Polaris Global Journal of Scholarly Research and Trends

Robust Statistical Anomaly Detection for Interpretable Intrusion Monitoring in Complex Network Systems

DOI: 10.58429/pgjsrt.v5n4a240

Ibrahim Mallam Fali[1*], Emeka Nwinyinya[2]
Accounting, Faculty of Administration and Management Sciences, University of Calabar, Calabar, Nigeria
Educational Management and Foundational Studies, Faculty of Education, Alex Ekwueme Federal University, Ndufu Alike, Ebonyi State, Abakaliki, Nigeria
Corresponding Author: faliibrahim7@gmail.com

ARTICLEINFO **VOLUME 5 |NO. 4 |2026 ISSUE**

ABSTRACT

Anomaly detection is essential for identifying abnormal observations that may signal operational errors, system failures, fraud, network misuse, or security threats in complex data systems. This study develops and evaluates a robust statistical framework for interpretable anomaly detection using the Canadian Institute for Cybersecurity Intrusion Detection System 2017 (CICIDS2017) network intrusion dataset as an empirical case. The framework combines descriptive profiling, data-quality auditing, benign-baseline calibration, robust location and scale estimation, interquartile range (IQR) boundaries, median absolute deviation (MAD) scoring, empirical probability-tail thresholds, and feature-count aggregation. The empirical dataset contains 2,830,743 observations distributed across eight flow-based files, including 2,273,097 benign records and 557,646 attack or anomaly records. The Monday benign-only traffic was used to calibrate the normal statistical baseline, while 2,300,825 observations from the remaining days were retained for holdout evaluation. Sixteen interpretable flow variables were selected for the statistical monitoring experiment after cleaning missing values, infinite values, and inconsistent labels. The most conservative probability-tail rule reached 84.80% accuracy, 87.30% precision, 43.63% recall, an F1-score of 0.5818, and a false-alarm rate (FAR) of 2.03%. The findings show that robust threshold monitoring is particularly useful when the main operational objective is high-confidence early warning with low false alarms rather than maximum attack recall. The study contributes a transparent, reproducible, and statistically grounded anomaly-detection procedure that can support cybersecurity monitoring and can be adapted to other risk-sensitive complex data systems.

KEYWORDS

anomaly detection; robust statistics; network intrusion; outlier detection; threshold monitoring

I. INTRODUCTION

Complex data systems generate large volumes of operational information from network devices, industrial equipment, financial transactions, health records, and digital platforms. Within these environments, most observations reflect routine activity, while a smaller portion may represent abnormal behavior. These abnormal observations are important because they may signal early system failure, unexpected process drift, fraudulent activity, unauthorized access, denial-of-service behavior, scanning activity, or other operational risks. Anomaly detection therefore has a direct practical role in early-warning systems, risk prevention, and security monitoring.

Network systems provide a particularly suitable domain for robust anomaly detection because traffic flows are large, noisy, heterogeneous, and frequently imbalanced. A network-flow dataset may contain millions of benign records but only a smaller portion of attack records, and some attack classes may be rare. In addition, flow variables such as bytes per second, packets per second, packet length, idle time, and active time often show extreme skewness and heavy-tailed behavior. These characteristics reduce the reliability of simple mean-standard deviation thresholds and make anomaly decisions sensitive to outliers, missing values, infinite values, and distributional irregularities.

Many computational intrusion-detection models aim to maximize predictive accuracy, yet high performance alone is not always sufficient for operational monitoring. Security analysts and decision-makers often need clear threshold logic, reproducible rules, and interpretable evidence explaining why a given observation is treated as abnormal. If a detection model is accurate but opaque, it may be difficult to justify alerts, calibrate warning levels, or understand false alarms. Statistical monitoring remains valuable because it can translate abnormality into transparent rules based on robust baselines and measurable deviations from normal behavior.

The central problem addressed in this study is that conventional statistical thresholds may be unstable in complex network data, while many high-performing predictive models may reduce interpretability. A practical anomaly-detection framework should therefore balance detection performance with robustness and transparency. Robust statistics offer this balance by using median-based location, resistant scale measures, quartile boundaries, median absolute deviation, and empirical tail probabilities instead of relying exclusively on the classical mean and standard deviation. Such measures are less distorted by extreme values and can be communicated as direct monitoring rules.

The aim of this study is to develop and empirically evaluate a robust statistical framework for interpretable anomaly detection in complex network systems. The empirical case uses CICIDS2017-style network intrusion data, where benign traffic is treated as normal behavior, and all attack categories are mapped into a binary anomaly class. The Monday benign-only file is used as the baseline for calibrating normal traffic behavior, and the remaining days are used as the holdout evaluation set. This design simulates a realistic monitoring situation in which normal historical behavior is used to set thresholds and later traffic is screened for abnormality.

The manuscript is organized as follows. Section 2 reviews the related literature and summarizes the main empirical and methodological studies. Section 3 identifies the research gap, and Section 4 explains the study contribution. Section 5 presents the proposed statistical framework. Section 6 describes the methodology, dataset, cleaning procedure, variables, equations, and evaluation metrics. Section 7 presents the empirical results. Section 8 discusses the findings regarding robustness, interpretability, operational monitoring, theoretical and practical implications, limitations, and future research. Section 9 concludes the manuscript.

II. LITERATURE REVIEW

Anomaly detection is widely treated in the literature as a core problem in outlier analysis, with the objective of identifying observations that meaningfully deviate from the dominant behavior of the data. This foundation is important for network-flow analysis because abnormal records may represent rare

events, data errors, or meaningful security signals rather than simple noise to be discarded [1]. In cybersecurity contexts, anomaly detection has been approached through statistical, classification-based, clustering-based, and information-theoretic methods, confirming that statistical monitoring remains an important family within a broader methodological landscape [2].

