This document contains the detailed characteristics of the datasets used in the literature, evaluating the use of synthetic benchmarks like CMU CERT versus real-world data sources. The authors utilized **NotebookLM** (processing 2–5 papers per prompt) to identify dataset versions, volumes, and the specific strategies used by researchers to address the class imbalance problem. Every entry was manually verified by the authors to ensure the accuracy of the reported data and to confirm the specific dataset utilized.

Prompt 3: Datasets and Data Sources (RQ3)

Task: You are a research assistant. Analyze the uploaded papers and fill the **Datasets and Data Sources Table** using the following definitions.

Instructions for Columns:

- **Paper ID:** The identifier for this study (Authurs,year).
- **Dataset Source:** Identify where the data came from. Look for: *CMU CERT (include version like r4.2 or r6.2), Enron Email, TWOS, LANL, or "Private/Real-world corporate logs."*
- **Data Type:** List the specific types of activity logs used. Look for: *HTTP/Web logs, Email, Logon/Logoff, File operations, USB/Device usage, or Psychometric (LDAP) data.*
- **Volume / Scale:** Extract the size of the dataset mentioned in the study. Look for: *Number of users (e.g., 1,000 users), number of events (e.g., 1M log rows), or duration (e.g., 18 months).*
- **Behavioral Focus:** In 3–5 words, describe the specific behavior being monitored (e.g., "Data exfiltration via USB," "After-hours login anomalies," or "Sentiment analysis of emails").

Output Format: Provide the results in a single-row table with these columns: | Authors (year) | Dataset Source | Data Type | Volume / Scale | Behavioral Focus |

Constraint: Use ONLY information found in this paper. If the scale or version is not mentioned, write "Not Specified."

Paper ID	Dataset Source	Data Type	Volume / Scale	Behavioral Focus
Adun (2023)	Kaggle (CERT v1 subset)	System logs	350 labeled samples	Malicious activity prediction
Alabdulkareem (2022)	Unspecified log source	Activity logs (USB, Email)	5,000 samples	Data leakage detection
Ahmadi (2025)	Simulated environment	User session data	10,000 sessions	Real-time risk scoring (ZTA)
Ahmed (2025)	Live network environment at the Canadian Institute of Technology, Tirana, Albania	Security logs, firewall logs, system logs from Windows machines (Windows 11 OS, Windows 2022 Server)	60-90 days of data collection	Insider user behavior anomalies
Al Hammadi et al. (2021)	Data collected from 17 people using an Emotiv Insight EEG device ; Images from the scientific international affective picture system (IAPS)	EEG brainwave signals (THETA, ALPHA, LOW_BETA, HIGH_BETA, GAMMA bands from AF3, T7, Pz, T8, AF4 electrodes)	17 subjects; 20 images; >10K raw EEG data	Emotional states; insider risk assessment
Ali et al. (2025)	Curated by an expert team of US Army Insider Threat hub analysts	Narrative text data (insider threat cases, analytical reports)	1,306 insider threat cases	Classifying insider threat levels; identifying threat topics
ALmihqani et al 2021	CMU CERT r4.2	Logon/Logoff, Device, Email, HTTP, Files	1000 employees, 18 months, 32,770,222 event records	Anomalous insider activities, user behavior
Almusawi 2024	CMU CERT r4.2	Logon/Logoff, Device usage, Email, HTTP/Web logs, Psychometric	1000 online users, 18-month duration, over 32 million entries	Insider risks, user behavior, network activity

AL-Mihqani 2022 A new intelligent multilayer	CMU CERT r4.2	Logon/Logoff, USB/Device usage, HTTP/Web logs, File operations, Email	1000 online users, 18-month duration, over 32 million entries	Malicious insider activities, risky behavior
Alshehari 2023 (IF)	CMU CERT r4.2	Logon/Logoff, USB/Device usage, HTTP/Web logs, File operations, Email	1000 employees, 18 Months	Anomaly detection of insider activities
Al-shehari & Alsawail 2023 (Random Undersampling)	CMU CERT r4.2	Logon/Logoff, USB/Device usage, HTTP/Web logs, File operations, Email	1000 employees, 18 Months	Data exfiltration, unusual device use
Al-shehari (CNN) (2024)	CMU CERT r4.2	Logon/Logoff, USB/Device usage, HTTP/Web logs, File operations, Email	1000 employees, 18 Months	Malicious insider activities, user behavior
AL-SHEHARI (LOF) (2024)	CMU CERT r4.2	Logon/Logoff, USB/Device usage, HTTP/Web logs, File operations, Email	1000 employees, 18 Months	Rare malicious activities, user behavior deviations
Alshehri (2022)	CERT 4.2	Log-in, log-off, device, and HTTP	>1,000 users; 32 million events; 502 days	Anomaly behavior in activity streams
Amiri-Zarandi (2023)	CERT v6.2	Logon, File, Email, HTTP, Device, Psychometric, and Organization Structure	4,000 employees; 11 local participant units	Insider threats in IoT environments
Al-Shehari (2021)	CMU-CERT R4.2	Logon/off, file operations, HTTP, email, and removable device	1,000 insiders; over 17 months	Data leakage before leaving organization
Amuda (2022)	CERT Insider Threat Synthetic dataset v4.2	Logon/logoff, device usage, and file access	1,000 users; 3,015,990 log lines; 3 months	Sequences of user actions
Anju (2024)	CERT R4.2 and R6.2	Logon, Device, HTTP, Email, File, Psychometric, and LDAP	1,000–4,000 users; 135,117,169 log entries	Multivariate time-series user behavior

