

This document contains the systematic extraction of data collection, transformation steps, and feature engineering techniques from the 82 primary studies. The authors utilized **NotebookLM** (processing 2–5 papers per prompt) to identify specific preprocessing steps. The extraction followed a specific prompt requiring the assistant to identify data preprocessing and feature dimensionality. All AI-generated results were cross-checked manually by the authors to ensure the accuracy of technical information and to verify that all extracted features were explicitly stated in the source. NotebookLM has a feature where it highlights where the text come from, but the authors still verified that information by reading the paper to ensure that the information is accurate and correct any misinformation.

Prompt 1: Data Preprocessing & Feature Engineering (RQ1)

Task: You are a research assistant specializing in cybersecurity data science. Analyze the uploaded papers and fill the **Preprocessing & Feature Engineering Table** based on the criteria below.

Instructions for Columns:

- **Data Modality:** Identify the format of the input data. (e.g., Structured system logs, Network traffic, Grayscale images, Natural Language/Text).
- **Specific Preprocessing Techniques:** List the exact steps taken to clean or transform the raw data. Look for: *Normalization, SMOTE (oversampling), Scaling, Tokenization, Anonymization, or Data Augmentation.*
- **Feature Selection / Dimensionality Reduction:** Identify specific algorithms used to reduce the number of variables. Look for: *PCA, t-SNE, RFE (Recursive Feature Elimination), LDA, or Information Gain.*
- **Features (N):** Extract the total number of features (with their names) used for training. If the model uses a "latent space" or "embeddings," specify the dimensionality (e.g., "128-D vector").

Output Format: Provide the results in a single-row table with these columns: | Authors (year) | Data Modality | Specific Preprocessing Techniques | Feature Selection / Dimensionality Reduction | Features (N) |

Constraint: Use ONLY information found in this paper. If a specific technique or the number of features is not mentioned, write "Not Specified."

	Paper ID	Data Modality	Specific Preprocessing Techniques	Feature Selection / Dimensionality Reduction	Features (N) and Names
1	Adun (2023)	Structured Logs	Redundancy removal; data cleaning.	Manual expunging of irrelevant features.	4
2	Alabdulkareem (2022)	Multi-source Logs	Refinement, Aggregation, and One-hot Encoding.	Behavioral feature extraction (e.g., after-hours activity).	Not Spec.
3	Ahmadi (2025)	Behavioral Logs	Data anonymization; IP reputation scoring. (Graph-based representation)	Recursive Feature Elimination (RFE).	Not Spec.
4	Ahmed (2025)	Log-based (Windows Event Logs, Custom Log File)	Data collection, general preprocessing (implied by "configuring the decoders" and "build the baselines") , baselines established days) , anonymization (user names and IPs changed)	Relevant features for each use case based on domain Expertise (manual)	Not Specified
5	Al Hammadi et al. (2021)	EEG brainwave signals	Signals preprocessed (to remove noise from physical movements and eye blinking)	Not Specified	Not Specified
6	Ali et al. (2025)	Natural Language/Text (narrative text data)	Removed entries with missing critical information, truncation, chunking, under-sampling negligible threat cases, over-sampling medium threat cases	Not Specified	Not Specified
7	ALmihqani 2021	Structured system logs	ADASYN (adaptive synthetic sampling approach)	DNN feature extraction	Not Specified
8	Almusawi 2024	Structured system logs	SMOTE for data balancing	Manual feature selection aggregations tools in NoSQL code	4 (After END, Blocked site Count activity, No_of _emails)
9	AL-Mihqani 2022 A new intelligent multilayer	Structured system logs	Not Specified	Manual feature selection	8 feature groups (activity , day, time, pc, user ID, User role, Unit, Depart)
10	Al-Shehari et al. (2024) comparative analysis	Insiders' Activity Logs (Login/Off, USB, files, etc.), user activity data observations	Files combining, Data cleaning (handling missing values: mean for continuous, mode for categorical), Feature extraction, Min-max scaling, One-hot encoding, Train-test split	Feature extraction (relevant features) Use One hot encoding	5 feature groups Hour of day, day of week, user, PC , Action
11	Al-Shehari alsawil. (2023) Random sma	Logs (logins, emails, file transfers, USB drives), user activity log data	Data preparation, Feature extraction, Encoding, Scaling, Random under-sampling, Random over-sampling, Hybrid resampling	Feature extraction manually based on observation (relevant features, most significant attributes) focus on data leakage	5 Feature Groups
12	Al-Shehari et al. (2024)	Human behavior-based IT activity data, log data, sequences of user activity	Data preparation, Feature selection, Encoding labels, StandardScaler, SMOTE, Borderline-SMOTE, ADASYN	Feature selection (relevant features)	Not Specified
13	Al-Shehari et al. (2023) Looks at data contamination	Insider data leakage dataset, log data	Data preprocessing, Feature extraction, Data refinement, Aggregation, Encoding Encode categorical variables into numerical representations. • Normalize or standardize features	Feature extraction, Removal of unrelated features (focus on data leakage)	(logon, logoff, device, http), Timestamps, User IDs, Events (e.g., login/logoff, USB connect/disconnect, http visits, etc.) and Targets (normal or malicious)

