

USER BEHAVIOR ANALYTICS FOR DETECTING AI-DRIVEN SOCIAL ENGINEERING ATTACKS

Nurmukhammad Nuriddinov

INHA University in Tashkent, Tashkent, Uzbekistan

Undergraduate student

n.nuriddinov@student.inha.uz

Abstract

Social engineering remains a persistent security problem because it exploits human decision-making instead of only software flaws. Recent evidence shows that identity misuse and human-driven mistakes continue to be major breach pathways: Verizon's 2025 Data Breach Investigations Report analyzed more than 22,000 incidents and 12,195 confirmed breaches, while credential abuse remained a leading initial attack vector at 22% [1]. This paper examines how behavioral security analytics, user and entity behavior analytics (UEBA), deep learning, deep evidential clustering (DEC), explainable AI (XAI), and federated learning (FL) can defend against social engineering and insider threats. This work synthesizes published results and compares methods, datasets, formulas, advantages, limitations, and deployment gaps. The review finds that AI-driven behavioral analytics can reduce false positives, detect subtle post-login behavior changes, and provide analyst explanations; however, real-world deployment is limited by dataset imbalance, privacy constraints, cold-start baselines, model poisoning, and the difficulty of inferring human intent from logs alone. The paper concludes that the strongest defense is not fully automated surveillance, but a human-centered system that combines contextual behavior modeling, privacy-preserving learning, explainable alerts, and continuous security awareness.

Keywords - social engineering, behavioral analytics, UEBA, insider threat detection, deep evidential clustering, federated learning, explainable AI, cybersecurity.

1. Introduction

Cybersecurity is often discussed as a technical contest between attackers and defenders, but social engineering shows that the human layer is often the easiest path into a secure system. Social engineering attacks use trust, authority, urgency, fear, curiosity, or routine work habits to influence a user into clicking a malicious link, approving an authentication request, sharing credentials, or moving data outside policy [2], [3]. Wang et al. describe social engineering as an interaction among effect mechanisms, human vulnerabilities, and attack methods, meaning that the attacker's success depends on psychology as much as on technology [2].

The growth of cloud work, SaaS tools, remote access, and AI-generated communication has made this problem more serious. A phishing email can now be grammatically polished, personally targeted, and timed to match a user's role. Once the user authenticates, many traditional defenses see a valid identity instead of an attacker. This is why behavior-aware defense is necessary. UEBA systems monitor patterns such as login time, device use, application access, file movement, email behavior, and network destinations to estimate whether an authenticated user is behaving like their usual self [4], [5].

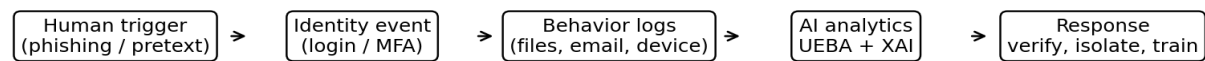


Figure 1. Review flow: linking social engineering signals to behavior-based AI defense.

The aim of this review is to connect two research streams that are often treated separately: human vulnerability in social engineering and AI-driven behavioral analytics for defense. The paper edits and refocuses the original review into a compact, referenced, review-style manuscript. It avoids presenting the results section as if a new experiment was conducted; instead, it interprets reported findings from the literature and explains why, where, and how those methods are useful.

2. Problem Statement and Research Gaps

The core problem is that many organizations still depend on static controls - passwords, signatures, fixed thresholds, and rule-based alerts - while social engineering and insider threats operate through legitimate identities and context-dependent behavior. If an employee is tricked into approving a login or if an insider abuses normal privileges, the session may look valid to perimeter controls. Verizon's 2025 report emphasizes the importance of credential abuse and third-party involvement in breaches, while IBM reports that the global average cost of a data breach in 2025 was USD 4.4 million [1], [6]. These figures support the need for earlier detection after authentication, not only stronger perimeter blocking.