Anakath (2022)	Not Specified (Open-source datasets used in Cooja simulator)	User interaction behavior (mouse movements/clicks and keystrokes)	20 nodes; data blocks of 500 to 3,000 instances	User interaction behavior patterns
Asha S (2023)	CMU CERT v3.2	Logon/logoff, HTTP, Email, File access, USB usage, Psychometric, and LDAP	1,000 users; 135,117,169 log events; 516 days	Wikileaks and career portal visits
Cai X (2024)	CERT r4.2 and r5.2	Logon, Web visits (HTTP), File operations, and Device connection	1,000–2,000 employees; 32.7M–79.8M activities; 18 months	Activity-level real-time behaviors
Dong J (2025)	CMU CERT r5.2 and r6.2	Five behavioral log types (Logon, File, HTTP, Email, and Device)	2,000–3,995 users; Various temporal granularities	Contextual employee behavioral patterns
Eshmawi (2026)	CMU-CERT r5.2	Logon/logoff, email, web (HTTP), file, thumb drive (device), and psychometric/LDAP	2,000 users; 18-month duration	Data exfiltration and sabotage
Feng (2025)	CERT R4.2	Logon/logoff, file activity, HTTP/browsing, email, and device (USB) usage	1,000 staff members; 32,770,222 total behaviors	Multi-granularity behavior anomalies
Ferraro (2025)	PicoDomain and CERT r5.2	Zeek logs (Connection, Kerberos), logon, file access, device usage, and psychometric indicators	PicoDomain: 3 days, 15 users; CERT: 18 months, 1,901 users	AI-reasoned log anomaly detection
Gayathri, R.G. (2025)	**CERT v4.2 and v5.2**,	Logon/logoff, email, web (HTTP), file, thumb drive (device), and psychometric (LDAP),,	1,000–2,000 users; 330,452 to 1,048,575 instances,	**Insider-driven poison-label backdoors**,,
Gayathri, B. (2025)	**NSL-KDD, UNSW-NB15**, and **CERT**	Behavioral logs (duration, protocol, service, flags, bytes, urgent, hot),	**Not Specified** (20% train/test split used for evaluation)	**Insider threats in cloud**,
Gayathri, R.G. (2024)	**CERT r4.2 and r5.2**	Logon/logoff, email, web, file, thumb drive (device), organizational structure, and psychometric	1,000–2,000 users; 18-month duration	**Scenario-based malicious insider activities**
Gonzales (2025)	**Synthetic datasets** and **UniProt database** (Real-world protein sequences used as a behavioral analogy)	User action sequences ({C,R,U,D,E}) and Amino acid single-letter codes	250,000 to 3,000,000 synthetic sequences; 100 UniProt sequences (lengths 20–140)	Contextual user action sequences
Gupta (2024)	**CMU CERT r4.2**	Behavioral logs (Profession, request count/type, leakage	10,000 training samples	Malicious data distribution intentions