14	Alshehri (2022) DL	Structured system logs (audit data/activity streams)	Parsing log lines into multivariate time series vectors; Normalization using the median and inter-quartile range; Sequence activity embedding	Not Specified (Uses Deep Learning/RNN for automatic feature extraction)	256-D embedding vector (optimal batch-size)
15	Amiri-Zarandi (2023)	Structured system logs (behavior logs), organization structure, and psychometric data	Feature extraction from daily behavior statistics; Data splitting into 11 local participant units	Autoencoder "latent representation" (compression)	892 features; 221-D latent space
16	Al-Shehari (2021)	Structured system logs (logon, file, HTTP, email, and device logs)	Label Encoding; One-Hot Encoding; Feature Scaling (Standard Scalar); SMOTE (oversampling); Unix Epoch Time conversion	Not Specified (Manual feature extraction based on threat scenario relevance)	5,481 features (post-encoding matrix columns)
17	Amuda (2022)	Structured cyberspace system logs (PC access, file access, and device logs)	Categorical attribute encoding; Min-Max normalization (scaling to 0–1 range); Standard scaler techniques	Mutual Information Gain technique	9 features
18	Anju (2024)	Multivariate time-series data (User activity logs: Logon, Device, HTTP, Email, File) and user information	One-Hot Encoding; Z-score normalization (outlier handling); Mean value imputation (missing values); Sequence padding and truncation	Not Specified (Uses Stacked CNN for temporal representation learning)	8 features out of 39 fields
19	Anakath (2022)	User interaction behavior (Mouse movements/clicks and keystrokes)	Feature extraction of mouse distance (N-grams and Jaccard coefficient), speed, keystroke duration, and latency	Not Specified (Deep Belief Network handles high-dimensional parameters)	Not Specified
20	Asha S (2023)	Structured system logs (Logon, Device, and HTTP)	Data aggregation (merging logs into a single table); Categorical encoding (numerical transformation); Sampling (Nearmiss-2 undersampling; SMOTE, ADASYN oversampling)	Not Specified	6 features: Insider Threat, Vector, Date, User, Pc, and Activity
21	Cai X (2024)	User activity sequences (logon, web visits, file operations)	Data aggregation; Session splitting; Repetitive HTTP filtering; Numeric token mapping (type ID × 24 + hour slot)	Not Specified	128-D vector (size of hidden layer and embeddings)
22	Dong J (2025)	Employee behavioral logs (logon, file, HTTP, email, device)	Context fusion (aggregating multiscale temporal context); Curriculum learning (progressive exposure to complex patterns); Dilated causal convolutions	Not Specified	Not Specified (Uses context-infused temporal embeddings)
23	Eshmawi (2025)	Structured, tabular event logs (logon, email, file, HTTP, device)	Data aggregation; Numerical mapping; Normalization (zero mean and unit variance); Oversampling (SMOTE); Sampling (to handle massive record counts)	Not Specified (Mentions extracting key features and discarding less essential ones but identifies no specific algorithm)	Not Specified (Extracts frequency elements, statistical features, and user information; examples include Email ID, size, PC, and activity)
24	Feng (2025)	User activity logs (Logon, File, HTTP, Email, Device) and LDAP organizational data	Data fusion; Numerical mapping; Chronological data partitioning (Extrapolation mode); Oversampling (SMOTE/Random) and Undersampling	Not Specified (Notes that Random Forest automatically identifies features with significant impact)	Not Specified (Lists features such as Id, User, Date, Filename, URL, CC/BCC, and Attachment)

25	Ferraro (2025)	Network-centric Zeek logs and behavior-rich logs	Log grouping (sharing same ID); Log partitioning (by category); JSON schema enforcement (for output consistency)	Not Specified	Not Specified (Analyzes fields including ts, uid, id.orig_h, request_type, client, and service via LLaMA-3.1 8B context)
26	Gayathri (2025) [Adversarial]	Tabular data (numerical, binary, and categorical)	Data augmentation (GAN-based: CGAN, ACGAN, CWGAN-GP); Oversampling (SMOTE, ROS); Self-normalization (SeLU activation)	SHAP (Shapley Additive Explanations) and Tree SHAP used for global model interpretation and identifying critical feature subspaces	Not Specified (Dimensionality of latent space is equal to the number of features; logs include Logon, Email, Web, File, and Device)
27	Gayathri (2025) [Cloud]	Behavioral log files; Sequential and standalone activities	Multivariate Imputation by Chained Equations (MICE); Golf Optimization Algorithm (GOA) for parameter tuning; One-hot encoding	Domain extraction (manual selection); Long Short-Term Memory-Autoencoder (LS-AE) for automatic feature extraction	Not Specified (Includes features such as duration, protocol_type, service, flag, src_bytes, dst_bytes, land, urgent, and hot)
28	Gayathri (2024) [Hybrid]	Heterogeneous data (structured/tabular) from resource logs	SPCAGAN augmentation; Sentiment analysis of text (email/web); Random Over Sampling (ROS); SMOTE	Linear manifold learning (SPCA); Feature selection (eliminating correlated/redundant features); t-SNE (for visualization) Manual and Automated feature selection	47 features (Logon/off: 10, Email: 12, Web: 8, File: 5, Device: 5, Psychometric: 5, Sentiment, and holiday details)
29	Gonzales (2025)	User action sequences (e.g., Create, Read, Update, Delete, Execute)	Label-encoding (actions to numerical); Padding (standardizing sequence length); Sub pattern pruning	Contiguous pattern mining (Stage 1); length-based pattern sorting for redundancy reduction	Not Specified (Models actions as bigrams and trigrams)
30	Gupta (2024)	Behavioral logs; Historical and live user details	Normalization/scaling (range); Missing data cleaning; Data encoding; Text-to-numeric conversion	Not Specified	13 features: Profession, number of requests, type of requests, data limit, user type, historical data, leakage record, how many times data leaked, how frequently asking data, data retention, leak ratio, and leak channel
31	Hafizu Rhman (2022)	User activity logs (CERT r4.2)	Dataset partitioning (Training/Testing/Validation in various ratios such as 80:15:5)	NARX-based trait selection (identifying dimensions related to detections parameters)	8 input features: Access, Motivation, Action, Intent, +Action, Method, Level of Trust, and Insider Knowledge
32	Haq (2022)	Natural Language/Text (E-mail messages) and Structured (Financial data).	Email transformation (headers, body, name extraction); Null handling (NumPy.nan for empty subjects, dropped missing rows); Label encoding; One-hot encoding; Tokenization; Padding (150 word max length).	Transfer Learning (using pre-trained Word2Vec and GloVe) to handle data sparsity; Feature selection via feature format function; GridSearchCV for parameter optimization.	20 financial/metadata features (e.g., salary, bonus, to_messages, expenses); 300-D word embeddings.
33	He (2022)	Natural Language (Email body/subject) and Structured Logs (User behavior sequences).	Word segmentation; Vectorization; MDFTS (Multi-Domain Feature Time Series) algorithm for heterogeneous logs; JSON formatting for virtual role modeling.	Bi-LSTM to obtain long-range dependent features; Attention mechanism to capture key sequence information; XGBoost to strengthen network structure.	Not Specified (Uses concatenated output vectors from LSTM/Bi-LSTM and numerical statistical features).

