4. Repository resources generated throughout the data-centric pipeline.

Repository Resource	Number of Records	Normal (0)	Anomalous (1)	Anomaly Rate (%)
Raw firewall traffic (D_0)	951,560	–	–	–
Processed dataset (D_p)	951,560	–	–	–
EFMS-KMeans pseudo-labeled dataset (D_l)	951,560	903,982	47,578	5.00
Derived Repository Resource				
Sequential Training Representation (repository resource)	554,647	353,557	201,090	36.25

2.11. Reproducibility Protocol

To facilitate independent verification and reproducible experimentation, the complete dataset generation workflow is documented through formal process definitions, parameter specifications, metadata, and publicly available published data resources. Figure 2 summarizes the successive transformations from raw traffic acquisition to the released behavioral dataset and the model-oriented representations, complementing the formal description of the proposed data-centric pipeline.

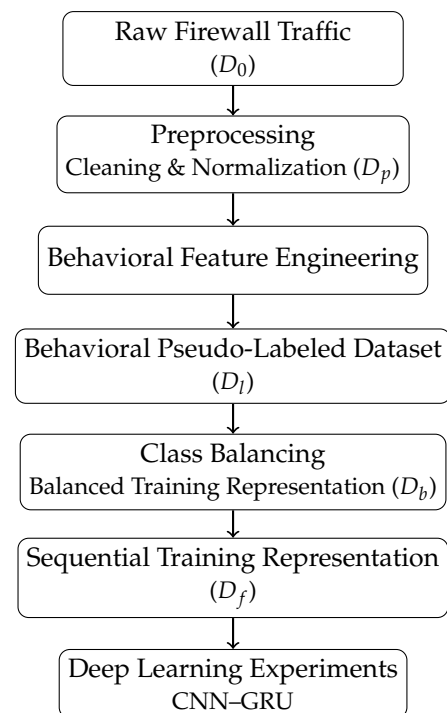


Figure 2. Workflow of the proposed data-centric pipeline illustrating the transformations from raw firewall traffic (D_0) to the published behavioral pseudo-labeled dataset (D_l), followed by the generation of the intermediate computational representations (D_b and D_f) used for supervised CNN-GRU training. Representation D_f is generated through an ordered sliding-window transformation and serves as a computational input for CNN-GRU rather than a chronological reconstruction of the operational network traffic. The consistency of the behavioral pseudo-labels is subsequently assessed through the statistical, geometric, and clustering validation procedures described in Section 4.

3. Data Records

The published repository contains the data resources generated throughout the proposed dataset construction pipeline, from the original firewall traffic logs to the processed behavioral datasets for anomaly detection. Together, these resources ensure the reproducibility and independent validation of the proposed methodology.

3.1. Raw Log Acquisition and Collection Campaigns

The published dataset was derived from three independent acquisition campaigns based on original FortiGate forward-traffic logs collected at the institutional gateway firewall between February and August 2025. Conducting the campaigns at different periods captures multiple operational traffic profiles variability in an operational university network, including changes in user behavior, application workloads, and traffic characteristics. This strategy improves the diversity and representativeness of the released dataset while reducing the risk of reflecting a single network operating condition.

The raw traffic logs constitute the primary data source from which all preprocessing, normalization, anonymization, feature engineering, and behavioral pseudo-labeling stages were performed. Consequently, the statistics presented in this section were computed directly from the original FortiGate logs, ensuring complete traceability between the acquisition process and the published dataset.

The statistics reported in Table 5 were computed directly from the original FortiGate forward-traffic logs prior to preprocessing. The three acquisition campaigns produced a total of 951,560 traffic records, exactly matching the number of observations released through Mendeley Data and thereby preserving a one-to-one correspondence between the original acquisition records and the published dataset. No invalid timestamps were detected during the acquisition-stage audit, confirming the temporal integrity of the collected data before feature extraction, anonymization, and subsequent processing. This consistency ensures that all descriptive statistics reported for the released dataset accurately reflect the characteristics of the original operational traffic without introducing sampling-related distortions during preprocessing.

Table 5. Summary of the original FortiGate acquisition campaigns used to construct the published dataset.

Collection Campaign	Collection Period	Raw Traffic Records	Invalid Timestamps
February 2025	25–26 February 2025	24,176	0
March 2025	6 March 2025	402,994	0
August 2025	7 August 2025	524,390	0
Overall	February–August 2025	951,560	0

The public release of the dataset followed the same anonymization policy applied throughout the preprocessing pipeline. Consequently, the published repository contains only anonymized flow-level metadata together with derived behavioral features required for reproducibility. No information allowing the identification of individual users, devices, institutional services, or communication contents is included in the released files. Furthermore, the repository preserves only the information necessary to support independent validation, comparative benchmarking, and reproducible machine learning experimentation while preventing the disclosure of operationally sensitive infrastructure details or personally identifiable information.

The exported forward-traffic logs contain metadata associated with application identification (App-ID), application risk, UTM actions, and Web-related inspection fields, indicating that multiple FortiGate traffic inspection mechanisms contributed metadata during

traffic processing. In particular, Application Control metadata (appcat, apprisk) and UTM actions were consistently present in the exported records. Because only forward-traffic logs were used to construct the released dataset, the complete runtime configuration of individual security modules (e.g., IPS and WebFilter policies) cannot be independently reconstructed from the available exports.

The encrypted-traffic analysis revealed noticeable differences between the proportion of encrypted flows and the corresponding encrypted byte volume across the acquisition campaigns (Table 6). During the February campaign, 51.80% of the recorded flows used destination port 443, accounting for 80.88% of the total bidirectional traffic volume. In contrast, the March and August campaigns exhibited similar proportions of port-443 flows (58.58% and 56.39%, respectively), whereas the corresponding byte proportions were considerably lower (18.57% and 25.80%). Such variability is expected in an operational university network, where the mix of applications, services, and user activity naturally changes over time.

Table 6. Measured proportion of encrypted traffic in the original FortiGate acquisition campaigns. Destination port 443 was used as a conservative flow-level proxy for HTTPS/TLS (TCP/443) and QUIC (UDP/443) communications.

Collection Campaign	Port-443 Flows (%)	Port-443 Bytes (%)
February 2025	51.80	80.88
March 2025	58.58	18.57
August 2025	56.39	25.80
Overall	57.20	22.87

These observations demonstrate that the released dataset captures multiple operational traffic profiles rather than a single static network condition.

The combination of multiple acquisition campaigns with different traffic compositions increases the diversity and representativeness of the released dataset, providing a realistic basis for the development, benchmarking, and reproducibility of network traffic analysis and anomaly detection methods.