A related strand of research emphasizes that anomaly detection should not only identify unusual observations but also support interpretation and explanation. Work on graph anomaly detection shows that abnormality becomes more useful when the detected case can be described and understood, even when the data structure differs from tabular network-flow data [3]. Reviews of network anomaly-detection systems also show that false alarms, scalability, operational usability, and system transparency remain persistent challenges, which supports the need for detection frameworks that combine performance reporting with clear rule logic [4].

Density-based and data-driven cybersecurity approaches have contributed substantially to anomaly-detection research. Local density-based methods are useful because they identify observations located in sparse regions of the data space, but their scoring logic can be less direct for simple threshold governance [5]. Broader reviews of data-driven intrusion detection further show the increasing use of computational and machine-learning models in cybersecurity, while also highlighting the continuing need for reproducible and interpretable alternatives to complex predictive systems [6].

Evaluation design is another central issue in anomaly-detection research. Studies on unsupervised outlier-detection benchmarking show that model quality cannot be judged by a single metric; accuracy, precision, recall, F1-score, and false-alarm rate must be interpreted together, especially when abnormal cases are rare [7]. Foundational anomaly-detection surveys also distinguish among point, contextual, and collective anomalies, which positions the present study primarily within point anomaly detection, as each network-flow record is evaluated against robust statistical thresholds [8].

Comparative algorithmic studies show that anomaly-detection performance varies substantially across domains, datasets, and data structures. This supports the use of threshold sensitivity analysis because no single anomaly rule can be assumed to perform best under all empirical conditions [9]. Research on unsupervised multivariate anomaly detection similarly demonstrates that algorithmic effectiveness depends strongly on data characteristics, reinforcing the importance of selecting interpretable variables and carefully reporting performance [10].

Temporal structure and class imbalance are especially important in intrusion-detection settings. Research on temporal and sequence-related outlier detection is relevant because network-flow records often contain duration, rate, active period, idle period, and timing-related variables that shape the interpretation of abnormal traffic [11]. At the same time, studies on imbalanced classification show that benign traffic usually dominates attack traffic, making accuracy alone insufficient and requiring additional attention to recall, precision, and false-alarm behavior [12].

Robust statistics provide a strong methodological foundation for interpretable anomaly detection. Robust location and scatter estimation show how resistant baselines can reduce the influence of extreme or contaminated observations when modeling normal behavior [13]. In network intrusion detection, comparative studies confirm that anomaly detection is a natural cybersecurity monitoring approach, although older benchmark work also shows the importance of cautious interpretation and transparent evaluation design [14].

Algorithmic anomaly-detection methods provide useful comparison points for statistical monitoring. Isolation-based anomaly detection identifies abnormal observations by their relative ease of separation from the rest of the data, which offers an influential model-based perspective [15]. However, benchmark studies using CICIDS2017 show that intrusion-detection performance depends

heavily on preprocessing, feature selection, and evaluation protocol, supporting the present study's focus on documented cleaning, benign-baseline calibration, and holdout testing [16].

The quality of empirical datasets is also central to anomaly-detection research. Modern intrusion-detection datasets reinforce the need for realistic benchmark data that include structured flow records and labeled attack categories [17]. At the same time, reviews of deep anomaly-detection methods show the growing role of complex computational models, but such models may reduce threshold-level interpretability, which strengthens the rationale for a transparent statistical monitoring framework [18].

Novelty-detection research is directly relevant to the present framework because it models normal behavior first and treats later deviations as potential abnormalities [19]. This logic supports the use of benign traffic to calibrate a normal baseline before evaluating later network-flow observations. Explainable decision-making research also emphasizes that automated systems are more useful when users can understand how outputs are produced, which aligns with the present study's traceable threshold-violation logic [20].

Dataset suitability and data cleaning are recurring concerns in cybersecurity benchmark research. Reviews of network-based intrusion-detection datasets stress that empirical studies must clearly describe dataset structure, attack labels, and evaluation conditions before making methodological claims [21]. Specific analyses of CICIDS2017 also identify quality and flow-extraction issues, showing that missing values, infinite values, and label inconsistencies should be addressed before anomaly thresholds are constructed [22].

Robust anomaly detection is further supported by statistical work showing that resistant measures, such as the median and median absolute deviation, provide stable baselines under contamination [23]. Broader methodological reviews also show that anomaly detection includes both shallow and deep model families, allowing interpretable statistical approaches to remain relevant even within a field increasingly influenced by complex machine-learning models [24].

Threshold design and benchmark justification are closely connected. Time-series anomaly-detection studies show that sensitivity analysis is necessary because monitoring rules must be aligned with operational objectives and acceptable false-alarm levels [25]. The CICIDS2017 dataset provides an important benchmark context because it includes generated benign and attack traffic, making it suitable for evaluating intrusion-monitoring methods under labeled empirical conditions [26].

Recent intrusion-detection literature also warns that benchmark performance should not be interpreted as immediate deployment readiness. Studies on real-world applicability show that models developed under controlled benchmark conditions may face transferability challenges in operational environments [27]. Earlier benchmark critiques also identify redundancy and evaluation limitations in intrusion-detection datasets, which supports the present study's emphasis on dataset auditing, class distribution analysis, and cautious interpretation of results [28].

Finally, high-dimensional numerical outlier-detection research shows that anomaly detection becomes more difficult when many correlated or noisy variables are available [29]. This issue is highly relevant to network-flow data because such datasets contain many numerical features.

Overall, the reviewed literature shows a clear need for interpretable statistical frameworks that combine robust baseline calibration, transparent threshold rules, threshold sensitivity analysis, and practical performance evaluation in complex network-flow environments.

Table 1. Literature review comparison table.