		records, leak ratio, and leak channel)		
Hafizu Rhman (2022)	**CMU CERT r4.2**	User activity logs (Logon/logoff, file operations, email, web browsing, and USB usage)	1,000 users; 18 months; 4,002,500 total entries	Unintentional insider threat activities
Haq (2022)	Enron Email Dataset	E-mail messages and detailed financial information	250,000 e-mail messages; 30 GB total data volume	Identifying persons of interest
He (2022)	Enron Email, and CMU-CERT r6.2	Email text (subject/body) and system logs (Logon, device, file, email, web, and psychometric)	4,000 users; 18 months duration	Phishing and behavior anomalies
He (2024)	Enron Email, and CMU-CERT r6.2	Email text/attachments and host logs (LDAP, logon, file, device, email, http, and psychometric)	4,000 users; 18 months; 186,331 normal and 8,794 attack sequences	Social engineering attack chains
Huang (2025)	**KDD-UEBA, NSL-UEBA, CIC-UEBA,** and **KDD-UNBLogs** (benchmark integrations)	Network traffic and user behavior records	**Not Specified** (Appendix B containing details was not provided in the text)	Masquerader attack detection
Jaiswal (2024)	**CERT r4.2** and **CERT r5.2**	HTTP, file, device, logon, and email	**r4.2:** 1,000 users, 7 months, 330,452 instances; **r5.2:** 2,000 users, 693,649 instances	Day-long activity sequences
Janjua (2021)	**Enron Email** and **TWOS**	Emails (Enron); Keystrokes, host display, cursor, emails, login, and network traffic (TWOS)	**Enron:** 151 employees, ~517,431 emails, 4-year window; **TWOS:** 24 users, 5-day span	Analysis of malicious email
Kamatchi (2025)	**Simulated dataset** and **X-IIoTID**	**IoT logs** (access patterns, login times, data transmission), **network traffic**, and **system resources**	**X-IIoTID**: 421,417 benign and 399,417 malicious records; **Simulated**: 100 edge nodes, 1 hour duration	**IoT behavior deviations**
Kong (2025)	**CERT r4.2**, **HCT**, **SEA**, and **Purdue UNIX (PU)**	**Device-level logs** (authentication, file operations, system calls, sensor events) and **shell commands**	**CERT r4.2**: 1,000 users, 18 months, 32M records; **SEA**: 750k logs; **HCT**: 13,446	**Dual-perspective behavior patterns**

			commands; **PU**: 2 years	
Kotb (2025)	**CMU-CERT r4.2**	**Host logs** (logon, email, device, file, HTTP) and **Psychometric (LDAP)** Big Five scores	1,000 employees over 17 months; **32,770,222 total events**	**Real and synthetic threats**
Lavanya (2024)	**CMU CERT r6.1 & r6.2**	Logon, device, HTTP, email, file, and psychometric	3,995 users; 17 months (01-11-2009 to 01-05-2011)	Legitimate access abnormal behavior
Lavanya (2025)	**CMU CERT r4.1 & r4.2**; **Private data** (LAN, Wi-Fi, Raspberry Pi)	File, device, HTTP, email, user logs, and psychometric	CERT: 1,000 users; 32,770,222 activities; 17 months. Private: 30 users; 102,457 activities; 3 days	Exfiltration, theft, and sabotage
Le & Zincir-Heywood (2021)	**CERT R4.2**, **CERT R6.2**, **LANL**, and **TWOS**	Logon/off, email, web, file, thumb drive, mouse, and keystroke	R4.2 (1,000 users), R6.2 (4,000 users), LANL (58 days), TWOS (24 users)	Transitions in user behaviors
Li et al. (2024)	**CMU CERT (r4.2, r6.2)** and **UMDWikipedia**	Logon/logoff, email, device, file, HTTP, and wiki edits	1,000 to 4,000 users; over 135M activities	**Exfiltration, theft, and sabotage**
Li et al. (2023)	**CMU CERT (r4.2, r6.2)**	Logon/logoff, email, device, file, and HTTP logs	1,000 to 4,000 users; up to 135,117,169 activities	**Exfiltration, intellectual property, sabotage**
Liu et al. (2025)	**Cert-r4.2, Cert-r5.2**, and **DARPA1999**	Logon/logoff, device usage, file operations, HTTP access, and email logs	**Cert-r4.2**: 1,000 users, 370k+ sessions; **Cert-r5.2**: 2,000 users, 215k+ sessions	**Multi-step evidence detection**
Mehmood et al. (2023)	**Customized CERT dataset**	System logs (PC, Activity), Email (Size, Attachments), and External Device logs	**10,000 samples** (per feature range) [125, Fig. 10]	**Cloud privilege escalation attacks**
Medvedev et al. (2025)	**CMU** and **KeyRecs**	Keystroke dynamics (timing of pressing and releasing keys)	**150 participants** (99 from KeyRecs, 51 from CMU); 400 repetitions per CMU user	**Authentication via typing rhythm**
Mehnaz & Bertino (2021)	**Wikipedia** file repository	**Block-level I/O traces** (blktrace) including sector/block numbers, access type, and timestamps	**77 users**; 560 files; **2-month** duration	Fine-grained file system accesses