3.2. Repository Information

The dataset generated in this study has been deposited in the Mendeley Data repository and is publicly available for research and educational purposes. The repository contains the complete set of published data resources generated throughout the proposed pipeline, including the processed dataset, the EFMS-KMeans behavioral pseudo-labeled dataset, the Sequential Training Representation, metadata files, preprocessing scripts, and the feature dictionary required to ensure reproducibility.

Although this representation is distributed through the repository to support supervised CNN-GRU experimentation, it is not considered an additional published dataset because it constitutes an intermediate computational representation generated specifically for model training.

The dataset can be accessed through the following repository: <http://doi.org/10.17632/g5dtm7349f>.

The repository is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, allowing reuse, redistribution, and adaptation with appropriate attribution. In addition to the dataset files, the repository provides documentation describing the dataset generation process, variable definitions, and dataset organization to support independent validation and facilitate dataset reuse.

The open publication of structured datasets represents an essential practice in data-centric research, as it enables independent validation, reproducible comparison, and reuse in subsequent machine learning and cybersecurity studies [25,26].

3.3. File Structure

The dataset repository is hierarchically organized to facilitate usability, traceability, and reproducibility across the different stages of the generation pipeline. The directory structure reflects the sequential transformation of the data from raw acquisition to the final sequential sliding-window representation:

The organization shown in Figure 3 reflects the successive processing stages described in Section 2, allowing each published resource to be directly associated with a specific transformation of the proposed pipeline.

The `raw_logs` directory contains the original firewall traffic records captured from the operational environment. The `processed` folder stores the cleaned and normalized dataset used for clustering. The repository also contains the Sequential Training Representation used for the CNN–GRU experiments. This intermediate computational representation is distributed exclusively to facilitate reproducible deep learning experimentation and should not be interpreted as an additional published dataset.

In addition, the `metadata` directory provides complementary documentation, including the dataset description and a detailed feature dictionary that defines variable semantics and transformations.

Figure 3 summarizes the hierarchical organization of the repository and the relationships among the published resources.

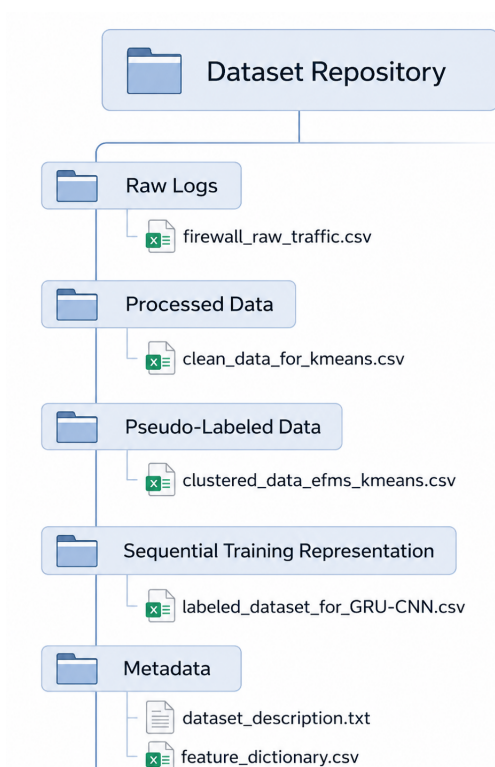


Figure 3. Hierarchical organization of the dataset repository and associated metadata files.

3.4. Data Schema

The final dataset consists of structured network flow records stored in tabular format. Each row represents an individual bidirectional traffic flow described by behavioral, volumetric, security-related, and temporal attributes extracted from the original operational

firewall logs. Together, these variables characterize the communication behavior of each observed flow while preserving the semantic relationships required for downstream statistical analysis, clustering, and supervised machine learning tasks. The schema was designed to provide a consistent and reproducible representation of network traffic suitable for anomaly detection research. The released dataset contains 29 behavioral and traffic-related variables describing volumetric, temporal, statistical, and security attributes extracted from operational bidirectional network flows. The variables included in the dataset comprise the following:

- Flow identifiers;
- Volume and duration metrics;
- Derived behavioral features;
- Temporal variables;
- Binary label (normal/anomalous).

Among the derived behavioral features, the published dataset also retains the auxiliary variable `distance_to_centroid`, which stores the Euclidean distance between each traffic flow and its assigned cluster centroid during the EFMS-KMeans pseudo-label generation stage. This variable is included exclusively to ensure the reproducibility and transparency of the published pseudo-labeling process. It is not incorporated into the feature set used to train the supervised CNN-GRU classifier, thereby preventing information leakage during model training.

The adoption of flow metadata-based representations constitutes a widely accepted strategy in encrypted traffic environments, where packet payload inspection is not feasible [27]. By leveraging statistical, behavioral, and temporal descriptors extracted from bidirectional flows, the dataset preserves sufficient discriminative information for anomaly detection tasks while maintaining compatibility with privacy-preserving monitoring approaches.

3.5. Variable Description Table

To facilitate dataset interpretation and reuse, Table 7 summarizes representative variables included in the published behavioral pseudo-labeled dataset together with their semantic role within the behavioral traffic representation. A complete description of all released variables is provided in the accompanying `feature_dictionary.csv` file available in the Mendeley Data repository.

Table 7. Representative variables included in the published behavioral pseudo-labeled dataset. A complete description of all released variables is provided in the accompanying `feature_dictionary.csv` file available in the Mendeley Data repository.

Variable	Type	Description
<code>proto</code>	Numeric	Transport protocol associated with the network flow.
<code>duration</code>	Numeric	Communication duration of the bidirectional flow.
<code>bytes_total</code>	Numeric	Total number of bytes exchanged during the bidirectional flow.
<code>pkt_total</code>	Numeric	Total number of packets exchanged during the bidirectional flow.
<code>bytes_per_sec</code>	Numeric	Normalized communication rate computed from transferred bytes and flow duration.
<code>pkt_ratio</code>	Numeric	Ratio between forward and backward transmitted packets.
<code>apprisk</code>	Numeric	Application risk level assigned by the FortiGate security platform.
<code>hour</code>	Temporal	Hour extracted from the original traffic timestamp.
<code>dayofweek</code>	Temporal	Day of the week extracted from the original traffic timestamp.
<code>distance_to_centroid</code>	Numeric	Euclidean distance to the assigned EFMS-KMeans cluster centroid used during pseudo-label generation.
<code>label</code>	Binary	Behavioral pseudo-label (0 = Normal, 1 = Anomalous).

3.6. Label Distribution

Table 8 presents the class composition of the Sequential Training Representation used in the CNN-GRU experiments. This representation contains 353,557 instances correspond-

ing to normal traffic and 201,090 instances corresponding to anomalous traffic, representing 63.75% and 36.25% of the total records, respectively.