No.	Author(s) and year	Focus	Limitation	Relevance
[1]	Aggarwal, C. C. (2017)	General outlier analysis; computational and statistical detection families.	Broad reference work; not specific to network-flow monitoring.	Supports the general outlier logic and terminology of the present framework.
[2]	Ahmed, M., Mahmood, A. N., & Hu, J. (2016)	Review of classification, statistical, information-theoretic, and clustering methods for network anomalies.	Survey work without a new empirical statistical framework.	Positions statistical monitoring among major network anomaly-detection families.
[3]	Akoglu, L., Tong, H., & Koutra, D (2015)	Survey of anomaly detection and explanation in graph-structured data.	Graph settings differ from flow-tabular intrusion data.	Strengthens the interpretability and explanation rationale.
[4]	Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2014)	Structured survey of network anomaly-detection systems and tools.	Mainly review-based and broad.	Identifies operational challenges such as false alarms and scalability.
[5]	Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000)	Introduced local density-based outlier scoring.	Density methods are less transparent for simple threshold governance.	Provides a contrast with threshold-based statistical rules.
[6]	Buczak, A. L., & Guven, E (2016)	Comprehensive review of data-driven cyber intrusion detection.	Emphasizes many computational models rather than robust statistical monitoring.	Highlights the need for transparent, reproducible alternatives.
[7]	Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenkova, B., Schubert, E., Assent, I., & Houle, M. E (2016)	Examined evaluation measures and benchmark issues in unsupervised outlier detection.	General benchmark focus rather than network-only study.	Supports the use of multiple metrics rather than accuracy alone.
[8]	Chandola, V., Banerjee, A., & Kumar, V. (2009)	Foundational survey of anomaly detection types and applications.	Pre-dates newer intrusion datasets and modern feature sets.	Provides theoretical grounding for point anomaly detection.
[9]	Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018)	Compared outlier algorithms across public and industrial datasets.	Focuses on algorithms more than interpretable threshold logic.	Supports empirical comparison and threshold sensitivity analysis.

No.	Author(s) and year	Focus	Limitation	Relevance
[10]	Goldstein, M., & Uchida, S. (2016)	Compared unsupervised anomaly-detection algorithms for multivariate data.	Not specialized to cybersecurity flows.	Supports multivariate anomaly scoring and fair evaluation.
[11]	Gupta, M., Gao, J., Aggarwal, C. C., & Han, J (2014)	Reviewed temporal and sequence-related outlier detection.	Network-flow records are treated as tabular flows in this study.	Shows why timing variables are important anomaly indicators.
[12]	He, H., & Garcia, E. A (2009)	Reviewed problems caused by imbalanced class distributions.	Focuses on learning models rather than statistical thresholds.	Supports the use of precision, recall, and false-alarm metrics.
[13]	Hubert, M., Rousseeuw, P. J., & Verdonck, T. (2012)	Developed robust location and scatter estimation.	Multivariate covariance estimation may be heavy for large network datasets.	Supports robust baseline estimation under contamination.
[14]	Lazarevic, A., Ertoz, L., Ozgur, A., Kumar, V., & Srivastava, J. (2003)	Compared anomaly-detection schemes in intrusion detection.	Older datasets and methods limit direct transferability.	Establishes the network intrusion relevance of anomaly detection.
[15]	Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008)	Introduced isolation-based anomaly detection.	Algorithmic isolation may be less transparent than direct thresholds.	Used as a benchmark family motivating interpretable statistical alternatives.
[16]	Maseer, Z. K., Yusof, R., Bahaman, N., Mostafa, S. A., & Foozy, C. F. M (2021)	Benchmarked several intrusion-detection models on CICIDS2017.	Model performance depends strongly on evaluation protocol and preprocessing.	Provides direct dataset-related comparison background.
[17]	Moustafa, N., & Slay, J (2015)	Introduced another modern network intrusion-detection dataset.	Different feature construction and attack taxonomy from CICIDS2017.	Supports the importance of realistic benchmark datasets.
[18]	Pang, G., Shen, C., Cao, L., & van den Hengel, A. (2021)	Reviewed deep anomaly-detection methods.	Complex models may reduce threshold-level interpretability.	Clarifies why this study emphasizes transparent statistical monitoring.
[19]	Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L (2014)	Reviewed novelty detection where normal data are modeled first.	Not specific to intrusion-detection feature engineering.	Supports the Monday benign-baseline calibration design.
[20]	Ribeiro, M. T., Singh, S., & Guestrin, C (2016)	Developed local explanation logic for predictions.	Explains classifiers rather than threshold rules.	Supports the value of explainable decision logic.

No.	Author(s) and year	Focus	Limitation	Relevance
[21]	Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019)	Reviewed network intrusion datasets and suitability criteria.	Dataset survey without new detection framework.	Supports the empirical choice of a labeled flow-based dataset.
[22]	Rosay, A., Cheval, E., Carlier, F., & Leroux, P (2022)	Analyzed CICIDS2017 quality and flow extraction issues.	Focuses on dataset diagnostics rather than robust statistical scoring.	Informs cleaning decisions for labels, missing values, and flow variables.
[23]	Rousseeuw, P. J., & Hubert, M. (2018)	Explained robust statistical anomaly detection principles.	Overview paper; empirical domain adaptation remains needed.	Core theoretical source for median, MAD, and robust deviation scoring.
[24]	Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., & Muller, K. R (2021)	Unified shallow and deep anomaly-detection perspectives.	Does not prioritize simple operational thresholds.	Helps locate robust statistics within broader anomaly-detection methods.
[25]	Schmidl, S., Wenig, P., & Papenbrock, T (2022)	Evaluated time-series anomaly-detection methods.	Current study treats flow records as cross-sectional holdout observations.	Supports careful evaluation and sensitivity analysis.
[26]	Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A (2018)	Introduced CICIDS2017 and described benign and attack traffic generation.	Dataset still requires cleaning and cautious validation.	Primary empirical dataset source for the present study.
[27]	Sommer, R., & Paxson, V (2010)	Discussed deployment challenges for intrusion-detection models.	Focused on modeling limitations rather than robust threshold statistics.	Supports the need for operationally cautious and interpretable detection.
[28]	Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A (2009)	Identified benchmark dataset redundancy and evaluation problems.	Analyzes older KDD data rather than CICIDS2017.	Supports dataset audit before empirical modeling.
[29]	Zimek, A., Schubert, E., & Kriegel, H. P (2012)	Reviewed high-dimensional outlier-detection challenges.	Broad survey rather than domain-specific evaluation.	Supports feature selection and robustness concerns in flow data.