34	He (2024)	Natural Language (Email) and Structured System Logs (Logon, Device, File, Email, Web, Psychometric).	Format conversion; Email parsing; MDFTS algorithm (transforming raw logs into ActivityList, ContentList, and CountsList); Tokenization; Padding.	MDFTS algorithm for multi-domain handling; Attention mechanism to capture "attack words" in implicit vectors; L2 Regularization and Dropout (0.4 ratio) to reduce complexity and overfitting.	L-XGB: 18 features (e.g., attachment counts, URL symbols, HTML forms); BiLA-ITD: 626 total features (210-D text input and 416-D numeric input)
35	Huang (2025)	Network traffic and user behavior records	ColumnTransformer for feature encoding; One-hot encoding for labels; Resampling of rare class samples; 1-D Batch Normalization (scaling and shifting)	Random Forest for dimensionality reduction to reduce noise and capture principal features	Not Specified
36	Jaiswal (2024)	Structured system logs (HTTP, file, device, logon, and email)	Day-long sequence transformation; Activity encoding; Null handling and missing value imputation; Min-max scaling ; Hybrid/Over/Undersampling (SMOTEENN, SMOTE-TL, ADASYN, SVMSMOTE, ENN)	Not Specified	30 attributes (excluding labels). Names include: Leaks, Job search, Keylogger, Restricted uploads, After hours, File downloads, Duration, Time window, Number of directories, Depth, Weekends/After-hours connected device, Login per day, Weekends logins, First/Last login difference, File open/write, Mailed outside, Attachment, Send/Receive
37	Janjua (2021)	Natural Language/Text (Emails)	Missing value imputation; Stop word removal; Stemming; TF-IDF vectorization	Latent Semantic Analysis (LSA) (utilized after TF-IDF matrix generation)	Not Specified
38	Kamatchi (2025)	Structured IoT behavior data (access patterns, login times, and data transmission metrics)	Normalization (scaling login times and transmission bytes to a common range) and Mean Imputation for handling missing values.	OPTICS-CR clustering (used to identify influential nodes and refine centroids to reduce communication overhead).	59 features extracted from network traffic, system resources, and security event logs (specific to X-IloTID).
39	Kong (2025)	Structured device-level logs (authentication, file operations, system calls, and sensor events)	Data cleaning (filtering heartbeats and duplicate logs); Timestamp parsing; Behavior aggregation; Session segmentation; and Symbolic Tagging of redundant fragments.	DPI-ITD Tagging mechanism (guided by Tagging Scores to filter uninformative components and reduce sequence length).	20-D embedding vector (generated via adaptive GloVe-based embedding with a co-occurrence window of 5).
40	Kotb (2025)	Structured event logs (logon, email, device, file, and HTTP) and Psychometric scores	Data cleaning (excluding null/string values); Feature scaling (standardization); and On-the-fly weighted random sampling to address class imbalance.	Deep Feature Synthesis (DFS) (automatically generates relational features using Gaussian Copula and mathematical aggregation/transform primitives).	430 features. Names include: "Big Five" personality traits (OCEAN), total email attachment size, and primitives like "time since last," "entropy," and "percent unique"
41	Lavanya (2024)	Structured system logs (logon, files, emails, devices, and HTTP)	Ordinal encoding; mining timestamps (hour, day, month, year); Robust scaling; and Data Augmentation using Enhanced Bidirectional Generative Adversarial Network (EBiGAN)	Improved Principal Component Analysis (IPCA) and outlier estimators of k-means clustering	Not Specified (Attributes include id, userid, timestamps, PC number, and activity logs across various layers)