Unlike the published EFMS–KMeans behavioral pseudo-labeled dataset, whose class proportions preserve the behavioral distribution observed in the operational traffic, these percentages correspond exclusively to the model-oriented training representation generated for the CNN–GRU architecture. Consequently, they should not be interpreted as the natural prevalence of normal and anomalous traffic in the monitored network, but rather as the class composition of an intermediate computational representation designed for supervised sequential learning. Accordingly, the behavioral pseudo-labeled dataset (DI) released with this work remains the appropriate resource for studies requiring the original operational traffic distribution, whereas the Sequential Training Representation is provided exclusively to facilitate reproducible deep learning experiments based on the proposed data-centric pipeline.

Table 8. Class composition of the Sequential Training Representation used in the CNN–GRU experiment.

Class	Records	Percentage
Normal (0)	353,557	63.75%
Anomalous (1)	201,090	36.25%

The class distribution documented in Table 8 provides a balanced basis for supervised and hybrid ML/DL experiments. Class imbalance can significantly affect model generalization and stability; therefore, the application of SMOTE ensures a more reliable training foundation for subsequent supervised and hybrid ML/DL models [26].

Overall, the published repository provides a complete and fully documented collection of operational traffic records, processed behavioral datasets, metadata, feature documentation, and model-oriented representations. Together, these resources support reproducible anomaly detection research while ensuring full traceability across every stage of the proposed data-centric dataset construction pipeline.

4. Technical Validation

The objective of this section is to evaluate the statistical properties, informational diversity, feature-level discriminative capability, and global geometric structure of the proposed dataset through complementary validation analyses. Consistent with the data-centric paradigm, the assessment focuses on the quality and reliability of the published dataset rather than on the predictive performance of a specific machine learning model. Accordingly, the validation aims to demonstrate that the released resource preserves meaningful statistical patterns, structural consistency, and reproducible behavioral representations suitable for methodological research in network anomaly detection.

Feature-level discriminative capability was quantified using Fisher Scores computed after feature standardization with `StandardScaler`, whereas global geometric separability was evaluated using dimensionality reduction techniques together with clustering quality metrics, including the Silhouette coefficient and the Davies–Bouldin Index. Complementary statistical analyses based on entropy measurements were incorporated to assess information diversity, while visual inspection of the feature space through PCA and UMAP provided additional evidence of the consistency and structural coherence of the generated behavioral pseudo-labels. Collectively, these analyses offer a multidimensional validation of the released dataset, supporting its robustness, reproducibility, and suitability as a benchmark for data-centric cybersecurity research.

4.1. Statistical Stability and Distributional Analysis

To assess structural stability, descriptive statistical measures were computed for each feature x_j :

$$\mu_j = \frac{1}{N} \sum_{i=1}^N x_{ij}. \quad (13)$$

$$\sigma_j^2 = \frac{1}{N} \sum_{i=1}^N (x_{ij} - \mu_j)^2. \quad (14)$$

The feature space exhibits positively skewed and heavy-tailed distributions, which are characteristic of real network traffic due to burstiness and heterogeneous communication patterns. No evidence of variance collapse, structural degeneration, or artificial smoothing introduced by preprocessing or balancing was observed.

Understanding distributional behavior is essential, as generalization in deep learning systems is tightly connected to dataset structure and complexity [28]. The preservation of natural traffic dispersion confirms that the dataset retains realistic statistical properties.

4.2. Informational Diversity via Shannon Entropy

To quantify informational diversity within the dataset, Shannon entropy was computed as

$$H(X) = - \sum_{i=1}^n p(x_i) \log p(x_i). \quad (15)$$

where $p(x_i)$ denotes the probability associated with the occurrence of feature value x_i . Shannon entropy is widely used to characterize the variability and distributional complexity of data, providing an information-theoretic measure of behavioral diversity within network traffic.

In the context of encrypted traffic analysis, entropy-based measurements are particularly useful because they capture variations in traffic behavior without relying on payload inspection. Consequently, entropy serves as an auxiliary indicator for assessing the diversity of traffic patterns present in different repository resource and for evaluating the impact of preprocessing and balancing operations on the overall feature space [29].

To complement the entropy-based analysis, a quantitative assessment of the published EFMS-KMeans behavioral pseudo-labeled dataset was performed using objective statistical and structural quality indicators. The evaluation encompasses dataset completeness, repeated behavioral-profile analysis, feature variability, informational diversity, dimensionality, and internal clustering consistency, providing a comprehensive characterization of the dataset from multiple complementary perspectives. Collectively, these quantitative indicators offer objective evidence of the dataset's statistical robustness, structural coherence, and reproducibility, thereby supporting the reliability of the released resource beyond qualitative visual analyses.

Table 9 shows that the published dataset is complete, with no missing values or statistically constant features. Exact-row analysis identified 20,513 observations (2.16%) belonging to repeated behavioral-profile groups, whereas 14,612 observations (1.54%) corresponded to additional repeated feature vectors after retaining one representative profile. A duplicate-removal sensitivity analysis reduced the behavioral anomaly proportion only from 5.00% to 4.79%, indicating that repeated behavioral profiles do not materially alter the overall pseudo-label distribution. Furthermore, the normalized variance, Shannon entropy, and PCA statistics demonstrate that the selected behavioral descriptors preserve heterogeneous information distributed across multiple complementary dimensions. Collectively, these quantitative indicators provide objective evidence supporting the structural consistency and reproducibility of the released dataset.

Table 9. Quantitative quality indicators of the published EFMS–KMeans behavioral pseudo-labeled dataset.

Quality Indicator	Measured Value	Interpretation
Missing values	0%	No incomplete observations were detected in the published dataset after preprocessing.
Constant features	0	No statistically constant variables were identified among the features evaluated for the quantitative assessment.
Observations belonging to repeated behavioral-profile groups	20,513 (2.16%)	A small proportion of records share identical released feature vectors, which is consistent with recurring operational communication patterns.
Additional repeated feature vectors	14,612 (1.54%)	These records correspond to additional occurrences after retaining one representative observation for each exact behavioral profile.
Behavioral pseudo-anomaly proportion	47,578 (5.00%)	The pseudo-anomalous class was generated using the cluster-specific 95th-percentile centroid-distance criterion.
Pseudo-anomaly proportion after exact-vector deduplication	44,868/936,948 (4.79%)	Removing additional exact repetitions reduced the pseudo-anomaly proportion by only 0.21 percentage points, indicating limited sensitivity to repeated profiles.
Mean normalized feature variance	0.023785	The feature space preserves measurable variability across the released behavioral descriptors.
Median normalized feature variance	0.000540	The difference between the mean and median indicates heterogeneous dispersion across features.
Mean Shannon entropy	2.0777 bits	The selected variables preserve heterogeneous informational content across the behavioral representation.
Median Shannon entropy	1.4057 bits	The median value confirms that informational contribution is distributed unevenly but complementarily among the variables.
Variance explained by PC1 and PC2	29.89%	The first two principal components capture less than one third of the total variance, indicating that the dataset cannot be adequately represented by only two dimensions.
Principal components required for 80% cumulative variance	12	A substantial number of components is required to preserve most of the dataset variability.
Principal components required for 90% cumulative variance	14	The information remains distributed across multiple complementary dimensions.
Principal components required for 95% cumulative variance	16	The feature space exhibits multidimensional structure and is not dominated by a small number of components.
Silhouette Score	0.193	The original EFMS–KMeans partition presents partial but measurable separation among behavioral groups.
Davies–Bouldin Index	2.23	The value indicates overlap among behavioral clusters, which is expected in heterogeneous operational network traffic.