Table 1 summarizes the literature base and shows how the present study moves from general anomaly-detection theory toward a robust statistical framework for interpretable network-flow monitoring.

III. RESEARCH GAP

The reviewed literature shows that anomaly detection is well established across outlier analysis, network security, novelty detection, and benchmark evaluation. However, three gaps remain

important for this study. First, many intrusion-detection studies emphasize predictive performance without giving sufficient attention to transparent statistical decision rules. Second, robust statistical methods are often discussed conceptually but are less frequently organized into a full empirical framework using large labeled network-flow data. Third, CICIDS2017-based studies commonly focus on classification models, while fewer studies use benign-baseline calibration to test interpretable threshold monitoring on later traffic.

This study addresses these gaps by proposing a robust statistical anomaly-detection framework that combines data audit, benign-baseline calibration, resistant location and scale measures, robust z-score analysis, IQR boundaries, empirical probability-tail thresholds, and feature-count aggregation. The framework does not claim to replace high-performance predictive classifiers; rather, it provides a transparent monitoring layer suitable for early warning, explainable alerting, and risk-sensitive environments where low false-alarm rates and clear threshold logic are important.

IV.RESEARCH QUESTIONS

The study is guided by four research questions. First, how can robust statistical methods be organized into an interpretable anomaly-detection framework for complex network data? Second, how effectively can robust threshold rules distinguish benign network traffic from attack traffic? Third, which threshold family provides the best balance between precision, recall, and false-alarm control? Fourth, how can the resulting framework support practical early-warning systems where transparent decision rules are required?

V. STUDY CONTRIBUTION

The contribution of this study is fourfold. First, it develops a robust and interpretable statistical framework for anomaly detection in complex network systems. Second, it calibrates normal traffic behavior using a benign-only baseline and applies resistant threshold rules to later holdout traffic. Third, it evaluates the framework using a large labeled CICIDS2017-style dataset with 2,830,743 records, multiple attack categories, and standard performance metrics. Fourth, it demonstrates the practical trade-off between high-confidence anomaly alerts and attack recall, showing how probability-tail monitoring can provide a low-false-alarm early-warning layer while remaining transparent to analysts.

VI. PROPOSED ROBUST STATISTICAL FRAMEWORK

The proposed framework follows a sequential monitoring logic. Data are first audited and cleaned. Benign observations are then used to construct a normal statistical baseline. Robust measures are calculated from the baseline, including median, IQR, MAD, and empirical tail quantiles. Each new observation is compared against these robust boundaries. A feature-count rule aggregates abnormal feature signals so that an observation is classified as anomalous when one or more selected variables violate the chosen threshold condition. The final detection output is evaluated against the known benign or attack label.

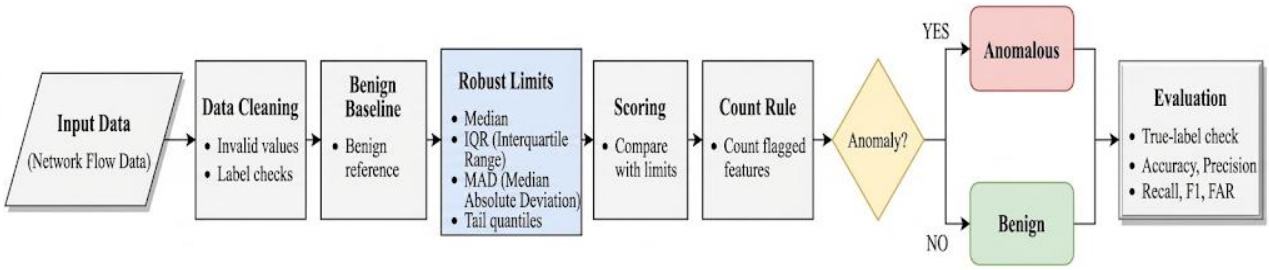

FIGURE 1. Proposed robust statistical anomaly-detection workflow.

FIGURE 1. shows the six-stage procedure used to move from raw network-flow records to interpretable anomaly classification and performance evaluation.

Table 2. Statistical threshold methods used in the framework.

Method	Rule	Interpretation
Robust z-score	$0.6745\|x - \text{median}\| / \text{MAD} > c$	Flags deviations from a robust benign baseline.
IQR rule	$x < Q1 - k(IQR)$ or $x > Q3 + k(IQR)$	Flags values outside robust quartile-based boundaries.
Empirical probability-tail rule	x below alpha or above 1-alpha benign quantiles	Flags rare values under the benign baseline distribution.
Feature-count aggregation	An observation is anomalous when at least m features violate thresholds	Controls noise by requiring several abnormal feature signals.

Table 2. shows that each threshold method has a direct interpretation, which is central to the transparency of the proposed framework.

VII. METHODOLOGY

1. RESEARCH DESIGN

The study uses a quantitative empirical design based on robust statistical anomaly detection. The design is not a purely theoretical framework; it applies statistical thresholds to a labeled network intrusion dataset and evaluates the resulting anomaly classifications. The analysis uses benign historical traffic for calibration and labeled holdout traffic for evaluation. This design is appropriate for early-warning settings because thresholds are derived from normal behavior and then applied to later observations.

2. DATASET AND EMPIRICAL SETTING

The empirical dataset is a CICIDS2017-style network intrusion dataset consisting of eight CSV files. The complete dataset contains 2,830,743 flow observations, 78 original numeric flow features, and one label column. The label column includes BENIGN traffic and multiple attack categories, including DDoS, PortScan, DoS Hulk, brute-force attacks, bot traffic, infiltration, SQL injection, and Heartbleed. For the binary anomaly-detection task, BENIGN records were coded as 0 and all attack records were coded as 1.

Table 3. Dataset file composition.