42	Lavanya (2025)	Multimodal/Structured network log records (file, device, http, email, and user logs)	Label encoding ; timestamp-based user activity filtering and aggregation; Z-score normalization ; and Data Augmentation using Adabelief Wasserstein Generative Adversarial Network (ABWGAN-EHI)	Chebyshev Graph Laplacian Eigenmaps (CGLE) solver and IS-DBSCAN clustering	Not Specified (Ablation study identifies that using v=20 eigenvectors achieved improved clustering accuracy)
43	Le & Zincir-Heywood (2021)	Structured activity logs (log on/off, e-mail, Web, file, and thumb drive connect)	Extraction of Frequency and Statistical features ; Temporal representations including Concatenation, Percentile, Mean difference, and Median difference	Not Specified (The study notes that models like Autoencoder use bottleneck layers and LODA uses sparse random projections for internal dimensionality reduction)	132 features (Extracted for a comparative study utilizing the time series feature extraction library tsfel)
44	Li et al. (2024)	Behavioral data (logon/logoff, email, device, file, and HTTP) transformed into graph structures	Dataset modification for feature extraction; identification of missing user-to-user/PC relationships; structural reconstruction via learnable stochastic augmentation	Graph-based Auto-Encoder (GAE) used to transform original features into low-dimension representations	128-D vector (nhid1) and 32-D vector (nhid2) for latent embeddings
45	Li et al. (2023)	Heterogeneous graphs converted from user log entries (emails, web, files, devices)	NLP content collection; Weighted feature similarity mechanism to construct fused structures; L2-normalization	Not Specified (uses convolutional layers to extract high-level representations/embeddings)	128-D and 256-D hidden layers. Features include logon times, file types, email recipients/size, and specific web categories (e.g., job huntin
46	Liu et al. (2025)	Timestamped operation sequences (Logon, Email, Device, File, HTTP)	Behavioral tokenization (converting actions and time into numerical values $e=b*24+t$); contrastive learning sample pairing (positive/negative pairs); HNSW indexing for retrieval efficiency.	Not Specified (Mentions BERT captures "refined operational semantic representation" in a latent space).	Not Specified (Uses BERT base architecture and embeddings of dimension d).
47	Mehmood et al. (2023)	Structured activity logs (System, Email, External Device)	Data aggregation and normalization ; Outlier removal via neighbor averaging (for 'size' feature); Filling missing values using dataset patterns (for 'File Copy').	Manual removal of irrelevant features (e.g., 'employee', 'file tree'); Mentions feature extraction to lower redundant data.	14 features (including: id, date, user, pc, to, cc, bcc, from, size, attachment, activity, to_removable_device, from_removable_device) [125, Fig 10].
48	Medvedev et al. (2025)	Keystroke dynamics (1D time-series timing data)	GAFMAT (Gabor Filter Matrix Transformation) to create 2D images; Interpolation-based data fusion for password length standardization; Bilinear interpolation for resizing; Outlier removal from user profiles.	Not Specified (CNN layers in the Siamese network reduce high-dimensional representations into embeddings).	256-D embedding vector ; Input standardized to 40 x 40 pixel images
49	Mehnaz & Bertino (2021)	Block-level I/O traces (OS kernel space)	blkparse formatting of kernel events; Online translation of sector numbers to file names, block numbers, and inodes; Extracting episodes (serial/parallel) to discover task patterns.	Extraction of a minimal set of interesting features from block-level information; Use of Finite State Automata to scan event sequences only once.	7 features : user ID, file name, access type, access size, accessed segment, timestamp, and sequence number.
50	Mladenovic et al. (2024)	Natural Language / Text (Email, HTTP, and File content)	TF-IDF encoding to transform text into numerical vectors; Downsampling of normal activity logs (1:10 for email/HTTP; 100:1 for file) to address class imbalance.	Feature space limited via Sklearn TF-IDF vectorizer ; Metaheuristic optimization (HARFO/RFO) used for hyperparameter tuning to ensure generalization.	250 features (limited feature space for simulations).
51	Nasir et al. (2021)	Structured system logs (CMU CERT r4.2)	Aggregation of CSV files into a "Master file" ; Missing value replacement with estimated mean values; Integer encoding for categorical	Manual identification of relevant fields based on attack scenarios; Redundant feature removal (e.g., ID fields); Use of	9 features : logon_time, day, user_id, activity, pc, logoff_time, user_role,