Several research gaps remain clear in the literature. First, many models detect anomalies but do not explain whether the anomaly is credential theft, privilege abuse, data leakage, or a stressed employee doing unusual but legitimate work [7]. Second, insider threat datasets are extremely imbalanced because malicious actions are rare compared with normal activity, which can bias machine learning models toward predicting “normal” behavior [4], [8]. Third, uncertainty is often missing: deterministic models may output a label without saying how confident they are, which increases alert fatigue [9]. Fourth, behavioral data is sensitive because it may include work patterns, keystroke rhythm, communication metadata, and location; centralized training can create privacy and compliance risks [10], [11]. Finally, many studies are evaluated on benchmark or synthetic datasets such as CERT, so the reported accuracy may not transfer directly to noisy, real-world organizations with different cultures, schedules, and roles [4], [7].

3. Literature Review

Social engineering research provides the psychological foundation. Wang et al. identify persuasion, social influence, trust, deception, emotion, and decision-making as mechanisms that explain why users may comply with a malicious request [2]. Pujari and Hussain also emphasize the human factor in phishing and social engineering, showing that training and technical controls must be combined because users respond differently under pressure [3].

Behavioral analytics research provides the technical defense. Nasir et al. use deep learning to detect insider threats from user behavior and show the value of modeling user activity sequences rather than isolated events [4]. Villarreal-Vasquez et al. propose LADOHD, an LSTM-based anomaly detection framework, to identify long-duration insider attack sequences that are mixed with normal actions [12]. Tian et al. move further by targeting specific threat scenarios - privilege abuse, identity theft, and data leakage - with behavior sequences and multi-head attention, reporting improvements over baseline deep learning methods [7].

Recent studies add three important improvements: uncertainty, privacy, and explanation. Ali et al. propose DEC for real-time insider threat detection and report 94.7% average accuracy with a 38% false-positive reduction on CERT and TWOS compared with traditional clustering [9]. Alazab et al. review FL for cybersecurity and show how decentralized learning can support authentication, privacy, trust management, and attack detection without sharing raw data [10]. Google's Gboard deployment is a practical example of FL and differential privacy: the keyboard improves language models while learning from devices without sending the user's typed text to Google [11], [13]. Alketbi and Mehmood survey XAI for malicious insider threat detection and argue that interpretability is necessary for analyst trust and operational decision-making [14].

Table 1. Focused Literature Matrix

Source	Main focus	Method / data	Reported value	Review relevance
Wang et al. [2]	Social engineering mechanisms	Conceptual review	6 mechanism groups	Explains human vulnerability behind attacks
Nasir et al. [4]	Insider behavior detection	Deep learning; user logs	IEEE Access study	Baseline for behavior sequence modeling
Villarreal-Vasquez et al. [12]	Long-duration insider attacks	LSTM anomaly detection	TDSC 2023	Shows why temporal context matters
Tian et al. [7]	Specific threat scenarios	Multi-head attention; CERT	+2%, +5%, +2% over baselines	Clarifies threat type, not only anomaly
Ali et al. [9]	Uncertainty-aware detection	DEC; CERT/TWOS	94.7% accuracy; 38% FP reduction	Addresses alert fatigue
Alazab et al. [10]	Privacy-preserving cybersecurity	Federated learning survey	IEEE TII review	Useful for sensitive behavioral data
Alketbi & Mehmood [14]	Explainable insider threat detection	XAI survey	IEEE Access 2025	Improves analyst trust

4. Methodology / Algorithm Description

The reviewed material includes the original submitted paper, the shorter reference-rich mini version, and recent academic and industry sources on social engineering, insider threat detection, UEBA, DEC, FL, XAI, and breach statistics. The selection priority was: (1) peer-reviewed or indexed academic sources, (2) papers with clear datasets and metrics, (3) industry reports for current breach statistics, and (4) real-world examples that make formulas understandable. Because this is a review paper, the formulas below describe methods reported in the literature; they are not new experiments conducted in this project.

A basic UEBA anomaly score can be written as:

$$z = |x_i - \mu| / \sigma$$

where x_i is the observed behavior, μ is the user's historical mean, and σ is the historical standard deviation. Real-life example: if a student employee normally downloads 20 files per day with a standard deviation of 5, but suddenly downloads 50 files after receiving a suspicious email, then z

$= |50 - 20| / 5 = 6$. A score six standard deviations from the baseline is highly unusual and should be reviewed, especially if it occurs outside normal hours.