File	Rows	BENIGN	Attack/anomaly	Main labels
Friday-WorkingHours-Afternoon-DDos	225,745	97,718	128,027	DDoS (128027), BENIGN (97718)
Friday-WorkingHours-Afternoon-PortScan	286,467	127,537	158,930	PortScan (158930), BENIGN (127537)

File	Rows	BENIGN	Attack/anomaly	Main labels
Friday-WorkingHours-Morning	191,033	189,067	1,966	BENIGN (189067), Bot (1966)
Monday-WorkingHours	529,918	529,918	0	BENIGN (529918)
Thursday-WorkingHours-Afternoon-Infilteration	288,602	288,566	36	BENIGN (288566), Infiltration (36)
Thursday-WorkingHours-Morning-WebAttacks	170,366	168,186	2,180	BENIGN (168186), Web Attack Brute Force (1507), Web Attack XSS (652), Web Attack SQL Injection (21)
Tuesday-WorkingHours	445,909	432,074	13,835	BENIGN (432074), FTP-Patator (7938), SSH-Patator (5897)
Wednesday-workingHours	692,703	440,031	252,672	BENIGN (440031), DoS Hulk (231073), DoS GoldenEye (10293), DoS slowloris (5796)

Table 3 shows that Monday contains only benign traffic and was therefore used as the normal calibration baseline, while the remaining days were used for holdout evaluation.

3. DATA CLEANING AND PREPROCESSING

The preprocessing procedure included cleaning column names, standardizing label text, replacing infinite values, imputing missing numerical values with baseline medians, and mapping labels into binary classes. The dataset contained 1,358 missing values in Flow Bytes/s and 4,376 infinite values mainly in Flow Bytes/s and Flow Packets/s. These values were corrected before threshold construction. Sixteen interpretable flow variables were selected for the monitoring experiment because they represent traffic duration, packet counts, byte volumes, rates, packet-length behavior, subflow volumes, and activity/idle timing.

Table 4. Data quality audit and analytical design.

Item	Value
CSV files	8
Total observations	2830743
Original numeric flow features available	78
Numeric flow features used in statistical framework	16
Label column	1
BENIGN observations	2273097
Attack/anomaly observations	557646
Calibration observations (Monday benign baseline)	529918
Evaluation observations (Tuesday-Friday holdout)	2300825
Evaluation BENIGN observations	1743179
Evaluation attack/anomaly observations	557646
Valid robust-MAD features	14
Missing values before imputation	1,358 (Flow Bytes/s)
Infinite values before replacement	4,376 (Flow Bytes/s and Flow Packets/s)

Table 4. documents the main data-quality and evaluation-design decisions before robust statistical modeling.

4. STATISTICAL EQUATIONS

The classical z-score was included as a reference point because it is widely used in threshold monitoring, although it is sensitive to extreme values.

The robust z-score replaces the mean and standard deviation with the median and MAD. The constant 0.6745 aligns the median absolute deviation with the standard deviation under normality.

$$rz_i = \frac{0.6745\,|x_i - \text{median}(x)|}{\text{MAD}(x)} \qquad (2)$$

The interquartile range rule defines lower and upper boundaries from quartiles rather than from mean-based assumptions.

$$\text{IQR} = Q_3 - Q_1, \quad L = Q_1 - k\text{IQR}, \quad U = Q_3 + k\text{IQR} \qquad (3)$$

The empirical probability-tail rule flags values located outside rare benign-baseline quantiles.

$$T_i = \begin{cases} 1, & x_i < Q_\alpha \text{ or } x_i > Q_{1-\alpha} \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

A feature-count aggregation rule was then used to convert individual feature violations into an observation-level anomaly decision.

$$A(x_i) = \begin{cases} 1, & \sum_{j=1}^{p} I\left(f_{ij} \text{ violates threshold}\right) \geq m \\ 0, & \text{otherwise} \end{cases} \qquad (5)$$

Performance was evaluated with standard binary metrics derived from true positives, false positives, true negatives, and false negatives.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F_1 = \frac{2(\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}} \qquad (6)$$

5. EVALUATION STRATEGY

The Monday file was used as the benign calibration set because it contains 529,918 BENIGN observations and no attack records. The remaining seven files formed the holdout evaluation set with 2,300,825 observations, including 1,743,179 benign observations and 557,646 attack observations. This separation reduces calibration contamination and reflects an operational setting in which a normal baseline is used to screen later network activity.

VIII. RESULTS

1. DATASET STRUCTURE AND CLASS DISTRIBUTION

The dataset is strongly imbalanced but still contains a substantial number of attack observations. BENIGN traffic represents 80.30% of the full dataset, while attack/anomaly traffic represents 19.70%. Some attack categories are common, such as DoS Hulk, PortScan, and DDoS, whereas other attacks are rare, including Heartbleed, SQL injection, and infiltration. This imbalance justifies reporting precision, recall, F1-score, and false-alarm rate rather than relying on accuracy alone.

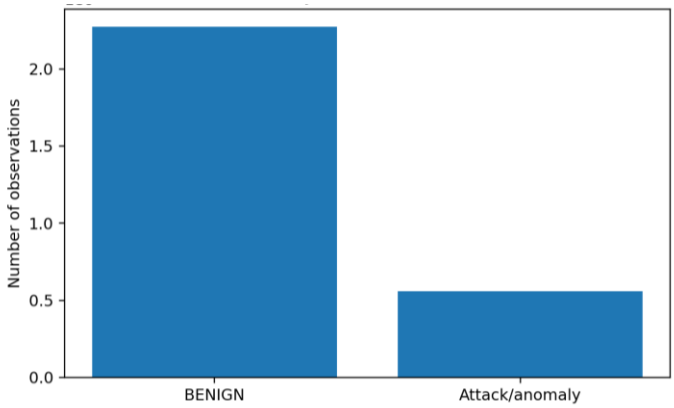

FIGURE 3. Binary distribution of benign and anomaly observations.

FIGURE 3. shows the overall imbalance between BENIGN records and attack/anomaly records in the complete dataset.