4.3. Geometric Class Separability

To evaluate structural divergence between normal and anomalous traffic, the normalized centroid distance was computed as

$$S = \frac{\|\mu_1 - \mu_0\|}{\sigma_1 + \sigma_0}. \quad (16)$$

where μ_0 and μ_1 denote the centroids of the normal and anomalous classes, respectively, and σ_0 and σ_1 represent their corresponding dispersion measures.

Feature-level discrimination was further evaluated using the Fisher Score:

$$F_j = \frac{(\mu_j^{(1)} - \mu_j^{(0)})^2}{\sigma_j^{(1)} + \sigma_j^{(0)}}. \quad (17)$$

where $\mu_j^{(1)}$ and $\mu_j^{(0)}$ correspond to the mean values of feature j for anomalous and normal traffic, respectively, and $\sigma_j^{(1)}$ and $\sigma_j^{(0)}$ denote the associated variances.

To provide a quantitative assessment of feature-level class separability, the Fisher Score was computed for the 13 behavioral descriptors after feature standardization using `StandardScaler`. Standardization was applied to eliminate the influence of heterogeneous feature scales and to ensure that the discriminative capability of each feature was evaluated under comparable conditions. The resulting scores were then summarized using descriptive statistics to characterize the distribution of feature-level discriminative information. Table 10 reports the measured values.

Table 10. Summary statistics of the Fisher Scores computed for the standardized behavioral features.

Statistic	Measured Value
Number of evaluated features	13
Maximum Fisher Score	100,395.100976
Minimum Fisher Score	54.407852
Mean Fisher Score	21,693.560499
Median Fisher Score	10,274.148864

The standardized Fisher Scores indicate that all evaluated behavioral descriptors contribute to distinguishing the two behavioral pseudo-labeled classes, although their relative discriminative capability varies substantially. The observed scores range from 54.407852 to 100,395.100976, confirming that certain variables provide considerably greater class-discriminative information than others within the same standardized feature space.

It is important to note that Fisher Scores evaluate the discriminative capability of individual features independently rather than the global geometric structure of the multidimensional dataset. Consequently, these results should not be interpreted as direct evidence of overall cluster separation. Instead, they provide a complementary feature-level assessment alongside the Silhouette coefficient and the Davies–Bouldin Index, which evaluate the global compactness and separation of the clustering solution.

Because Fisher Scores depend on the statistical distribution of each feature and are also influenced by the sample size, their absolute magnitudes should not be interpreted using universal quality thresholds. Instead, they are employed to rank the relative contribution of the evaluated behavioral descriptors under a common standardized representation. This analysis demonstrates that the generated behavioral pseudo-labels are supported by multiple informative features while remaining consistent with the moderate cluster separation observed in the geometric validation of the operational network traffic.

Table 11 further details the individual Fisher Scores obtained for the standardized behavioral descriptors. The ranking indicates that `dstport`, `dstreputation_cat`, and packet- and byte-related traffic statistics provide the greatest discriminative capability between the behavioral pseudo-labeled classes, whereas variables such as `srcport` and `apprisk` contribute to a lesser extent. Overall, the results demonstrate that class discrimination is supported by multiple complementary behavioral features rather than by a single dominant descriptor, reinforcing the robustness of the proposed dataset for downstream anomaly detection tasks.

The quantitative Fisher Score distribution confirms that feature relevance is heterogeneous across the evaluated behavioral representation. A small subset of descriptors accounts for the largest discriminative contributions, whereas the remaining features provide lower but non-zero complementary information. This result is consistent with the multidimensional nature of operational network traffic, in which class separability is not expected to depend uniformly on all available descriptors. The use of statistical feature-relevance measures is consistent with established practices in feature selection and dataset validation for machine learning applications [30].

Table 11. Ranking of behavioral features according to the standardized Fisher Score.

Feature	Fisher Score
dstport	100,395.100976
dstreputation_cat	79,332.941988
pkt_total	21,175.599682
rcvdpkt	19,576.660866
sentpkt	16,378.911313
duration	11,478.987927
bytes_total	10,274.148864
rcvdbyte	8486.957729
hour	6739.774708
dayofweek	4731.025291
sentbyte	3150.087452
srcport	241.681836
apprisk	54.407852

4.4. Linear Structural Validation via Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was applied to quantitatively examine the global structure and intrinsic dimensionality of the published behavioral feature space under linear dimensionality reduction. Let X denote the standardized feature matrix. The PCA transformation is expressed as

$$Z = XW. \quad (18)$$

where W contains the eigenvectors associated with the covariance structure of X , and Z represents the resulting projection onto the principal-component space.

The first two principal components explained 29.89% of the total variance. Accordingly, the two-dimensional PCA projection should be interpreted as a partial representation of the global dataset geometry rather than as a complete description of the original behavioral feature space. The limited cumulative variance captured by PC1 and PC2 indicates that a substantial proportion of the information remains distributed across higher-order components.

The cumulative explained-variance analysis showed that 12 principal components were required to preserve 80% of the total variance, whereas 14 and 16 components preserved 90% and 95%, respectively. These results indicate that the released behavioral representation exhibits a multidimensional structure in which relevant information remains distributed across multiple features rather than being concentrated in a small number of linear components.

Therefore, the PCA analysis provides quantitative evidence of the intrinsic dimensionality of the dataset. Rather than demonstrating complete class separation in two dimensions, the projection indicates that normal and behavioral pseudo-anomalous observations occupy a complex and partially overlapping feature space. This interpretation is consistent with the clustering validity results reported in Section 4.6, including a Silhouette Score of 0.193 and a Davies–Bouldin Index of 2.23.

The quantitative PCA results are summarized in Table 9, complementing the statistical quality indicators by objectively characterizing the intrinsic dimensionality of the released behavioral feature space [31].

These findings support the suitability of the released dataset for multivariate anomaly detection research, demonstrating that meaningful behavioral information is distributed across multiple complementary features rather than concentrated in a reduced subset of variables.