Autoencoders use reconstruction error to detect behavior that does not match learned normal patterns:

$$E = \|x - \hat{x}\|^2$$

where x is the original behavior vector and $\hat{x}$ is the reconstructed vector. Real-life example: if the model learned normal behavior such as morning login, internal repository access, and low external upload volume, it will reconstruct those patterns well. If the same account suddenly connects from a new device and uploads many compressed files, the reconstruction error increases, producing an anomaly alert [5].

Federated learning aggregates local models without centralizing raw data:

$$w_{t+1} = \sum_{k=1}^K (n_k / n) w_t^k$$

Here, w_t^k is the model update from client k , n_k is that client's sample size, and n is the total sample size. Real-life example: several branches of a bank can train a phishing-risk model locally on their own employee logs; the central server receives only model updates, not the private logs. Google's Gboard is a widely known non-security example: it improves keyboard language models from user devices while avoiding upload of raw typed content [11], [13].

DEC adds uncertainty estimation. In simple form, a model predicts evidence e_k for each cluster; Dirichlet parameters can be represented as $\alpha_k = e_k + 1$, total evidence $S = \sum_k \alpha_k$, and epistemic uncertainty:

$$u = K / S$$

where K is the number of clusters [9]. Real-life example: if a user starts using a new approved tool, the event may be unusual but uncertain, so the system can route it to low-priority review. If the same user also downloads payroll data to an external destination at midnight, the model may produce high confidence and a stronger alert.

5. Key Concepts and Technical Approach

This section intentionally avoids explaining every cybersecurity concept in detail. It focuses only on concepts needed to understand the review. Social engineering is the manipulation of a person into unsafe action, such as approving MFA fatigue prompts or sharing credentials. UEBA is the technical countermeasure that asks: "Is this authenticated user behaving like the real user?" For example, if a finance employee logs in from Tashkent at 09:00 as usual, that may be normal; if the same account accesses source-code repositories at 03:00 from a new device, UEBA raises risk even if the password is correct.

Federated learning is included because behavioral logs are sensitive. Google FL for Gboard is a clear real-world example: learning occurs on devices, and only learning updates are combined centrally [11], [13]. In enterprise security, the same idea can allow multiple departments or hospitals to learn shared threat patterns without sending employee activity logs to one central database. XAI is included because a security analyst should know why an alert occurred. Instead of saying only "risk = 0.91," an XAI-supported alert can show that login time, file volume, destination domain, and new device were the most important factors [14].

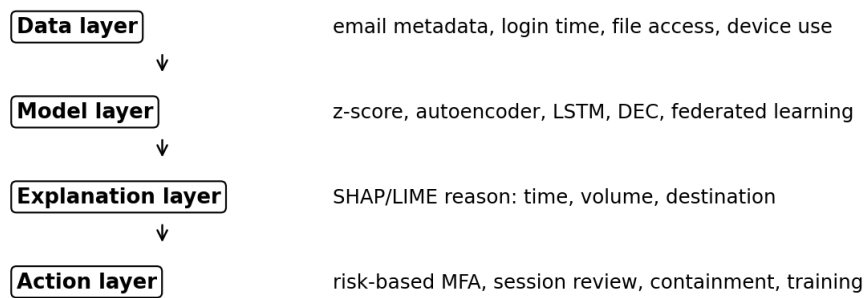
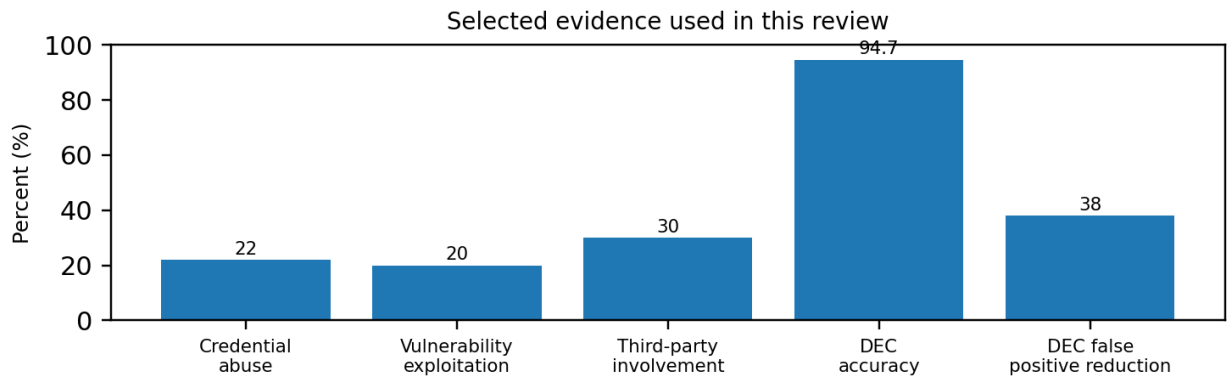