Mladenovic et al. (2024)	**CMU** CERT Insider Threat Test Dataset (Scenarios **4.1, 4.2, 4.3**)	**Email, HTTP, and File content** (textual/NLP-based)	Aggregate logs over a **500-day** interval	Sentiment and context analysis
Nasir et al. (2021)	**CMU CERT r4.2**	**Logon/Logoff, Device (USB), HTTP, Email, File activity**, and **LDAP**	**1,000 users**; **32,770,220** total log rows	Session-based user behavioral analysis
Nikiforova et al. (2024)	Private/Real-world corporate logs (from **e-StepControl**),	**Audit records** (Login, exit, system menu viewing, sending email, subsystem access)	**488 users**; **2,776,770** total actions	Deviations from group behavior,
Pal et al. (2023)	**CMU CERT** (versions **v4.2, v5.2, v6.2**),	**Logon/Logoff, Device (USB), File, Email**, and **HTTP** logs	**1,000–4,000 users**; **17 months** duration	Sequential single-day activity patterns,
Patel & Iyer (2025)	Private/Real-world corporate logs (from **Zenodo Cloud-based UEBA set**),	**Link shares, file accesses, logins**, and **config changes**	**4,000+ users**; **60 million** events; 15 GB	Cloud-based user interaction patterns,
Peccatiello et al. (2023)	**CMU CERT (r4.2)**,	**Device, Email, HTTP, Logon, File**, and **LDAP** data	**1,000 users**; **16.5 months**; **470,608** sessions,	Stream-based session-level anomalies,
Pennada (2024)	**CMU CERT (r5.2)**	**Not Specified** (830 behavioral features mentioned)	**693,649 samples**	Anomalous patterns and behaviors
Pennada (2025)	**CMU CERT**	**Login, File operations, and Email interactions**	**693,649 samples**	Explicit and implicit indications
Perez-Miguel et al. (2025)	**SPEDIA** (Real exercise, role-based simulation, and CERT)	**File, Command, Session, Email, HTTP, and Device**	**17 users; 72,250 logs; 24 days**	Disgruntled employee attack narrative
Qawasmeh & AlQahtani (2025)	**Synthetic organizational logs**	**Logins, Files, Software, Web, Network, and Social Engineering**	**500 employees; 10,000+ records; 4 weeks**	Real-time abnormal daily activities
Randive et al. (2023)	**CMU CERT (r4.2)**	**Logon, Device, File, Email, and HTTP**	**1,000 users; 17 months; 32.7M events**	Scenario-specific visual patterns
Senevirathna et al. (2025)	CERT Insider Threat Identification Dataset (Version Not Specified)	Logon/logoff, device interactions, file actions, psychometric traits, and CCTV video data	Not Specified	Integrated cyber-physical threat profiling
Song et al. (2024)	CMU CERT v6.2	Login, HTTP, file, email, and device logs	4,000 employees over 516 days	Time-aware user behavior rhythms

Tabassum et al. (2024)	Private/Real-world corporate logs (Hospital in North England)	Audit logs (Date, Device, User ID, Routine, Patient ID, Duration, Discharge dates)	1,007,727 audit log entries	Contextual anomalies in EHR
Tian T et al. (2025) [ITDSTS]	CMU CERT r4.2	Logon, device, email, file, and weblogs	1,000 insiders over 17 months (Jan 2010 – May 2011)	Scenario-oriented specific threat classification
Tian Z et al. (2024) [DSDLITD]	CERT Insider Threat Dataset v6.2	Device, email, file, HTTP, and Logon/logoff activity	4,000 users, 516 days, 135,117,169 events	Fusing multichannel communication anomalies
Villarreal-Vasquez et al. (2023)	Commercial network (EDR)	Process, Module, File, Directory, and Registry events	30 computers, 38.9 million events over 20 days	Order-aware persistent attack detection
Wall & Agrafiotis (2021)	CMU CERT r4.1	Device, email, file, http, and logon	1,000 employees from Jan 2010 to June 2011	Mapping causal behavioral dependencies
Wang & El Saddik (2023)	CMU CERT r4.2 and r6.2,	Logon, device, file, email, http, and psychometric data,	1,000–4,000 users; 32M–135M activities	Digital twin self-attention profiling,
Wang Zhi et al. (2024) FedITD	CMU CERT r6.2	Logon, email, http, file, device, and LDAP	4,000 users; 135,117,169 activities	Federated privacy-preserving threat detection,
Wang Jiarong et al. (2023) Deep Cluster	CMU CERT r4.2	Logins, file access, email, web, and USB	1,000 employees over 17 months	Deep clustering multi-source events,
Wei Yichen et al. (2021)	CMU CERT r6.2,	Device, email, file, network, logon, and LDAP,	4,000 users over 516 days	Proactive forensic anomaly prediction,
Wei Zhiyuan et al. (2024)	CMU CERT r4.2,	Network, system, and user activity logs,	1,000 employees	Personalized hybrid behavioral monitoring,
Wen et al. (2023)	Enron Email,	Email content and timestamps	151 employees; Jan 2000 – Dec 2001,	Sentiment analysis of emails
Xiao Junchao et al. (2023)	CMU CERT r6.2,	Logon, email, http, file, and device logs	4,000 users; 18 months	Contextual user behavior relationships
Xiao Fengrui et al. (2025)	LANL; CMU CERT r6.2; TWOS	Authentication, NetFlow, DNS, Logon, File, Mouse, and Keystrokes	504 users (LANL); 1.4M events (TWOS); 40k events (CERT)	Multi-timescale user interaction behaviors
Xiao Haitao et al. (2024)	CMU CERT r5.2,	Logon, device, file, email, http, and psychometric	2,000 employees over 18 months	Integrating statistical/sequential info
Ye Xiaoyun et al. (2025)	CMU CERT v4.2 and v5.2	Login, device, http, file, email, and LDAP	1,000–2,000 users over 500–518 days,	Personalized federated behavior analysis