4.5. Nonlinear Manifold Consistency via Uniform Manifold Approximation and Projection (UMAP)

To investigate the nonlinear geometric organization of the released behavioral feature space, Uniform Manifold Approximation and Projection (UMAP) was applied. The nonlinear embedding is represented as

$$Y = \text{UMAP}(X). \quad (19)$$

where X denotes the standardized behavioral feature matrix and Y corresponds to its two-dimensional nonlinear embedding.

Unlike PCA, which characterizes the distribution of global linear variance, UMAP is primarily designed to preserve local neighborhood relationships while representing nonlinear structure. To quantitatively evaluate the quality of the embedding, the Trustworthiness metric was computed for multiple neighborhood sizes. Trustworthiness measures the extent to which observations that are close in the low-dimensional embedding were also neighbors in the original high-dimensional feature space.

Table 12 reports the measured Trustworthiness values for the PCA and UMAP projections. Across all evaluated neighborhood sizes, UMAP obtained higher values than PCA. For neighborhoods of 5, 10, and 15 nearest neighbors, UMAP achieved Trustworthiness values of 0.995629, 0.993872, and 0.991276, respectively. For larger neighborhoods, the values decreased gradually to 0.980786 for $k = 30$ and 0.966282 for $k = 50$, while remaining above the corresponding PCA measurements.

Table 12. Trustworthiness comparison between the PCA and UMAP low-dimensional embeddings.

Neighborhood Size (k)	PCA Trustworthiness	UMAP Trustworthiness
5	0.960726	0.995629
10	0.957032	0.993872
15	0.954453	0.991276
30	0.948729	0.980786
50	0.944067	0.966282

The measured values show that UMAP preserves local neighborhood relationships more effectively than the two-dimensional PCA projection for every evaluated value of k . The difference between both methods was 0.034903 for $k = 5$, 0.036840 for $k = 10$, 0.036823 for $k = 15$, 0.032056 for $k = 30$, and 0.022214 for $k = 50$. The smaller difference observed at $k = 50$ indicates that the relative advantage of UMAP decreases as increasingly broader neighborhoods are evaluated. Nevertheless, UMAP maintained a Trustworthiness value greater than 0.96 throughout the analysis.

Figure 4 reveals a heterogeneous nonlinear organization composed of multiple local neighborhoods and transitional regions. Behavioral normal observations occupy most of the embedding, whereas behavioral pseudo-anomalous observations are less frequent and distributed across both peripheral and overlapping regions, indicating partial rather than complete class separation.

The visual interpretation of Figure 4 is supported by the Trustworthiness values reported in Table 12. The scores, ranging from 0.966282 to 0.995629, indicate that the two-dimensional UMAP embedding preserves most local neighborhood relationships from the original feature space, supporting its use as a reliable exploratory representation of local behavioral organization.

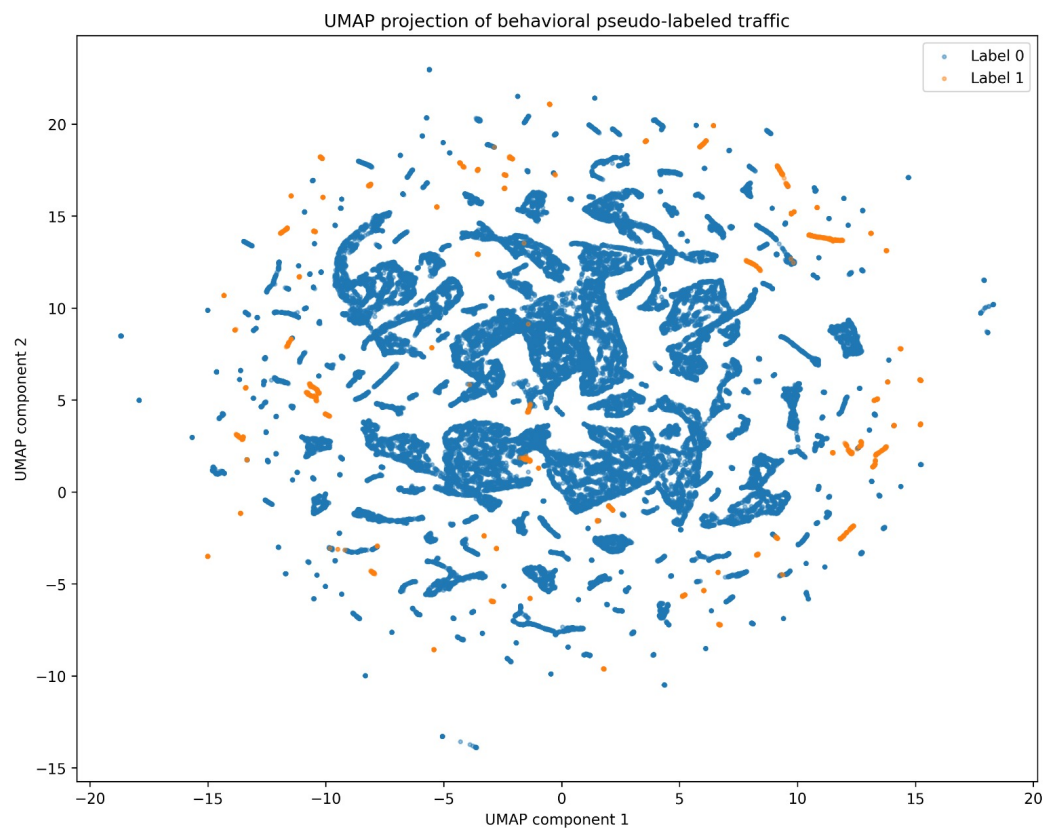


Figure 4. Two-dimensional UMAP projection of the released behavioral pseudo-labeled dataset. Blue observations represent the behavioral normal class, whereas orange observations represent the behavioral pseudo-anomalous class.

PCA and UMAP therefore provide complementary structural validation. PCA quantifies the distribution of global variance across linear components, whereas UMAP preserves nonlinear local relationships. The first two PCA components explain only 29.89% of the total variance, while the high Trustworthiness values confirm strong local neighborhood preservation. Together, these complementary analyses demonstrate that the released behavioral dataset preserves both global and local structural properties, supporting its use for exploratory analysis, clustering, and anomaly detection research under realistic encrypted traffic conditions [31].

4.6. Clustering Validity Assessment

Because the anomaly labels are generated through an unsupervised clustering process rather than manual expert annotation, their validity cannot be established through comparison with ground-truth attack labels. Instead, the consistency of the generated behavioral pseudo-labels is assessed indirectly using complementary validation strategies, including internal clustering validity indices together with the statistical, geometric, and structural analyses presented throughout Section 4. Taken together, these analyses evaluate whether the pseudo-labels represent coherent behavioral structures within the feature space rather than confirming malicious activity or demonstrating maximally separated clusters. Instead,

they provide complementary evidence that the generated behavioral pseudo-labels are internally consistent and suitable for methodological research and benchmarking.