Figure 3. Minimal architecture for behavior-aware defense.

6. Results and Performance Evaluation

This section is a review synthesis, not a claim that a new experiment was performed. The values below are reported by the cited papers or reports and are used to compare trends across methods. The most important pattern is that behavior-aware models perform best when they combine temporal context, uncertainty estimation, and explainability. However, reported benchmark results should be treated carefully because datasets, feature sets, and evaluation protocols differ across studies.

Method	Purpose	Evidence / dataset	Reported result	Interpretation for review
DEC [9]	Uncertainty-aware anomaly detection	CERT and TWOS	94.7% average accuracy; 38% FP reduction	Good for reducing SOC alert fatigue
LSTM / LADOHD [12]	Long attack sequences	System event sequences	Detects slow attacks mixed with normal actions	Useful for “low and slow” insider behavior
Specific-scenario model [7]	Privilege abuse, identity theft, data leakage	CERT	Approx. +2%, +5%, +2% over strongest baselines	Better than generic anomaly labels
Deep learning [4]	Behavior-based insider detection	User activity logs	IEEE Access evaluation	Shows value of modeling user behavior
FL for cybersecurity [10]	Privacy-preserving training	Surveyed FL security cases	Identifies privacy, trust, and poisoning challenges	Useful where logs cannot be centralized
XAI survey [14]	Interpretable alerts	ML/DL insider threat literature	Explains need for SHAP/LIME and transparent models	Improves analyst confidence



The reviewed evidence suggests three practical conclusions. First, identity is a central attack path: credential abuse accounted for 22% of initial attack vectors in Verizon DBIR 2025, and this directly supports post-login behavioral monitoring [1]. Second, uncertainty-aware models are promising because they do not treat every unusual event as equally dangerous; Ali et al. report a 38% reduction in false positives for DEC compared with traditional clustering [9]. Third, scenario-aware modeling improves clarity. Tian et al. show that modeling privilege abuse, identity theft, and data leakage separately improves performance over generic baselines, which is important because “anomaly” alone is not enough for incident response [7].

A simple real-life interpretation is the following: an employee clicks a convincing AI-generated phishing email and logs into a fake portal. The attacker then uses the valid password. A signature-based tool may not detect malware because no malware is used. UEBA can still detect the account's abnormal behavior: different device, unusual login time, abnormal file access, and new network destination. XAI then explains the alert to the analyst, and FL can help train similar detection models across departments without moving raw employee data.

7. Advantages and Limitations

The enterprise transition toward AI-driven behavioral defense introduces major operational advantages, but it remains constrained by structural, mathematical, and ethical limitations that require continuous refinement.

Advantages:

Evasion Resistance: Unlike traditional signature-based systems that rely on known malware hashes or static IPs, behavioral analytics track the immutable, multidimensional “fingerprint” of user activity. [1] Even if an attacker utilizes zero-day exploits or legitimately acquired credentials, their subsequent interactions with the network - such as the speed of file access, or command-line syntax - will inevitably deviate from the victim's historical baseline, triggering a high-fidelity alert.[4]

Contextual peer grouping: Advanced algorithmic clustering groups users by specific job function and workflow rather than generic organizational roles. [17] If a software engineer routinely accesses code repositories and compilers at 2:00 AM, the UEBA system mathematically recognizes this as normal peer behavior, dramatically reducing the false alarm rate that plagues static threshold systems. [17]

Scalability via Dimensionality Reduction: Techniques like PCA allow deep learning models to ingest, compress, and analyze millions of log entries across hundreds of distinct features with sub-second latency, facilitating true, continuous real-time defense without paralyzing network bandwidth. [5]