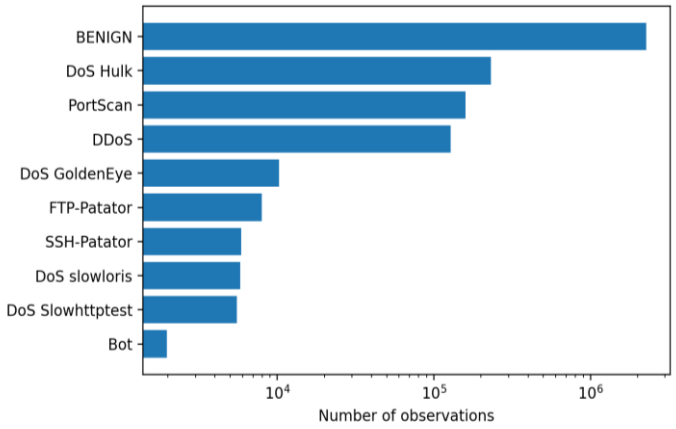

FIGURE 4. Major traffic labels in the dataset.

FIGURE 4 uses a logarithmic scale to show both common and relatively smaller traffic labels without hiding rare categories.

Table 5. Class distribution of traffic labels.

Class	Count
BENIGN	2,273,097
DoS Hulk	231,073
PortScan	158,930
DDoS	128,027
DoS GoldenEye	10,293
FTP-Patator	7,938
SSH-Patator	5,897
DoS slowloris	5,796
DoS Slowhttptest	5,499
Bot	1,966
Web Attack Brute Force	1,507
Web Attack XSS	652
Infiltration	36
Web Attack Sql Injection	21
Heartbleed	11

Table 5. confirms that the binary anomaly class is composed of heterogeneous attack types rather than a single abnormal event.

2. DESCRIPTIVE AND DISTRIBUTIONAL RESULTS

The descriptive statistics indicate that selected network-flow variables are highly skewed and heavy-tailed. Several variables have median values much smaller than their means, indicating that a small number of very large observations strongly affect classical averages. For example, packet counts, byte totals, Flow Bytes/s, and subflow byte measures show very high skewness and kurtosis. These patterns support the use of robust statistics because mean-standard deviation thresholds would be strongly influenced by extreme values.

Table 6. Descriptive statistics of selected flow variables.

Feature	Mean	Median	Std	IQR	MAD	Skewness	Kurtosis
Flow Duration	15,743,678.22	31,326.00	34,541,139.86	4,423,840.00	31,305.00	2.03	2.47

Feature	Mean	Median	Std	IQR	MAD	Skewness	Kurtosis
Total Fwd Packets	9.25	2.00	714.03	4.00	1.00	234.63	57,556.85
Total Backward Packets	10.17	2.00	941.34	3.00	1.00	237.94	59,840.59
Total Length of Fwd Packets	560.44	60.00	6,795.01	181.00	56.00	154.69	33,936.40
Total Length of Bwd Packets	16,472.53	120.00	2,281,157.44	562.00	120.00	239.23	59,726.59
Flow Bytes/s	1,534,579.58	4,285.09	26,352,712.40	159,881.58	4,285.09	40.75	2,208.91
Flow Packets/s	71,851.31	106.61	258,213.92	23,806.82	106.45	5.50	35.57
Fwd Packets/s	64,632.48	59.67	250,885.86	12,344.15	59.59	5.68	37.22
Bwd Packets/s	7,318.72	16.89	39,625.88	7,751.84	16.89	20.88	594.45
Packet Length Mean	187.08	56.60	327.08	124.40	53.00	2.20	3.97
Packet Length Std	329.70	24.65	683.65	216.83	24.65	2.63	6.98
Average Packet Size	208.53	72.00	356.39	152.25	67.00	2.29	4.64
Subflow Fwd Bytes	560.44	60.00	6,795.01	181.00	56.00	154.69	33,936.40
Subflow Bwd Bytes	16,472.37	120.00	2,281,295.29	562.00	120.00	239.25	59,736.61
Active Mean	84,147.86	0.00	648,645.70	0.00	0.00	27.20	1,733.79
Idle Mean	9,373,507.37	0.00	25,229,609.95	0.00	0.00	2.82	6.49

Table 6 shows strong skewness and heavy-tailed behavior, confirming the need for the median, IQR, MAD, and empirical tail thresholds.

3. DETECTION PERFORMANCE

The detection results show that the empirical probability-tail rule produced the strongest balance for conservative early warning. The best-performing rule flagged an observation when at least one selected feature fell outside the 0.1% empirical tails of the benign baseline. This rule achieved 84.80% accuracy and 87.30% precision, with a low false-alarm rate of 2.03%. Its recall was 43.63%, indicating that the rule is conservative: it identifies fewer attacks than aggressive thresholds but produces high-confidence alerts with limited benign misclassification.

Table 7. Leading threshold performance results.

Method	Threshold	TP	FP	TN	FN	Accuracy	Precision	Recall	F1-score	False-alarm rate
Empirical probability-tail rule	0.1% tails; count>=1	243,287	35,401	1,707,778	314,359	0.8480	0.8730	0.4363	0.5818	0.0203

Method	Threshold	TP	FP	TN	FN	Accuracy	Precision	Recall	F1-score	False-alarm rate
Empirical probability-tail rule	1.0% tails; count>=1	270,503	146,316	1,596,863	287,143	0.8116	0.6490	0.4851	0.5552	0.0839
IQR feature-count rule	20xIQR; count>=2	287,542	230,157	1,513,022	270,104	0.7826	0.5554	0.5156	0.5348	0.1320
IQR feature-count rule	10xIQR; count>=4	244,110	114,850	1,628,329	313,536	0.8138	0.6800	0.4378	0.5326	0.0659
IQR feature-count rule	15xIQR; count>=2	290,162	263,819	1,479,360	267,484	0.7691	0.5238	0.5203	0.5220	0.1513
IQR feature-count rule	8xIQR; count>=4	245,911	150,231	1,592,948	311,735	0.7992	0.6208	0.4410	0.5157	0.0862
Robust z-score feature-count rule	z>50; count>=2	457,907	763,708	979,471	99,739	0.6247	0.3748	0.8211	0.5147	0.4381
IQR feature-count rule	12xIQR; count>=2	296,023	301,443	1,441,736	261,623	0.7553	0.4955	0.5308	0.5125	0.1729
IQR feature-count rule	20xIQR; count>=1	330,687	409,768	1,333,411	226,959	0.7233	0.4466	0.5930	0.5095	0.2351
Robust z-score feature-count rule	z>40; count>=2	459,028	790,055	953,124	98,618	0.6138	0.3675	0.8232	0.5081	0.4532

Table 7. shows the trade-off among threshold methods. Conservative empirical tails reduce false alarms, while aggressive robust z-score rules increase recall at the cost of many false positives.