The consistency of the EFMS–KMeans behavioral pseudo-labels was initially evaluated using standard clustering validity indices, including the Silhouette Score and the Davies–Bouldin Index. These metrics provide quantitative evidence of cluster compactness and separation, which are essential for assessing the internal consistency of the unsupervised segmentation process.

Table 13 summarizes the values obtained for the clustering validation metrics. As observed, the Silhouette score reached a value of 0.193, while the Davies–Bouldin index obtained a value of 2.23.

Table 13. Clustering validity indices obtained for the EFMS–KMeans algorithm.

Metric	Value	Interpretation
Silhouette Score	0.193	Consistent with heterogeneous operational traffic
Davies–Bouldin Index	2.23	Consistent with overlapping behavioral patterns

The obtained internal validity indices should not be interpreted as evidence of highly separated clusters. Instead, they reflect the intrinsic complexity of real operational network traffic, where normal and anomalous behavioral patterns frequently overlap due to legitimate administrative activities, heterogeneous applications, encrypted communications, and continuously evolving user behavior. Thus, moderate values of the Silhouette Score and the Davies–Bouldin Index are not unexpected in heterogeneous operational traffic environments, where legitimate and anomalous behavioral patterns naturally overlap due to encrypted communications, diverse user activities, and continuously evolving network behavior.

The use of complementary clustering validity indices contributes to the methodological rigor of the proposed behavioral pseudo-labeling strategy by providing quantitative evidence of the internal consistency of the generated pseudo-labels. These analyses complement the statistical and geometric validation presented throughout Section 4 and should therefore be interpreted as supporting evidence of label reliability rather than standalone proof of label correctness [30].

Although manually verified attack annotations are unavailable for the operational traffic analyzed in this study, the reliability of the behavioral pseudo-labels is supported through complementary validation procedures rather than a single clustering metric. Specifically, standardized Fisher Score analysis provides feature-level evidence of the discriminative capability of the behavioral descriptors, whereas PCA, UMAP, centroid-distance distributions, and internal clustering validity indices evaluate the global structural consistency of the behavioral feature space. Collectively, these complementary analyses support the reliability of the generated behavioral pseudo-labels for methodological research and benchmarking rather than as definitive attack annotations.

Therefore, label reliability in the present study should be interpreted as the consistency of behavioral pseudo-labels across multiple independent validation perspectives rather than as the outcome of a single clustering quality index.

4.7. Overall Dataset Validity

The validation results consistently support the quality of the proposed dataset. Entropy analysis confirmed the preservation of informational diversity after balancing, whereas standardized Fisher Score analysis identified multiple behavioral descriptors with significant discriminative capability. Complementary centroid-distance, PCA, UMAP, and cluster-

ing validity analyses further demonstrated the structural consistency of the generated behavioral pseudo-labels under realistic operational network conditions.

Taken together, these results indicate that the dataset preserves realistic traffic characteristics while maintaining sufficient discriminative structure for anomaly detection research in encrypted network environments [28].

5. Usage Notes

This section provides practical guidance for the reuse of the published dataset across different cybersecurity research scenarios. Particular emphasis is placed on controlled experimentation, methodological transparency, reproducibility, and the appropriate interpretation of results derived from the different dataset versions released through the proposed data-centric pipeline.

5.1. Research Applications

The published dataset supports a broad range of cybersecurity research scenarios involving encrypted network traffic analysis. Potential applications include anomaly detection, traffic characterization, behavioral profiling, clustering validation, feature selection evaluation, supervised classification, and sequential deep learning experiments.

Because the dataset preserves real operational traffic characteristics and multiple intermediate dataset versions, it can also be used to benchmark alternative preprocessing strategies, pseudo-labeling approaches, class balancing techniques, and temporal modeling frameworks. These characteristics make the dataset suitable for both methodological research and comparative evaluation studies in data-centric cybersecurity environments [32,33].

5.2. Published Resources and Recommended Usage

The repository is organized to reflect the complete dataset lifecycle and facilitate controlled reuse across different research scenarios. Multiple dataset versions are provided to support distinct experimental objectives, ranging from traffic characterization and unsupervised anomaly detection to supervised learning and sequential deep learning analysis.

Table 14 summarizes the recommended usage scenarios associated with each published dataset version.

Table 14. Recommended usage scenarios for the published dataset versions.

Dataset Version	Primary Purpose	Recommended Applications
Raw Traffic Dataset	Traffic characterization and realism analysis	Traffic profiling, prevalence estimation, statistical analysis under natural class imbalance
Processed Dataset	Preprocessing and feature engineering evaluation	Alternative feature extraction, clustering, and unsupervised labeling experiments
Pseudo-Labeled Dataset	Label validation and anomaly detection research	Clustering validation, robustness analysis, and comparison of segmentation strategies
Balanced Dataset	Supervised learning experiments	Random Forest, XGBoost, CNN, and other classification models requiring balanced classes
Sequential Training Representation	Model-oriented sequential learning	Computational training representation for CNN–GRU and other sequential deep learning experiments

The pseudo-labeled dataset is particularly suitable for evaluating clustering-based labeling strategies through internal validation metrics and cluster validity indices, which provide quantitative measures of cluster compactness and separation [34].

The availability of multiple published resources generated throughout the proposed pipeline promotes methodological transparency, reproducibility, and controlled comparison across different stages of the data engineering workflow [32,33].

5.3. Experimental Reproducibility Guidelines

To facilitate meaningful comparison across studies, researchers are encouraged to preserve the original preprocessing workflow and clearly report any modifications introduced during feature engineering, balancing, or temporal restructuring stages.

When alternative labeling approaches are investigated, experiments should explicitly distinguish between the original EFMS–KMeans pseudo-labels and newly generated annotations. Similarly, studies employing balanced data should document the balancing strategy adopted, as class distribution can substantially influence anomaly detection outcomes.

Clear reporting of dataset versions and processing choices contributes to methodological consistency and improves the interpretability of comparative evaluations across different experimental settings [16,26].

The principal contribution of the released resource lies in the combination of real operational traffic, transparent pseudo-label generation, comprehensive metadata, and a fully reproducible construction pipeline. These characteristics distinguish the dataset from many existing network traffic benchmarks that rely on synthetic traffic generation or limited methodological documentation.

5.4. Limitations, Ethical Considerations, and Appropriate Use

Although the dataset was collected from a real operational environment, its characteristics may influence the behavior and generalization capability of machine learning models, as learning performance is inherently dependent on the available data representation and experimental conditions [35]. Researchers should therefore consider these factors when comparing results across different datasets and detection approaches.