Decentralized Privacy Protection: The maturation of Federated Learning allows multinational corporations, healthcare conglomerates, or allied governmental institutions to train a collective "super-model" on global threat behaviors. This is achieved without ever transmitting sensitive intellectual property or violating employee privacy regulations, fundamentally solving the cross-jurisdictional data-sharing crisis. [16]

Limitations:

The "Cold Start" Vulnerability: Unsupervised UEBA models require vast quantities of historical telemetry to establish a mathematically accurate baseline. When a new employee is onboarded, the system lacks sufficient historical data to construct a profile, creating a dangerous vulnerability window (often lasting 14 to 30 days) where malicious behavior cannot be accurately benchmarked against normal activity. [5]

Adversarial Model Poisoning: Machine learning baselines utilizing continuous updates (like EWMA) are highly susceptible to "boiling the frog" attacks. If an advanced insider acts slightly maliciously over an extended temporal period, the dynamic baseline may slowly self-adjust to accept this gradual malicious behavior as the new "normal," effectively blinding the detection model to eventual data exfiltration. [5]

Biometric Volatility: Systems relying heavily on keystroke dynamics or mouse movements suffer from high intra-class variance. An employee's typing rhythm can drastically change due to acute fatigue, emotional stress, physical injury, or hardware alterations (e.g., switching from a mechanical office keyboard to a flat laptop keyboard). These natural variations routinely lock out legitimate users, causing severe operational disruptions. [5]

Computational Overhead: Processing complex multi-modal data continuously using Deep Autoencoders and transformers consumes massive graphical processing unit (GPU) resources. Smaller enterprises often find the required hardware procurement and cloud-compute costs financially prohibitive, limiting advanced UEBA deployment to the Global 1000.[5]

8. Critical Review

The strongest contribution of AI-driven behavioral analytics is that it moves security from "who logged in?" to "does this behavior make sense?" This is essential for social engineering because the attacker often tries to make the victim open the door voluntarily. However, the literature also shows a risk of overconfidence. High accuracy on CERT or TWOS does not automatically mean high accuracy in a real company. Real employees change schedules, work under stress, travel, switch devices, and perform new duties. Therefore, a useful security model must be adaptive, but not so adaptive that an attacker can slowly poison the baseline.

Another issue is human intent. A log can show that a person downloaded many files, but it cannot fully prove whether the person is malicious, careless, pressured by a scammer, or simply doing urgent work. This is why the review should not treat UEBA as a complete replacement for human judgment. The best approach is decision support: AI prioritizes suspicious events, XAI explains them, and a trained analyst checks context before punishment or containment. This protects both security and fairness.

Finally, social engineering defense should not become only surveillance. If an organization only monitors employees but does not improve training, workload, reporting culture, and safe verification habits, it treats symptoms instead of causes. The future of behavioral security should

combine AI detection with human-centered security culture: easy reporting, non-punitive phishing education, phishing-resistant authentication, and privacy-preserving analytics.

9. Real-World Application

In finance, UEBA can monitor privileged administrators and detect unusual bulk export of customer records, risky payment changes, or access from unfamiliar devices. In healthcare, behavior analytics can detect unusual access to patient records or account takeover attempts, especially where staff are under time pressure and social engineering may exploit urgency. In universities, the same ideas can protect student information systems by detecting abnormal downloads, suspicious login patterns, or compromised staff accounts. In SaaS platforms, behavioral analytics supports continuous authentication by combining device fingerprinting, geography, activity rhythm, and session behavior.

A practical implementation should be proportional. Low-risk anomalies may trigger step-up authentication or user verification. Medium-risk anomalies may create an analyst ticket with XAI explanations. High-risk events, such as external transfer of sensitive data from a new device after a suspicious email, may temporarily restrict the session while preserving evidence. This layered approach is safer than fully automatic blocking for every anomaly because it reduces both false positives and user disruption.

10. Future Scope

Future work should focus on five directions. First, scenario-specific models should classify social engineering outcomes such as credential theft, MFA fatigue, malicious link follow-up, and data leakage instead of only producing a generic anomaly score. Second, FL should be hardened against poisoning and dishonest clients, because privacy-preserving learning is not automatically secure [10]. Third, XAI should move from post-hoc explanations to interpretable-by-design models where possible [14]. Fourth, benchmarks should include more realistic cross-organization data, role changes, and stress-driven behavior variation. Fifth, governance should be built into systems through data minimization, retention limits, employee transparency, and clear escalation rules.