			strings; Random oversampling to balance classes.	session-based flexible time windows instead of fixed windows.	functional_unit, and department.
52	Nikiforova et al. (2024)	Structured audit records (information system logs)	Creation of a matrix of user activities and sequential action pairs; Anonymization of user and action names (e.g., female names, generic action labels).	"Embedded" clustering approach combining sets of user actions with information about transitions between actions; Elbow method to determine optimal cluster counts.	Not Specified (Uses unique actions such as "generate_report," "login/exit," and "view_main_menu" to model graphs).
53	Pal et al. (2023)	Structured activity logs (CMU CERT v4.2, v5.2, v6.2)	Single-day activity sequence generation ; One-hot encoding for categorical embeddings; Equally-Weighted Random Sampling (EWRS) (undersampling normal and oversampling malicious instances).	Robust categorization of user activities into unique action-IDs based on time, system type, and environment.	168 action-IDs (v4.2) or 282 action-IDs (v5.2/v6.2); 100-D hidden feature vector .
54	Patel & Iyer (2025)	Cloud-based user behavior logs	Binary data encoding to quantify interactions; Extraction of parametric features from raw log files.	Frequent Pattern (FP) Growth algorithm used to analyze and assess behavior sequences for anomaly identification.	3 categories (Administer, Authentication, Comments); Input vector size is 784 .
55	Peccatiello et al. (2023)	Structured system logs (CMU CERT v4.2)	Session-based segmentation (logon to logoff); Implicit logoff insertion for missing events; Contamination (deliberate insertion of malicious samples into training sets).	Grid Search algorithm for initial hyperparameter optimization over labeled training sets.	7 features : diff_begin_first, device_count_out, device_count, count_dwn_exe_file, url_blocklist, unit_code, and tfidf_jobsites
56	Pennada et al. (2024)	Structured system logs (CERT dataset),	Verification for missing values/outliers; Random Oversampling , SMOTE , ADASYN , Random Undersampling, Cluster Centroids, and ENN ; Normalization ,.	Principal Component Analysis (PCA) ,.	830 features ; Target variable: "insider".
57	Pennada et al. (2025)	Structured activity logs (CERT dataset),	Cleaning (null/inconsistent values); StandardScaler ; IQR method for outlier removal; SMOTE (k=5), Random Undersampling, and ENN ,.	Latent feature extraction via Deep Autoencoders (DAE) and Variational Autoencoders (VAE),.	50 behavioral features ; 16-D latent space (from autoencoders),.
58	Perez-Miguel et al. (2025)	Event-level logs (File manipulation, command execution, network behavior)	Log normalization ; Harmonization with SPEDIA schema; Dataset cleaning and handling of null values,.	Manual removal of internal system data columns to retain study-relevant variables.	25 curated attributes (e.g., Agent_name, User, Scrip, Content, Filename, Command, Activity, Action),.
59	Qawasmeh & AlQahtani (2025)	Synthetic network/user records ,	Imputation (missing values replaced with 1); Outlier adjustment (frequency values clamped to 1); SMOTEENN (SMOTE + Edited Nearest Neighbors),.	Not Specified	22 features including PII (Name, Email, SSN) and 17 anomalous types (e.g., Login Attempts, Data Transfer, Secure Printing),-.
60	Randive et al. (2023)	User activity logs transformed into grayscale images ,	Min-Max Normalization ,; SMOTEENN ,; Bilinear interpolation (resizing to 64x64); Horizontal flip data augmentation,.	Not Specified	20 features (L1-L9 Logon, D1-D2 Device, E1-E5 Email, F1 File, H1-H3 HTTP),.
61	Rauf et al. (2021)	Unstructured log-feeds (Logon, web, devices).	Conversion of unstructured logs to structured activity vectors (<i>avij</i>); One Hot Encoding for activity attributes; mapping timestamps to cyclic temporal variables (0–24 Hrs).	Not Specified (utilizes a weighted distance metric for unsupervised learning).	Not Specified (Activity vectors representing sequences of malicious activities).
62	Roy & Chen (2024)	Host logs (action sequences) and	Log parsing into tokens ; Word vectorization via FastText (word2vec); One-hot encoding for attributes.	TF-IDF weighting for word vectors; n-hop neighbor sampling via Random Walk with Restart (RWR).	100-D embedding vector derived from 862 unique log fields.