Similarly, the anomaly labels are generated through an unsupervised EFMS–KMeans pseudo-labeling strategy rather than manual expert annotation. Consequently, the dataset should primarily be considered a benchmark resource for anomaly detection research, clustering validation, and comparative experimentation rather than a source of definitive ground-truth attack labels.

The balanced dataset generated through SMOTE is intended for supervised learning experiments and should not be used for prevalence estimation or studies requiring naturally occurring traffic distributions. Researchers are encouraged to consider these characteristics when interpreting results and comparing findings with datasets generated under different operational conditions [17,34].

All records were anonymized before publication to remove direct identifiers while preserving the statistical and behavioral properties required for scientific analysis. The dataset is intended exclusively for academic research, reproducible experimentation, and methodological evaluation in encrypted traffic environments [36].

Because the dataset was derived from an operational institutional environment, privacy preservation constituted a primary design requirement throughout the publication process. Consequently, potentially identifying operational information was anonymized before public release. Although this approach protects institutional confidentiality and enables open scientific dissemination, it also limits analyses requiring the original network topology or host identities.

6. Conclusions

This work presented the construction, technical validation, and public release of a data-centric network traffic dataset derived from real operational FortiGate firewall logs collected in an environment predominantly characterized by encrypted HTTPS/TLS communications. The released resource is supported by a fully documented and reproducible pipeline comprising preprocessing, behavioral feature engineering, EFMS–KMeans behavioral pseudo-labeling, class balancing, and model-oriented sequential representations, enabling reproducible research across multiple anomaly detection scenarios.

The statistical, structural, and geometric validation results demonstrate that the dataset preserves meaningful behavioral diversity, internal consistency, and discriminative feature representations. Complementary analyses based on entropy, Fisher Scores, PCA, UMAP, and clustering quality metrics collectively support the reliability of the published behavioral pseudo-labels and the suitability of the dataset as a reproducible benchmark for methodological research in encrypted traffic analysis.

Although the dataset reflects the characteristics of a single operational academic environment and its labels correspond to behavioral pseudo-labels rather than expert-validated attacks, these design choices are consistent with the objectives of data-centric cybersecurity, where transparent data generation, reproducibility, and methodological traceability constitute primary scientific contributions.

Overall, the released dataset provides a reusable benchmark for anomaly detection research under realistic encrypted traffic conditions, facilitating reproducible experimentation, comparative evaluation, and future methodological developments. Future work will extend the dataset through additional operational environments and complementary pseudo-labeling and validation strategies to further improve its representativeness and long-term scientific value.

Author Contributions: Conceptualization, D.Q.Y.; Methodology, D.Q.Y. and D.F.I.; Software, D.Q.Y.; Validation, D.F.I. and F.J.N.; Formal analysis, D.Q.Y. and D.F.I.; Investigation, D.Q.Y.; Data curation, D.Q.Y. and F.J.N.; Writing—original draft preparation, D.Q.Y.; Writing—review and editing, D.F.I., F.J.N. and D.G.; Visualization, D.Q.Y.; Supervision, D.F.I., F.J.N. and D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out at CITIC within the framework of project PID2023-150794OB-I00, funded by the Ministry of Science, Innovation and Universities (MICIU), the State Research Agency (AEI) (Grant No. 10.13039/501100011033), and co-funded by the European Regional Development Fund (ERDF), European Union. The authors also acknowledge support from the Xunta de Galicia and the European Union (FEDER Galicia 2021–2027 Programme) through grants ED431B 2024/21, ED431B 2024/02, and CITIC ED431G 2023/01. Partial funding was also received through the ERDF EU Interreg VI-A Spain–Portugal (POCTEP) 2021–2027 Programme under the project “Quantum IBER_IA: Impulso estratégico de las capacidades en tecnologías cuánticas e inteligencia artificial en el espacio ibérico transfronterizo”.

Institutional Review Board Statement: Institutional Review Board approval was not required, as this study used only anonymized operational network metadata and did not involve human participants or personally identifiable information. The operational firewall logs were provided for research purposes with the authorization of the institution’s Director of Information Technology.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is publicly available through the Mendeley Data repository at <http://doi.org/10.17632/g5dtm7349f>. The released files contain only anonymized flow-level metadata derived from operational firewall logs. All identifiers capable of revealing operational infrastructure or individual users were anonymized prior to publication. No packet payloads or personally identifiable information are included in the repository.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
DL	Deep Learning
EFMS	Enhanced Feature Mapping and Selection
GRU	Gated Recurrent Unit
HTTPS	Hypertext Transfer Protocol Secure
ICMP	Internet Control Message Protocol
IDS	Intrusion Detection System
IoT	Internet of Things
ML	Machine Learning
PCA	Principal Component Analysis
QUIC	Quick UDP Internet Connections
SMOTE	Synthetic Minority Over-sampling Technique
TCP	Transmission Control Protocol
TLS	Transport Layer Security
UDP	User Datagram Protocol
UMAP	Uniform Manifold Approximation and Projection