The most promising future model is a human-centered AI defense loop: detect unusual behavior, explain the reason, verify context, respond proportionally, and update training. In this loop, AI does not replace human responsibility; it improves the speed, consistency, and context of security decisions.

11. References

- [1] Verizon Business, “2025 Data Breach Investigations Report: Alarming surge in cyberattacks through third-parties,” Apr. 2025.
- [2] Z. Wang, H. Zhu, and L. Sun, “Social Engineering in Cybersecurity: Effect Mechanisms, Human Vulnerabilities and Attack Methods,” *IEEE Access*, vol. 9, pp. 11895-11910, 2021, doi: 10.1109/ACCESS.2021.3051633.
- [3] S. Pujari and M. Hussain, “Human Factor in Cybersecurity: Behavioral Insights into Phishing and Social Engineering Attacks,” *Nanotechnology Perceptions*, 2024.
- [4] R. Nasir, M. Afzal, R. Latif, and W. Iqbal, “Behavioral Based Insider Threat Detection Using Deep Learning,” *IEEE Access*, vol. 9, pp. 143266-143274, 2021, doi: 10.1109/ACCESS.2021.3118297.

- [5] J. Fuentes, I. Ortega-Fernandez, N. M. Villanueva, and M. Sestelo, "Cybersecurity threat detection based on a UEBA framework using Deep Autoencoders," *AIMS Mathematics*, vol. 10, no. 10, pp. 23496-23517, 2025.
- [6] IBM Security, "Cost of a Data Breach Report 2025," IBM, 2025.
- [7] T. Tian, C. Zhang, B. Jiang, H. Feng, and Z. Lu, "Insider threat detection for specific threat scenarios," *Cybersecurity*, vol. 8, article 17, 2025, doi: 10.1186/s42400-024-00321-w.
- [8] P. Manoharan, J. Yin, H. Wang, Y. Zhang, and W. Ye, "Insider threat detection using supervised machine learning algorithms," *Telecommunication Systems*, 2023, doi: 10.1007/s11235-023-01085-3.
- [9] A. Ali, M. Husain, and P. Hans, "Real-Time Detection of Insider Threats Using Behavioral Analytics and Deep Evidential Clustering," *arXiv:2505.15383*, 2025, doi: 10.48550/arXiv.2505.15383.
- [10] M. Alazab, S. P. R. M., P. M., P. R. Maddikunta, T. R. Gadekallu, and Q.-V. Pham, "Federated Learning for Cybersecurity: Concepts, Challenges and Future Directions," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3501-3509, 2022, doi: 10.1109/TII.2021.3119038.
- [11] Google, "Learn how Gboard gets better," Google Gboard Help, accessed May 2026.
- [12] M. Villarreal-Vasquez, G. Modelo-Howard, S. Dube, and B. K. Bhargava, "Hunting for Insider Threats Using LSTM-Based Anomaly Detection," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 1, pp. 451-462, 2023, doi: 10.1109/TDSC.2021.3135639.
- [13] Z. Xu et al., "Federated Learning of Gboard Language Models with Differential Privacy," *ACL Industry Track*, 2023.
- [14] K. S. Alketbi and A. Mehmood, "A Comprehensive Survey of Explainable Artificial Intelligence Techniques for Malicious Insider Threat Detection," *IEEE Access*, vol. 13, pp. 121772-121798, 2025, doi: 10.1109/ACCESS.2025.3587114.
- [15] D. Le and N. Zincir-Heywood, "Anomaly Detection for Insider Threats Using Unsupervised Ensembles," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1152-1164, 2021, doi: 10.1109/TNSM.2021.3071928.
- [16]. Sarraf, G. (2026). *Behavioral Analytics for Continuous Insider Threat Detection in Zero-Trust Architectures*. ResearchGate.
- [17]Ribeiro, J. P., Petrova, E., & Oladele, O. K. (2025). *AI-Based Behavioral Analytics for Insider Threat Detection in Enterprise Information Systems*