		Psychological behavioral data.			
63	Bin Sarhan & Altwaijry (2023)	Relational system logs (CERT r4.2: logon, device, email, http, file).	Deep Feature Synthesis (DFS) for automated feature engineering; One-hot encoding ; StandardScaler ; SMOTE balancing.	Principal Component Analysis (PCA) (preserving 95% variance).	69,738 features (original) reduced to 553 features (post-PCA).
64	Senevirathna et al. (2025)	Structured behavioral logs (logon, device, psychometric) and CCTV video streams .	Imputation (backward/forward filling, mean/median), One-hot encoding , Min-Max normalization , Frame extraction , Pixel normalization , and Data augmentation .	Dimensionality reduction of multi-dimensional tensor outputs to uniform feature vectors for temporal modeling.	Not Specified (Includes login trends, psychometric traits via OCEAN model, and graph-based interactions).
65	Song et al. (2024)	Behavioral feature sequences (Implicitly encoded time and behavior tokens).	Behavior tokenization , temporal aggregation into chronological sequences , and maximum normalization to account for user-day rhythm fluctuations.	User-adaptive selection of sequence construction based on maximizing covariance of feature sequences.	129 distinct types of behavior tokens.
66	Tabassum et al. (2024)	Structured Electronic Health Records (EHR) / audit logs.	Cleaning , Missing value management (mean/mode), One-hot encoding , and Min-max scaling (standardizing to range 0–1).	Cross-correlation analysis to identify redundant features and relevant subsets.	90,385 distinct identifiers used after refinement. Primary fields include Date, Time, Device, User ID, Routine, Patient ID, Duration, and Discharge dates.
67	Tian et al. (2025) [ITDSTS]	Structured system logs (login, device, email, file, weblogs) formatted as text strings.	Min-Max scaling ; chronological integration via unique identifiers; One-hot encoding .	TF-IDF algorithm (text topic analysis); Isolation Forest (outlier detection); K-means (frequency segregation).	Not Specified (Uses 14 categories in Table 1, dynamically updated with textual topics and high-frequency behaviors).
68	Tian Z et al. (2024) [DSDLITD]	Network/Data traffic (Event log entries: Device, Email, File, HTTP, Logon).	Data cleansing; sampling; transforming raw logs into a vectorized matrix ; one-hot to embedding vector mapping.	Multi-head attention mechanism used to learn and categorize features (basic, content-based, and traffic-based) before concatenation.	Not Specified (Features grouped into three types: Basic, content-based, and traffic-based).
69	Villarreal-Vasquez et al. (2023)	Structured system logs (EDR high-dimensional activity events).	Enumeration of activities to form a "Vocabulary of Events"; Feature transformation $FT(\bullet)$ to control granularity (low vs. high).	Feature Selection (F) applied to 6 categorical and continuous features (Actor, Event type, Action, Target, Network, User).	175 categorical symbols (Vocabulary size); 16-D embedding vector.
70	Wall & Agrafiotis (2021)	Structured system logs (device.csv, email.csv, file.csv, http.csv, logon.csv).	Daily temporal partitioning; Mean and Standard Deviation mapping (into 7 finite intervals); Laplace smoothing .	Pearson's chi-squared test (structural learning to determine pairwise dependencies and edges).	Approximately 70 features (including 'counter' types for device, email, file, and web actions).
71	Wang & El Saddik (2023)	Structured system logs (logon, device, file, email, http) and user profile data (HR/psychometric).	Data cleansing, imputation of missing values, and grouping/sorting by user and day to form chronological sequences. Data augmentation using BERT (contextual word embedding insert/substitution) and GPT-2 (sentence-level generation). Tomek Links for down-sampling normal data.	Not Specified.	Numeric tokens identifying user behavior based on activity type, occurrence hour, work hour status, and PC ownership. Behaviors are transformed into dense word embeddings and padded for uniform length.
72	Wang Zhi et al. (2024) FedITD	Structured system logs (login, device, file, email, HTTP) and user profile data (identities, psychometric traits).	Data cleaning, Anonymization/Obfuscation of sensitive information, and chronological sorting into daily behavior sequences. NLP data augmentation using pre-trained LLMs for contextual word insertion.	Not Specified.	Numeric behavior tokens representing activity types and temporal information. Tokens are converted into word

					embeddings via a custom vocabulary.
73	Wang Jiarong et al. (2023) Deep Cluster	Multi-source behavioral event sequences (logs: host login, file, email, web, USB).	Record-to-event transformation; discretization of action counts into action magnitude levels . Fixed-length sequence segmentation (length $l=8$).	k-means clustering is used to initialize centroids in the feature space during the pre-training stage.	5 entities per event: Time, Host, User, Action, and Action Magnitude. Uses 10-D embedding vectors for event entities.
74	Wei Yichen et al. (2021)	Behavior sequences from raw logs (device, email, file, http, logon) and metadata (role, psychometric scores).	Log line aggregation on a per-user/day basis; One-hot encoding of behaviors. Unsupervised data purification using Cascaded Autoencoders (CAEs) to drop samples with high reconstruction error.	Dimension reduction via the non-linear latent representation in the "code layer" (middle layer) of the final Autoencoder.	164 behaviors encoded numerically. BiLSTM extracts abstract features from sequences, which are then concatenated with user meta-data (role, team, supervisor, psychometric score).
75	Wei Zhiyuan et al. (2024)	Unstructured logs (system, user, and network activities).	Time-interleaving-based aggregation and user-based segregation. One-Hot Encoding to convert unstructured data to structured formats.	Information Gain (entropy-based labeling) and Variance threshold-based filtering to exclude low-variance/irrelevant variables. Local Outlier Factor (LOF) for parametric filtering.	Activity vectors representing structured user data rows. Total number (N) is Not Specified .
76	Wen et al. (2023)	Natural Language/Text (Email content) and metadata,.	Data cleaning (removing non-Enron members/empty content); filtering by rational timestamps; weekly snapshot aggregation; NLTK (VADER) sentiment analysis to compute normalized composite scores,.	Singular Value Decomposition (SVD) to decompose spatial-temporal sequences into base networks; eigenvector centrality for node ranking,.	103×2076 matrix representing weekly sentiment snapshots of directed communication paths,.
77	Xiao Junchao et al. (2023)	Structured system logs (Numerical and Text fields),.	Factorization for activity fields; Word2Vec for text fields (segmentation of file addresses/URLs); Min-Max Normalization ; transformation into Visibility Graphs (VG) and Horizontal Visibility Graphs (HVG),.	Arithmetic average aggregation of word vectors; linear transformations to reduce dimension from 112 to 57, then 6; t-SNE for visualization,.	112 node features (derived from file tree, activity type, and 100-word content embeddings) per log record,.
78	Xiao Fengrui et al. (2025)	Structured interaction logs (Authentication, NetFlow, DNS),.	Uniform log format creation (padding with 0); Ordinal encoding for discrete text; Word2Vec for continuous text; sliding time windows; Negative sampling ,.	Hierarchical contextual attention (HCA) ; Edge Graph Attention (EGAT) for feature representation; latent node embedding dimension d ,.	K distinct fields cascaded into a single feature vector F_{ij} ; hidden state dimension of 512 ,.
79	Xiao Haitao et al. (2024)	Multisource behavior logs (Audit logs + User profiles),.	Multisource log integration; time granularity division (daily segments); Min-Max Normalization ; behavior mapping into Action IDs ,.	Statistical analysis module reshapes vectors into 18×18 matrices; flatten layers for mapping to 1-D vectors,.	20 action factors and 8 PC-time offsets combined to form identifiers; max sequence length of 320 ,.
80	Ye Xiaoyun et al. (2025)	Numerical/Structured logs transformed into Grayscale images ,.	Category encoding; weekly aggregation; Z-score normalization ; conversion to images via DeepInsight (mapping similar features closer together),.	Principal Component Analysis (PCA) ; t-SNE used to reduce feature embeddings to 2D coordinates for pixel discretization,.	1098 features extracted from CERT v5.2, mapped onto a fixed-size 64×64 pixel grid ,.
81	Yildirim & Anarim (2022)	Behavioral biometrics (Mouse usage patterns),.	Action segmentation (lower-level events to higher-level MM, DA, DC actions); silence thresholding; Lomb-Scargle periodogram for unevenly sampled signals,.	Feature masking via Gradient Boosting (XGBoost); Pruning (max tree depth); grid search for optimal parameters,.	105 movement features (54 time-domain, 48 frequency-domain, and 3 scalar features) plus 6 features for double-click actions,.
82	Zhu et al. (2024)	Multisource ERP system logs ,.	Numerical behavior coding; variable time-window session creation (login to logout);	Encoder (TCN + LSTM) transforms high-dimensional vectors to low-dimensional	me (event frequency dimension) and mt (time