References

1. Ferriyan, A.; Thamrin, A.H.; Takeda, K.; Murai, J. Generating Network Intrusion Detection Dataset Based on Real and Encrypted Synthetic Attack Traffic. *Appl. Sci.* **2021**, *11*, 7868. [\[CrossRef\]](#)
2. Alshaibi, A.; Al-Ani, M.; Al-Azzawi, A.; Konev, A.; Shelupanov, A. The Comparison of Cybersecurity Datasets. *Data* **2022**, *7*, 22. [\[CrossRef\]](#)
3. Choi, D.-S.; Kim, T.; Kang, B.; Im, E.G. Image-Based Malicious Network Traffic Detection Framework: Data-Centric Approach. *Appl. Sci.* **2025**, *15*, 6546. [\[CrossRef\]](#)
4. Ayodele, B.; Buttigieg, V. The VNF Cybersecurity Dataset for Research (VNFCYBERDATA). *Data* **2024**, *9*, 132. [\[CrossRef\]](#)
5. Cabezas-Quinto, J.J.; Vidal-Silva, C.; Serrano-Malebrán, J.; Márquez, N. A Dataset and Experimental Evaluation of a Parallel Conflict Detection Solution for Model-Based Diagnosis. *Data* **2025**, *10*, 139. [\[CrossRef\]](#)
6. Sufi, F. A New Time Series Dataset for Cyber-Threat Correlation, Regression and Neural-Network-Based Forecasting. *Information* **2024**, *15*, 199. [\[CrossRef\]](#)
7. Zhou, Y.; Shi, H.; Zhao, Y.; Ding, W.; Han, J.; Sun, H.; Zhang, X.; Tang, C.; Zhang, W. Identification of Encrypted and Malicious Network Traffic Based on One-Dimensional Convolutional Neural Network. *J. Cloud Comput.* **2023**, *12*, 53. [\[CrossRef\]](#)
8. Liu, X.; Liu, J. Malicious Traffic Detection Combined Deep Neural Network with Hierarchical Attention Mechanism. *Sci. Rep.* **2021**, *11*, 12363. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Ferrag, M.A.; Maglaras, L.; Moschoyiannis, S.; Janicke, H. Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study. *J. Inf. Secur. Appl.* **2020**, *50*, 102419. [\[CrossRef\]](#)
10. Papernot, N.; McDaniel, P.; Sinha, A.; Wellman, M. SoK: Towards the Science of Security and Privacy in Machine Learning. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*; IEEE: London, UK, 2018; pp. 1–19.
11. Paullada, A.; Raji, I.D.; Bender, E.M.; Denton, E.; Hanna, A. Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns* **2021**, *2*, 100336. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Quirumbay Yagual, D.; Fernández Iglesias, D.; Nóvoa, F.J. A Hybrid Deep Learning-Based Architecture for Network Traffic Anomaly Detection via EFMS-Enhanced KMeans Clustering and CNN-GRU Models. *Appl. Sci.* **2025**, *15*, 10889. [\[CrossRef\]](#)
13. Papadogiannaki, E.; Ioannidis, S. A Survey on Encrypted Network Traffic Analysis: Applications, Techniques, and Countermeasures. *ACM Comput. Surv.* **2021**, *54*, 123. [\[CrossRef\]](#)
14. Maciá-Fernández, G.; Camacho, J.; Magán-Carrión, R.; García-Teodoro, P.; Therón, R. UGR'16: A New Dataset for the Evaluation of Cyclostationarity-Based Network IDSs. *Comput. Secur.* **2018**, *73*, 411–424. [\[CrossRef\]](#)
15. Zha, D.; Bhat, Z.P.; Lai, K.H.; Yang, F.; Jiang, Z.; Zhong, S.; Hu, X. Data-Centric Artificial Intelligence: A Survey. *ACM Comput. Surv.* **2025**, *57*, 129. [\[CrossRef\]](#)
16. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in ML-Based Science. *Patterns* **2023**, *4*, 100804. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; Aroyo, L. Everyone Wants to Do the Model Work, Not the Data Work: Data Cascades in High-Stakes AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery (ACM): New York, NY, USA, 2021; Article 39, pp. 1–15. [\[CrossRef\]](#)

18. Ashmore, R.; Calinescu, R.; Paterson, C. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *ACM Comput. Surv.* **2021**, *54*, 1–39. [\[CrossRef\]](#)
19. Nguyen, T.T.T.; Armitage, G. A Survey of Techniques for Internet Traffic Classification Using Machine Learning. *IEEE Commun. Surv. Tutor.* **2008**, *10*, 56–76. [\[CrossRef\]](#)
20. Aceto, G.; Ciuonzo, D.; Montieri, A.; Pescapé, A. Toward Effective Mobile Encrypted Traffic Classification through Deep Learning. *Neurocomputing* **2020**, *409*, 306–315. [\[CrossRef\]](#)
21. Ring, M.; Wunderlich, S.; Scheuring, D.; Landes, D.; Hotho, A. A Survey of Network-Based Intrusion Detection Data Sets. *Comput. Secur.* **2019**, *86*, 147–167. [\[CrossRef\]](#)
22. Douzas, G.; Bacao, F.; Last, F. Improving Imbalanced Learning through a Heuristic Oversampling Method Based on K-Means and SMOTE. *Inf. Sci.* **2018**, *465*, 1–20. [\[CrossRef\]](#)
23. Choi, K.; Yi, J.; Park, C.; Yoon, S. Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines. *IEEE Access* **2021**, *9*, 120043–120055. [\[CrossRef\]](#)
24. Patel, H.; Guttula, S.; Gupta, N.; Hans, S.; Mittal, R.S.; Lokesh, N. A Data-centric AI Framework for Automating Exploratory Data Analysis and Data Quality Tasks. *ACM J. Data Inf. Qual.* **2023**, *15*, 44. [\[CrossRef\]](#)
25. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Daumé, H.; Crawford, K. Datasheets for Datasets. *Commun. ACM* **2021**, *64*, 86–92. [\[CrossRef\]](#)
26. Pineau, J.; Vincent-Lamarre, P.; Sinha, K.; Larivière, V.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E.; Larochelle, H. Improving Reproducibility in Machine Learning Research: A Report from the NeurIPS 2019 Reproducibility Program. *J. Mach. Learn. Res.* **2021**, *22*, 1–20.
27. Lashkari, A.H.; Draper-Gil, G.; Mamun, M.S.I.; Ghorbani, A.A. Characterization of Encrypted and VPN Network Traffic Using Time-Related Features. *J. Inf. Secur. Appl.* **2021**, *58*, 102425.
28. Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding Deep Learning (Still) Requires Rethinking Generalization. *Commun. ACM* **2021**, *64*, 107–115. [\[CrossRef\]](#)
29. Aceto, G.; Ciuonzo, D.; Montieri, A.; Pescapé, A. MIMETIC: Mobile Encrypted Traffic Classification Using Multimodal Deep Learning. *Comput. Netw.* **2019**, *165*, 106944. [\[CrossRef\]](#)
30. Moslemi, A. A Tutorial-Based Survey on Feature Selection: Recent Advancements on Feature Selection. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107136. [\[CrossRef\]](#)
31. Espadoto, M.; Martins, R.M.; Kerren, A.; Hirata, N.S.T.; Telea, A.C. Toward a Quantitative Survey of Dimension Reduction Techniques. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 2153–2173. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Mazumder, M.; Banbury, C.; Yao, X.; Karlaš, B.; Ciro, A.; Damos, G.; Damos, S.; He, L.; Kerkouche, R.; Park, Y.; et al. DataPerf: Benchmarks for Data-Centric AI Development. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 74773–74804.
33. Jakubik, J.; Vössing, M.; Kühl, N.; Walk, J.; Satzger, G. Data-Centric Artificial Intelligence. *Bus. Inf. Syst. Eng.* **2024**, *66*, 507–515. [\[CrossRef\]](#)
34. Ikotun, A.M.; Habyarimana, F.; Ezugwu, A.E. Cluster Validity Indices for Automatic Clustering: A Comprehensive Review. *Heliyon* **2025**, *11*, e41953. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Domingos, P. A Few Useful Things to Know about Machine Learning. *Commun. ACM* **2012**, *55*, 78–87. [\[CrossRef\]](#)
36. Whang, S.E.; Roh, Y.; Song, H.; Lee, J.-G. Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. *VLDB J.* **2023**, *32*, 791–813. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.