Yildirim & Anarim (2022)	Balabit Dataset (Extended),	Mouse usage patterns (coordinates/buttons)	20 users; 5 training sessions per user	Mouse dynamics behavioral biometrics
Zhu et al. (2024)	CMU CERT (r4.2 implied by user/threat count),	Logon, email, http, file, and device logs	1,000 employees over 1 year	Time/event feature reconstruction

The following table is produced to confirm that all studies that were reported to have used CERT dataset (58 Studies) are accurate as reported by NotebookLM

Paper ID	Dataset Source	Data Type	Volume / Scale	Behavioral Focus
Adun (2023)	CERT v1 subset	System logs	350 labeled samples	Malicious activity prediction
ALmihqani et al (2021)	CMU CERT r4.2	Logon/Logoff, Device, Email, HTTP, Files	1000 employees, 18 months, 32M+ events	Anomalous insider activities
Almusawi (2024)	CMU CERT r4.2	Logon, Device, Email, HTTP, Psychometric	1000 users, 18 months, 32M+ entries	Insider risks, network activity
AL-Mihqani (2022)	CMU CERT r4.2	Logon, USB, HTTP, File, Email	1000 users, 18 months, 32M+ entries	Malicious insider activities
Alshehari (2023)	CMU CERT r4.2	Logon, USB, HTTP, File, Email	1000 employees, 18 Months	Anomaly detection
Al-shehari & Alsawail (2023)	CMU CERT r4.2	Logon, USB, HTTP, File, Email	1000 employees, 18 Months	Data exfiltration, unusual device use
Al-shehari (CNN) (2024)	CMU CERT r4.2	Logon, USB, HTTP, File, Email	1000 employees, 18 Months	Malicious insider activities
AL-SHEHARI (LOF) (2024)	CMU CERT r4.2	Logon, USB, HTTP, File, Email	1000 employees, 18 Months	Rare malicious activities
Alshehri (2022)	CERT 4.2	Log-in, log-off, device, HTTP	>1,000 users; 32M events	Anomaly behavior
Amiri-Zarandi (2023)	CERT v6.2	Logon, File, Email, HTTP, Device, Psychometric	4,000 employees; 11 units	Insider threats in IoT environments
Al-Shehari (2021)	CMU-CERT R4.2	Logon/off, file, HTTP, email, device	1,000 insiders; 17 months	Data leakage
Amuda (2022)	CERT Synthetic v4.2	Logon, device, file access	1,000 users; 3M log lines	Sequences of user actions