			Event frequency counting and time recording,.	latent features (z) ; Adversarial learning aligns latent distribution with a Gaussian prior,.	feature dimension) totaling all defined behavior codes,.
--	--	--	--	--	--

The is the table used in the SLR but with references as numbers [x]

Feature Engineering Approach	Description	Count (N)	Studies (Grouped by Specific Technique)
Traditional & Statistical	Converts logs into fixed vectors using One-Hot/Label Encoding , Min-Max/Z-Score Scaling , and Sampling (SMOTE, ADASYN). Includes Deep Feature Synthesis (DFS) for automated statistical aggregation.	43	Automated (DFS) : Bin Sarhan (2023) [32], Kotb (2025)[31] Manual / Statistical : Adun (2023) [35], Ahmed (2025) [36], Alabdulkareem (2022), ALmihqani (2021), Almusawi (2024), AL-Mihqani (2022), Al-Shehari & Alsawail (2021), Al-Shehari (2023), Al-Shehari (2024), Al-Shehari & Alsawail (2023), Amiri-Zarandi (2023), Amuda (2022), Anakath (2022), Asha S (2023), Eshmawi (2026), Feng (2025), Gayathri (2025, Cloud), Gayathri (2025, Adv), Gupta (2024), Huang (2025), Kamatchi (2025), Lavanya (2024), Le & Zincir-Heywood (2021), Mehmood (2023), Nasir (2021), Nikiforova (2024), Patel & Iyer (2025), Peccatiello (2023), Pennada (2024), Pennada (2025), Perez-Miguel (2025), Qawasmeh (2025), Rauf (2021), Senevirathna (2025), Tabassum (2024), Tian Z (2024), Wall & Agrafiotis (2021), Wei Zhiyuan (2024), Xiao Haitao (2024), Yildirim (2022)
NLP-Based Embeddings	Treats logs or content as text. Uses TF-IDF or Sentiment Analysis for keyword weighting, and Embeddings (Word2Vec, BERT, RoBERTa, LLaMA) to capture semantic context.	14	TF-IDF / Sentiment : Janjua (2021), Mladenovic (2024), Tian T (2025), Wen (2023) Embeddings / LLM : Ali (2025), Ferraro (2025), Gayathri (2024 Hybrid), Haq (2022), He (2022), He (2024), Kong (2025), Liu (2025), Wang & El Saddik (2023), Wang Zhi (2024)
Sequence & Time-Series	Aggregates actions into time-ordered sequences (\$t_1\$ to \$t_2\$) using Sliding Windows and Event Vocabularies to preserve order for RNNs/LSTMs.	14	Temporal Sequences : Alshehri (2022), Anju (2024), Cai X (2024), Dong J (2025), Gonzales (2025), Hafizu Rhman (2022), Jaiswal (2024), Mehnaz (2021), Pal (2023), Song (2024), Villarreal-Vasquez (2023), Wang Jiarong (2023), Wei Yichen (2021), Zhu (2024)
Graph Construction	Models entities (User, File) as nodes and actions as edges to capture structural relationships via Heterogeneous Graphs or Adjacency Matrices .	8	Graph Structures : Ahmadi (2025), Lavanya (2025), Li et al. (2023), Li et al. (2024, GMFITD), Li et al. (2024, DD-GCN), Roy & Chen (2024), Xiao Junchao (2023), Xiao Fengrui (2025)
Image Transformation	Converts numerical log data or sequences into Grayscale Images or matrices (using DeepInsight, GAFMAT) to utilize CNNs for visual pattern recognition.	4	Visual / Pixel Mapping : Al Hammadi (2021), Medvedev (2025), Randive (2023), Ye Xiaoyun (2025)

We used Google Gemini and manual cross-referencing to confirm the references are accurate as numbered in the SLR

Seq.	Feature Engineering Category	Study Name (Author & Year)	Final Ref #
1	Automated (DFS)	Bin Sarhan (2023)	[32]
2	Automated (DFS)	Kotb (2025)	[31]
3	Manual/Statistical	Adun (2023)	[35]
4	Manual/Statistical	Ahmed (2025)	[36]