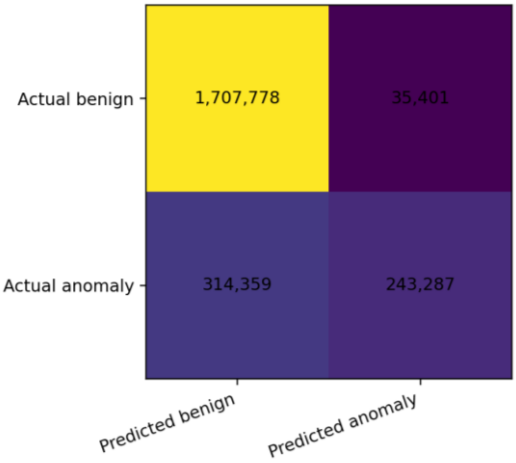

FIGURE 5. Confusion matrix for the best statistical threshold.

FIGURE 5. shows that the leading probability-tail rule correctly classified most benign holdout records while detecting 243,287 attack observations.

4. THRESHOLD SENSITIVITY ANALYSIS

Threshold sensitivity analysis shows that different robust rules serve different operational objectives. The 0.1% empirical probability-tail threshold provides the best precision and lowest false-alarm rate among the leading rules. The 1.0% empirical tail increases recall slightly but raises false alarms. IQR feature-count rules provide moderate recall and precision depending on the multiplier and required number of abnormal feature signals. Robust z-score rules with very large thresholds can produce higher recall, but their false-alarm rates become too high for conservative early-warning use.

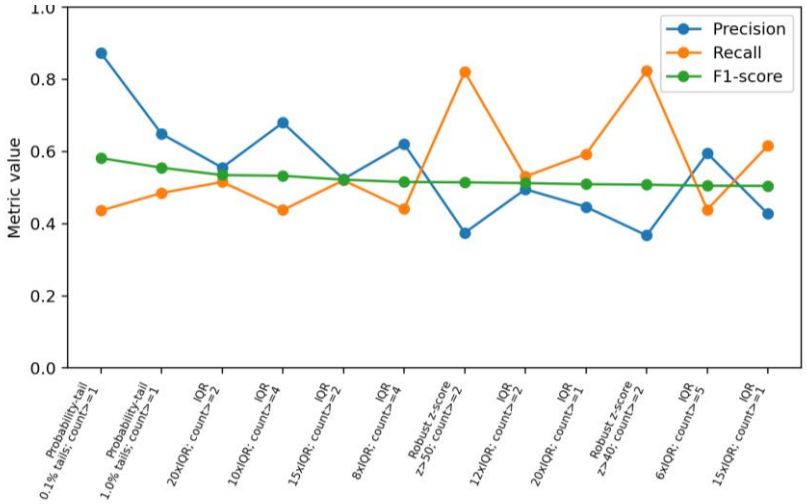

FIGURE 6. Threshold sensitivity across leading statistical rules.

FIGURE 6. shows that precision, recall, and F1-score move differently across thresholds, confirming that threshold choice should be aligned with operational monitoring priorities.

IX. DISCUSSION

The empirical findings support the value of robust statistical monitoring as an interpretable early-warning layer for complex network systems. The strictest probability-tail rule produced high precision and a very low false-alarm rate, which is important in practical monitoring because excessive false alarms can overload analysts and reduce trust in alerting systems. Although recall was moderate, the rule identified a large number of attack records while maintaining conservative alert quality. This makes the framework more suitable for high-confidence anomaly flagging than for exhaustive attack detection.

The results also confirm that network-flow data are not well represented by simple symmetric assumptions. The selected features show extreme skewness, heavy tails, and large differences between means and medians. Under such conditions, mean-standard deviation rules can be unstable because extreme values influence both the center and scale of the distribution. Robust medians, MAD values, quartiles, and empirical benign-tail quantiles provide more defensible baselines for operational monitoring.

The Monday calibration design is a methodological strength because thresholds were estimated from benign traffic only and then applied to later holdout files. This design avoids using attack records to define the normal baseline and supports a realistic interpretation of monitoring. However, benign network behavior may shift from day to day, which can affect threshold performance. The low false-alarm rate of the leading rule suggests that the empirical-tail boundary remained selective across holdout benign traffic, but the moderate recall indicates that not all attack categories appear as extreme deviations in the selected statistical features.

The framework should therefore be interpreted as a transparent monitoring layer rather than a complete intrusion-detection replacement. In operational environments, such a statistical layer can be used to prioritize high-confidence alerts, support rule auditing, and provide interpretable evidence for abnormal traffic. For maximum attack coverage, the framework could be combined with additional domain rules, multi-class attack modeling, or complementary predictive classifiers. Its main contribution is not maximal recall but robust, explainable, and reproducible anomaly screening.

The performance trade-off is particularly important for risk-sensitive contexts. A high-precision rule is useful when false positives are costly and analyst attention is limited. A higher-recall rule may be useful when missing attacks is more costly than investigating benign alerts. The framework allows these choices to be made explicitly by changing tail probability, IQR multiplier, robust z-score threshold, or feature-count requirement. This transparency is a major advantage over detection systems whose threshold behavior is difficult to interpret.