Anju (2024)	CERT R4.2 and R6.2	Logon, Device, HTTP, Email, File, Psychometric	1,000–4,000 users; 135M logs	Multivariate time-series behavior
Asha S (2023)	CMU CERT v3.2	Logon, HTTP, Email, File, USB, Psychometric	1,000 users; 135M events	Wikileaks and career portal visits
Cai X (2024)	CERT r4.2 and r5.2	Logon, HTTP, File, Device	1,000–2,000 employees; 32M–79M activities	Activity-level real-time behaviors
Dong J (2025)	CMU CERT r5.2 and r6.2	Logon, File, HTTP, Email, Device	2,000–3,995 users	Contextual employee behavioral patterns
Eshmawi (2026)	CMU-CERT r5.2	Logon, email, HTTP, file, device, psychometric	2,000 users; 18 months	Data exfiltration and sabotage
Feng (2025)	CERT R4.2	Logon, file, HTTP, email, USB	1,000 staff; 32M behaviors	Multi-granularity anomalies
Ferraro (2025)	PicoDomain and CERT r5.2	Zeek logs, logon, file, device, psychometric	CERT: 18 months, 1,901 users	AI-reasoned log anomaly detection
Gayathri, R.G. (2025)	CERT v4.2 and v5.2	Logon, email, HTTP, file, device, psychometric	1,000–2,000 users	Insider-driven poison-label backdoors
Gayathri, B. (2025)	NSL-KDD, UNSW-NB15, CERT	Behavioral logs	Not Specified	Insider threats in cloud
Gayathri, R.G. (2024)	CERT r4.2 and r5.2	Logon, email, web, file, device, psychometric	1,000–2,000 users; 18 months	Scenario-based malicious activities
Gupta (2024)	CMU CERT r4.2	Behavioral logs (Profession, request count, leakage)	10,000 training samples	Malicious data distribution intentions
Hafizu Rhman (2022)	CMU CERT r4.2	Logon, file, email, web, USB	1,000 users; 4M entries	Unintentional insider threat
He (2022)	Enron Email, CMU-CERT r6.2	Email text, system logs	4,000 users; 18 months	Phishing and behavior anomalies
He (2024)	Enron Email, CMU-CERT r6.2	Email text, host logs	4,000 users; 18 months	Social engineering attack chains
Jaiswal (2024)	CERT r4.2 and r5.2	HTTP, file, device, logon, email	r4.2: 1k users; r5.2: 2k users	Day-long activity sequences

Kong (2025)	CERT r4.2, HCT, SEA, PU	Device-level logs, shell commands	CERT r4.2: 1k users, 32M records	Dual-perspective behavior patterns
Kotb (2025)	CMU-CERT r4.2	Host logs and Psychometric	1,000 employees; 32M events	Real and synthetic threats
Lavanya (2024)	CMU CERT r6.1 & r6.2	Logon, device, HTTP, email, file, psychometric	3,995 users; 17 months	Legitimate access abnormal behavior
Lavanya (2025)	CMU CERT r4.1 & r4.2; Private	File, device, HTTP, email, user logs	CERT: 1,000 users; 32M activities	Exfiltration, theft, and sabotage
Le & Zincir-Heywood (2021)	CERT R4.2, R6.2, LANL, TWOS	Logon, email, web, file, thumb drive, mouse	R4.2 (1k users), R6.2 (4k users)	Transitions in user behaviors
Li et al. (2024)	CMU CERT (r4.2, r6.2), Wikipedia	Logon, email, device, file, HTTP	1,000–4,000 users; 135M activities	Exfiltration, theft, and sabotage
Li et al. (2023)	CMU CERT (r4.2, r6.2)	Logon, email, device, file, HTTP	1,000–4,000 users; 135M activities	Exfiltration, IP theft, sabotage
Liu et al. (2025)	Cert-r4.2, Cert-r5.2, DARPA	Logon, device, file, HTTP, email	r4.2: 370k+ sessions; r5.2: 215k+	Multi-step evidence detection
Mehmood et al. (2023)	Customized CERT dataset	System logs, Email, External Device	10,000 samples	Cloud privilege escalation
Mladenovic et al. (2024)	CMU CERT (Scenarios 4.1-4.3)	Email, HTTP, File content (NLP)	Aggregate logs over 500 days	Sentiment and context analysis
Nasir et al. (2021)	CMU CERT r4.2	Logon, Device, HTTP, Email, File, LDAP	1,000 users; 32M rows	Session-based analysis
Pal et al. (2023)	CMU CERT (v4.2, v5.2, v6.2)	Logon, Device, File, Email, HTTP	1,000–4,000 users; 17 months	Sequential single-day patterns
Peccatiello et al. (2023)	CMU CERT (r4.2)	Device, Email, HTTP, Logon, File, LDAP	1,000 users; 470k sessions	Stream-based session anomalies
Pennada (2024)	CMU CERT (r5.2)	830 behavioral features	693,649 samples	Anomalous patterns
Pennada (2025)	CMU CERT	Login, File, Email	693,649 samples	Explicit and implicit indications
Perez-Miguel et al. (2025)	SPEDIA (Simulation and CERT)	File, Command, Session, Email, HTTP	17 users (SPEDIA)	Disgruntled employee narrative