5	Manual/Statistical	Alabdulkareem (2022)	[37]
6	Manual/Statistical	ALmihqani (2021)	[38]
7	Manual/Statistical	Almusawi (2024)	[64]
8	Manual/Statistical	AL-Mihqani (2022)	[63]
9	Manual/Statistical	Al-Shehari & Alsowail (2021)	[29]
10	Manual/Statistical	Al-Shehari (2023)	[62]
11	Manual/Statistical	Al-Shehari (2024, Isolation Forest)	[13]
12	Manual/Statistical	Al-Shehari (2024, CNN)	[61]
13	Manual/Statistical	Al-Shehari & Alsawail (2023)	[39]
14	Manual/Statistical	Amiri-Zarandi (2023)	[40]
15	Manual/Statistical	Amuda (2022)	[41]
16	Manual/Statistical	Anakath (2022)	[42]
17	Manual/Statistical	Asha S (2023)	[43]
18	Manual/Statistical	Eshmawi (2026)	[44]
19	Manual/Statistical	Feng (2025)	[23]
20	Manual/Statistical	Gayathri (2025, Cloud)	[9]
21	Manual/Statistical	Gayathri (2025, Adv)	[16]
22	Manual/Statistical	Gupta (2024)	[45]
23	Manual/Statistical	Huang (2025)	[46]
24	Manual/Statistical	Kamatchi (2025)	[47]
25	Manual/Statistical	Lavanya (2024)	[48]
26	Manual/Statistical	Le & Zincir-Heywood (2021)	[24]
27	Manual/Statistical	Mehmood (2023)	[30]
28	Manual/Statistical	Nasir (2021)	[49]
29	Manual/Statistical	Nikiforova (2024)	[50]
30	Manual/Statistical	Patel & Iyer (2025)	[51]
31	Manual/Statistical	Peccatiello (2023)	[10]
32	Manual/Statistical	Pennada (2024)	[52]
33	Manual/Statistical	Pennada (2025)	[14]
34	Manual/Statistical	Perez-Miguel (2025)	[53]
35	Manual/Statistical	Qawasmeh (2025)	[54]

36	Manual/Statistical	Rauf (2021)	[28]
37	Manual/Statistical	Senevirathna (2025)	[55]
38	Manual/Statistical	Tabassum (2024)	[56]
39	Manual/Statistical	Tian Z (2024)	[57]
40	Manual/Statistical	Wall & Agrafiotis (2021)	[58]
41	Manual/Statistical	Wei Zhiyuan (2024)	[22]
42	Manual/Statistical	Xiao Haitao (2024)	[59]
43	Manual/Statistical	Yildirim (2022)	[60]
44	NLP: TF-IDF / Sentiment	Janjua (2021)	[65]
45	NLP: TF-IDF / Sentiment	Mladenovic (2024)	[66]
46	NLP: TF-IDF / Sentiment	Tian T (2025)	[67]
47	NLP: TF-IDF / Sentiment	Wen (2023)	[68]
48	NLP: Embeddings / LLM	Ali (2025)	[26]
49	NLP: Embeddings / LLM	Ferraro (2025)	[8]
50	NLP: Embeddings / LLM	Gayathri (2024 Hybrid)	[69]
51	NLP: Embeddings / LLM	Haq (2022)	[70]
52	NLP: Embeddings / LLM	He (2022)	[71]
53	NLP: Embeddings / LLM	He (2024)	[25]
54	NLP: Embeddings / LLM	Kong (2025)	[27]
55	NLP: Embeddings / LLM	Liu (2025)	[72]
56	NLP: Embeddings / LLM	Wang & El Saddik (2023)	[73]
57	NLP: Embeddings / LLM	Wang Zhi (2024)	[74]
58	Temporal Sequence	Alshehri (2022)	[75]
59	Temporal Sequence	Anju (2024)	[76]
60	Temporal Sequence	Cai X (2024)	[77]
61	Temporal Sequence	Dong J (2025)	[78]
62	Temporal Sequence	Gonzales (2025)	[79]
63	Temporal Sequence	Hafizu Rhman (2022)	[80]
64	Temporal Sequence	Jaiswal (2024)	[81]
65	Temporal Sequence	Mehnaz (2021)	[82]
66	Temporal Sequence	Pal (2023)	[83]

67	Temporal Sequence	Song (2024)	[84]
68	Temporal Sequence	Villarreal-Vasquez (2023)	[85]
69	Temporal Sequence	Wang Jiarong (2023)	[86]
70	Temporal Sequence	Wei Yichen (2021)	[87]
71	Temporal Sequence	Zhu (2024)	[15]
72	Graph Structure	Ahmadi (2025)	[88]
73	Graph Structure	Lavanya (2025)	[34]
74	Graph Structure	Li et al. (2023, DD-GCN)	[33]
75	Graph Structure	Li et al. (2024, GMFITD)	[89]
76	Graph Structure	Roy & Chen (2024)	[90]
77	Graph Structure	Xiao Junchao (2023)	[91]
78	Graph Structure	Xiao Fengrui (2025)	[92]
79	Image Transformation	Al Hammadi (2021)	[93]
80	Image Transformation	Medvedev (2025)	[94]
81	Image Transformation	Randive (2023)	[95]
82	Image Transformation	Ye Xiaoyun (2025)	[96]