Taken together, the study makes theoretical, practical, and future-oriented contributions by operationalizing robust statistical concepts within a complete empirical framework for network-flow monitoring and by providing cybersecurity analysts with an auditable method for building early-warning thresholds from benign historical traffic. The framework shows how normal-baseline calibration, resistant estimates, probability-tail boundaries, and feature-count aggregation can produce transparent alerts that are traceable to feature-level violations. However, the empirical scope remains limited to network intrusion data, binary anomaly mapping, and sixteen interpretable flow variables; future research should therefore test all available features, develop attack-specific threshold profiles, compare the framework against supervised and unsupervised benchmarks, and adapt the procedure to streaming traffic with controlled baseline updating.

X. CONCLUSION

This study developed and evaluated a robust statistical framework for interpretable anomaly detection in complex network systems using CICIDS2017-style intrusion data. The framework integrated data audit, benign-baseline calibration, robust estimation, IQR boundaries, MAD-based deviation scoring, empirical probability-tail monitoring, feature-count aggregation, and standard binary performance evaluation. The empirical dataset contained 2,830,743 observations, including 2,273,097 benign records and 557,646 attack/anomaly records. Monday benign traffic was used as the calibration baseline, while the remaining days formed a holdout evaluation set of 2,300,825 observations. The results showed that the most conservative empirical probability-tail rule achieved 84.80% accuracy, 87.30% precision, an F1-score of 0.5818, and a false-alarm rate of 2.03%. This result indicates that robust statistical monitoring is especially valuable when the operational objective is high-confidence alerting and low false-alarm burden. The framework is transparent because each anomaly decision can be traced to a statistical threshold and a feature-level violation. It is also reproducible because the thresholds are derived from observable benign baseline distributions rather than opaque model parameters. The study concludes that robust statistical anomaly detection can provide a useful foundation for early-warning systems in cybersecurity and other complex data environments. Future work should extend the approach to all available features, multi-class attack profiles, streaming traffic, and comparative hybrid frameworks that combine statistical interpretability with broader detection coverage.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

[1]. Aggarwal, C. C. (2017). Outlier analysis (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-47578-3

[2]. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31. https://doi.org/10.1016/j.jnca.2015.11.016

[3]. Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. Data Mining and Knowledge Discovery, 29(3), 626-688. https://doi.org/10.1007/s10618-014-0365-y

[4]. Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2014). Network anomaly detection: Methods, systems and tools. IEEE Communications Surveys & Tutorials, 16(1), 303-336. https://doi.org/10.1109/SURV.2013.052213.00046

[5]. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the ACM SIGMOD International Conference on Management of Data, 93-104. https://doi.org/10.1145/335191.335388

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

[7]. Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenkova, B., Schubert, E., Assent, I., & Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. Data Mining and Knowledge Discovery, 30(4), 891-927. https://doi.org/10.1007/s10618-015-0444-8

[8]. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1-58. https://doi.org/10.1145/1541880.1541882

[9]. Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition, 74, 406-421. https://doi.org/10.1016/j.patcog.2017.09.037

[10]. Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLOS ONE, 11(4), e0152173. https://doi.org/10.1371/journal.pone.0152173

[11]. Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and Data Engineering, 26(9), 2250-2267. https://doi.org/10.1109/TKDE.2013.184

[12]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

[13]. Hubert, M., Rousseeuw, P. J., & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. Journal of Computational and Graphical Statistics, 21(3), 618-637. https://doi.org/10.1080/10618600.2012.672100

[14]. Lazarevic, A., Ertoz, L., Ozgur, A., Kumar, V., & Srivastava, J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. Proceedings of the Third SIAM International Conference on Data Mining, 25-36. https://doi.org/10.1137/1.9781611972733.3

[15]. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. Proceedings of the IEEE International Conference on Data Mining, 413-422. https://doi.org/10.1109/ICDM.2008.17

[16]. Maseer, Z. K., Yusof, R., Bahaman, N., Mostafa, S. A., & Foozy, C. F. M. (2021). Benchmarking of machine learning for anomaly-based intrusion detection systems in the CICIDS2017 dataset. IEEE Access, 9, 22351-22370. https://doi.org/10.1109/ACCESS.2021.3056614

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

[18]. Pang, G., Shen, C., Cao, L., & van den Hengel, A. (2021). Deep learning for anomaly detection: A review. ACM Computing Surveys, 54(2), 1-38. https://doi.org/10.1145/3439950

[19]. Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. Signal Processing, 99, 215-249. https://doi.org/10.1016/j.sigpro.2013.12.026

[20]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

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

[22]. Rosay, A., Cheval, E., Carlier, F., & Leroux, P. (2022). Network intrusion detection: A comprehensive analysis of CIC-IDS2017. Proceedings of ICISSP 2022, 25-36. https://doi.org/10.5220/0010774000003120

[23]. Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(2), e1236. https://doi.org/10.1002/widm.1236

[24]. Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., & Muller, K. R. (2021). A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE, 109(5), 756-795. https://doi.org/10.1109/JPROC.2021.3052449

[25]. Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: A comprehensive evaluation. Proceedings of the VLDB Endowment, 15(9), 1779-1797. https://doi.org/10.14778/3538598.3538602

[26]. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of ICISSP 2018, 108-116. https://doi.org/10.5220/0006639801080116

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

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

[29]. Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining, 5(5), 363-387. https://doi.org/10.1002/sam.11161

Appendix A. Main Analytical Variables

Table A1. Main features used in the robust statistical framework.

Variable	Role	Interpretation
Flow Duration	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Total Fwd Packets	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Total Backward Packets	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Total Length of Fwd Packets	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Total Length of Bwd Packets	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Flow Bytes/s	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Flow Packets/s	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Fwd Packets/s	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Bwd Packets/s	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Packet Length Mean	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Packet Length Std	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Average Packet Size	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Subflow Fwd Bytes	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Subflow Bwd Bytes	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Active Mean	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.
Idle Mean	Input feature	Network-flow timing, volume, rate, packet-length, or activity indicator.