Randive et al. (2023)	CMU CERT (r4.2)	Logon, Device, File, Email, HTTP	1,000 users; 32.7M events	Scenario-specific visual patterns
Senevirathna et al. (2025)	CERT Insider Threat ID Dataset	Logon, device, file, psychometric, CCTV	Not Specified	Cyber-physical threat profiling
Song et al. (2024)	CMU CERT v6.2	Login, HTTP, file, email, device	4,000 employees; 516 days	Time-aware user behavior rhythms
Tian T et al. (2025)	CMU CERT r4.2	Logon, device, email, file, weblogs	1,000 insiders; 17 months	Scenario-oriented threat classification
Tian Z et al. (2024)	CERT Insider Threat v6.2	Device, email, file, HTTP, Logon	4,000 users; 135M events	Fusing multichannel anomalies
Wall & Agrafiotis (2021)	CMU CERT r4.1	Device, email, file, http, logon	1,000 employees	Mapping causal dependencies
Wang & El Saddik (2023)	CMU CERT r4.2 and r6.2	Logon, device, file, email, http, psychometric	1,000–4,000 users	Digital twin self-attention profiling
Wang Zhi et al. (2024)	CMU CERT r6.2	Logon, email, http, file, device, LDAP	4,000 users; 135M activities	Federated privacy-preserving detection
Wang Jiarong et al. (2023)	CMU CERT r4.2	Logins, file access, email, web, USB	1,000 employees; 17 months	Deep clustering multi-source events
Wei Yichen et al. (2021)	CMU CERT r6.2	Device, email, file, network, logon, LDAP	4,000 users; 516 days	Proactive forensic anomaly prediction
Wei Zhiyuan et al. (2024)	CMU CERT r4.2	Network, system, and user activity logs	1,000 employees	Personalized hybrid behavioral monitoring
Xiao Junchao et al. (2023)	CMU CERT r6.2	Logon, email, http, file, device	4,000 users; 18 months	Contextual user behavior relationships
Xiao Fengrui et al. (2025)	LANL; CMU CERT r6.2; TWOS	Auth, NetFlow, DNS, Logon, File, Mouse	40k events (CERT)	Multi-timescale behaviors
Xiao Haitao et al. (2024)	CMU CERT r5.2	Logon, device, file, email, http, psychometric	2,000 employees; 18 months	Integrating statistical/sequential info
Ye Xiaoyun et al. (2025)	CMU CERT v4.2 and v5.2	Login, device, http, file, email, LDAP	1,000–2,000 users	Personalized federated behavior analysis
Zhu et al. (2024)	CMU CERT (r4.2 implied)	Logon, email, http, file, device	1,000 employees; 1 year	Time/event feature reconstruction

The table below was produced to cross reference between the name and numbers (IEEE format) and ensure the information is accurate

Seq.	Dataset Category	Ref #	Author Name (Reference List)
1	1. CMU CERT (Synthetic)	[8]	Ferraro (2025)
2	1. CMU CERT (Synthetic)	[9]	Gayathri (2025, Cloud)
3	1. CMU CERT (Synthetic)	[10]	Peccatiello (2023)
4	1. CMU CERT (Synthetic)	[13]	Al-Shehari (2024, Isolation Forest)
5	1. CMU CERT (Synthetic)	[14]	Pennada (2025)
6	1. CMU CERT (Synthetic)	[15]	Zhu (2024)
7	1. CMU CERT (Synthetic)	[16]	Gayathri (2025, Adv)
8	1. CMU CERT (Synthetic)	[22]	Wei Zhiyuan (2024)
9	1. CMU CERT (Synthetic)	[23]	Feng (2025)
10	1. CMU CERT (Synthetic)	[24]	Le & Zincir-Heywood (2021)
11	1. CMU CERT (Synthetic)	[27]	Kong (2025)
12	1. CMU CERT (Synthetic)	[29]	Al-Shehari & Alsowail (2021)
13	1. CMU CERT (Synthetic)	[30]	Mehmood (2023)
14	1. CMU CERT (Synthetic)	[31]	Kotb (2025)
15	1. CMU CERT (Synthetic)	[32]	Bin Sarhan (2023)
16	1. CMU CERT (Synthetic)	[33]	Li et al. (2023, DD-GCN)
17	1. CMU CERT (Synthetic)	[34]	Lavanya (2025)
18	1. CMU CERT (Synthetic)	[37]	Alabdulkareem (2022)
19	1. CMU CERT (Synthetic)	[38]	ALmihqani (2021)
20	1. CMU CERT (Synthetic)	[39]	Al-Shehari & Alsowail (2023)
21	1. CMU CERT (Synthetic)	[40]	Amiri-Zarandi (2023)
22	1. CMU CERT (Synthetic)	[41]	Amuda (2022)
23	1. CMU CERT (Synthetic)	[43]	Asha S (2023)
24	1. CMU CERT (Synthetic)	[44]	Eshmawi (2026)
25	1. CMU CERT (Synthetic)	[45]	Gupta (2024)
26	1. CMU CERT (Synthetic)	[46]	Huang (2025)
27	1. CMU CERT (Synthetic)	[48]	Lavanya (2024)
28	1. CMU CERT (Synthetic)	[49]	Nasir (2021)

29	**1. CMU CERT (Synthetic)**	**[52]**	Pennada (2024)
30	**1. CMU CERT (Synthetic)**	**[55]**	Senevirathna (2025)
31	**1. CMU CERT (Synthetic)**	**[57]**	Tian Z (2024)
32	**1. CMU CERT (Synthetic)**	**[58]**	Wall & Agrafiotis (2021)
33	**1. CMU CERT (Synthetic)**	**[59]**	Xiao Haitao (2024)
34	**1. CMU CERT (Synthetic)**	**[61]**	Al-Shehari (2024, CNN)
35	**1. CMU CERT (Synthetic)**	**[62]**	Al-Shehari (2023)
36	**1. CMU CERT (Synthetic)**	**[63]**	AL-Mihqani (2022)
37	**1. CMU CERT (Synthetic)**	**[64]**	Almusawi (2024)
38	**1. CMU CERT (Synthetic)**	**[67]**	Tian T (2025)
39	**1. CMU CERT (Synthetic)**	**[69]**	Gayathri (2024 Hybrid)
40	**1. CMU CERT (Synthetic)**	**[72]**	Liu (2025)
41	**1. CMU CERT (Synthetic)**	**[73]**	Wang & El Saddik (2023)
42	**1. CMU CERT (Synthetic)**	**[74]**	Wang Zhi (2024)
43	**1. CMU CERT (Synthetic)**	**[75]**	Alshehri (2022)
44	**1. CMU CERT (Synthetic)**	**[76]**	Anju (2024)
45	**1. CMU CERT (Synthetic)**	**[77]**	Cai X (2024)
46	**1. CMU CERT (Synthetic)**	**[78]**	Dong J (2025)
47	**1. CMU CERT (Synthetic)**	**[80]**	Hafizu Rhman (2022)
48	**1. CMU CERT (Synthetic)**	**[81]**	Jaiswal (2024)
49	**1. CMU CERT (Synthetic)**	**[83]**	Pal (2023)
50	**1. CMU CERT (Synthetic)**	**[84]**	Song (2024)
51	**1. CMU CERT (Synthetic)**	**[86]**	Wang Jiarong (2023)
52	**1. CMU CERT (Synthetic)**	**[87]**	Wei Yichen (2021)
53	**1. CMU CERT (Synthetic)**	**[89]**	Li et al. (2024, GMFITD)
54	**1. CMU CERT (Synthetic)**	**[90]**	Roy & Chen (2024)
55	**1. CMU CERT (Synthetic)**	**[91]**	Xiao Junchao (2023)
56	**1. CMU CERT (Synthetic)**	**[92]**	Xiao Fengrui (2025)
57	**1. CMU CERT (Synthetic)**	**[95]**	Randive (2023)
58	**1. CMU CERT (Synthetic)**	**[96]**	Ye Xiaoyun (2025)
59	**2. Real-World & Private**	**[36]**	Ahmed (2025)

60	**2. Real-World & Private**	**[50]**	Nikiforova (2024)
61	**2. Real-World & Private**	**[85]**	Villarreal-Vasquez (2023)
62	**2. Real-World & Private**	**[56]**	Tabassum (2024)
63	**2. Real-World & Private**	**[93]**	Al Hammadi (2021)
64	**2. Real-World & Private**	**[26]**	Ali (2025)
65	**2. Real-World & Private**	**[51]**	Patel & Iyer (2025)
66	**2. Real-World & Private**	**[82]**	Mehnaz (2021)
67	**2. Real-World & Private**	**[94]**	Medvedev (2025)
68	**3. Other Simulated/Lab**	**[28]**	Rauf (2021)
69	**3. Other Simulated/Lab**	**[35]**	Adun (2023)
70	**3. Other Simulated/Lab**	**[42]**	Anakath (2022)
71	**3. Other Simulated/Lab**	**[47]**	Kamatchi (2025)
72	**3. Other Simulated/Lab**	**[53]**	Perez-Miguel (2025)
73	**3. Other Simulated/Lab**	**[54]**	Qawasmeh (2025)
74	**3. Other Simulated/Lab**	**[60]**	Yildirim (2022)
75	**3. Other Simulated/Lab**	**[79]**	Gonzales (2025)
76	**3. Other Simulated/Lab**	**[88]**	Ahmadi (2025)
77	**4. NLP & Communication**	**[25]**	He (2024)
78	**4. NLP & Communication**	**[65]**	Janjua (2021)
79	**4. NLP & Communication**	**[66]**	Mladenovic (2024)
80	**4. NLP & Communication**	**[68]**	Wen (2023)
81	**4. NLP & Communication**	**[70]**	Haq (2022)
82	**4. NLP & Communication**	**[71]**	He (2022)