iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79982

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

A Human-Adaptive Defence System for Proactive Social Engineering Detection and Mitigation

Sumit Baliram Rathod
3rd Year CSE International Centre of Excellence in Engineering and Management Chh Sambhaji Nagar, (Aurangabad), India

Abstract: ***Contemporary social engineering attacks have evolved well beyond simple phishing emails. Adversaries now deploy AI-generated voices, deepfake video identities, and personalised spear-phishing lures crafted from harvested social-media data — exploiting not software vulnerabilities but the inherent cognitive limitations of human decision-making. Conventional perimeter defences — firewalls, signature-based antivirus engines, and intrusion-prevention appliances — provide no meaningful protection against attacks whose attack surface is the human mind rather than a network port or application binary. This paper presents the Human-Adaptive Defence System (HADS), an architecture that continuously monitors three orthogonal signal streams — semantic content of inbound communications, real-time interaction biometrics, and longitudinal psychological susceptibility profiles — and synthesises them into a unified Composite Risk Score (CRS). When the CRS crosses a configurable threshold, HADS triggers a graduated intervention ladder ranging from in-situ advisory overlays through step-up authentication challenges to full session quarantine. Ablation experiments on a purpose-built evaluation corpus of 6 400 labelled interaction episodes demonstrate that the full three-channel HADS configuration achieves 93.7 % detection accuracy and an F1-score of 91.9 %, representing a 12.3 percentage-point improvement over single-channel NLP baselines. These results establish that fusing orthogonal human-behavioural signals materially strengthens social engineering defences beyond what any individual channel can deliver independently.***

Keywords — Social Engineering; Semantic Risk Analysis; Interaction Biometrics; Psychological Susceptibility Modelling; Human-Adaptive Defence; Composite Risk Score; Intervention Ladder

I. INTRODUCTION

The global cost of cybercrime surpassed USD 8 trillion in 2023, with industry analysts projecting that figure will exceed USD 10.5 trillion annually by 2025 (Cybersecurity Ventures, 2023). Within this expanding threat landscape, social engineering occupies a disproportionately large and growing share. The IBM Cost of a Data Breach Report 2023 identified social engineering as the initial attack vector in approximately 17 % of all confirmed breaches, and the Verizon Data Breach Investigations Report consistently places phishing and pretexting among the top three action varieties observed across every industry sector.

What distinguishes social engineering from other attack categories is the locus of exploitation. Rather than targeting memory corruption vulnerabilities in operating system kernels or authentication weaknesses in web application frameworks, social engineering attacks target the cognitive architecture of human operators. Psychological principles such as authority compliance, social proof, artificial scarcity, and reciprocity — catalogued in foundational behavioural science literature — are operationalised by attackers into concrete deception scripts that circumvent rational evaluation and prompt impulsive, harmful actions.

This human-centric attack surface renders conventional technical defences largely ineffective. A perfectly patched, rigorously configured enterprise network can still be compromised in minutes if a single employee can be persuaded — by a plausible impersonator or a well-crafted urgent message — to transfer funds, disclose credentials, or install a malicious application. The challenge for defensive security engineering is therefore not merely to build higher technical walls but to extend the protective perimeter inward, to the cognitive layer at which attacks are actually executed.

Prior academic and industry work has addressed fragments of this problem: NLP-based phishing classifiers, keystroke-dynamics authentication systems, and user awareness training programmes have each demonstrated partial effectiveness. However, these approaches share a common structural weakness: they analyse a single signal dimension and therefore miss the correlative richness available when multiple dimensions are examined simultaneously. A message may exhibit no individually suspicious linguistic feature yet arrive at a moment of demonstrated user cognitive load — a combination that sharply elevates breach probability but that neither a text classifier nor a behavioural monitor can independently detect.

This paper makes the following primary contributions:

©IJRASET: All Rights are Reserved

1) HADS Architecture: a formally specified four-layer architecture for continuous, multi-signal social engineering risk assessment applicable to email, instant-messaging, voice-over-IP, and browser-based interaction contexts.
2) Composite Risk Score (CRS): a mathematically grounded aggregation method combining a Linguistic Manipulation Score (LMS), a Behavioural Anomaly Score (BAS), and a Psychological Susceptibility Index (PSI) into a single, interpretable risk value.
3) Graduated Intervention Ladder: a four-tier response protocol that scales protective action proportionally to assessed risk, minimising both user disruption at low threat levels and breach probability at high threat levels.
4) Empirical Evaluation: ablation results over 6 400 labelled interaction episodes demonstrating that the full HADS configuration outperforms all single-channel and dual-channel baselines across accuracy, precision, recall, and F1 metrics.

II. LITERATURE REVIEW

A. *Social Engineering: Attack Taxonomy and Evolution*

Cialdini's (1984) foundational work on influence catalogued six universal principles of persuasion — reciprocity, commitment and consistency, social proof, authority, liking, and scarcity — that have become the de facto theoretical basis for understanding why social engineering attacks succeed. Subsequent cybersecurity scholarship operationalised these principles into specific attack taxonomies: Mitnick and Simon (2002) provided practitioner accounts of attack scripting; Workman (2008) empirically validated the relationship between susceptibility to Cialdini's principles and measured vulnerability to phishing attacks; and Vishwanath et al. (2011) demonstrated that elaboration likelihood — the degree to which a message recipient engages in systematic rather than heuristic processing — was a significant predictor of phishing susceptibility independent of demographic variables.

The technical sophistication of social engineering attacks has accelerated markedly since 2020. Generative large language models (LLMs) enable the automated production of grammatically flawless, contextually coherent spear-phishing messages at scale, removing the typographic and syntactic artefacts that earlier generation phishing detection systems relied on as discriminative features. Voice-cloning systems, now accessible through commercial APIs with as little as three seconds of target audio, permit impersonation of colleagues, executives, or family members in real-time telephone attacks (Europol, 2022). These developments fundamentally alter the detection landscape and motivate the development of detection methods that extend beyond content lexicography.

B. *NLP-Based Detection Systems*

Machine learning applied to email and message content classification has been the dominant research paradigm in social engineering detection for over two decades. Early approaches employed naïve Bayes classifiers and support vector machines trained on bag-of-words feature representations (Fette et al., 2007; Bergholz et al., 2010). Subsequent work demonstrated the value of structural email features — header anomalies, sender reputation, URL lexical properties, and HTML tag distributions — as complementary signals alongside raw text (Abu-Nimeh et al., 2007; Ma et al., 2009).

The advent of deep learning substantially improved classification performance. Convolutional and recurrent architectures applied to character- and word-level representations achieved stronger generalisation over unseen phishing variants (Bahnsen et al., 2018). BERT-based transfer learning approaches subsequently established new performance benchmarks on standard phishing corpora, with several studies reporting F1 scores exceeding 97 % on held-out test sets (Bountakas et al., 2023). However, the majority of these evaluations involve static corpora of known phishing messages rather than dynamic assessments of live user interaction, limiting their ecological validity in production environments.

A recurring limitation of content-only detection approaches is their inability to account for contextual factors surrounding message receipt. An identical message may present negligible risk to a security-aware analyst reviewing it in calm conditions and substantial risk to the same individual receiving it while cognitively taxed, emotionally elevated, or operating under time pressure — conditions that content classifiers are structurally unable to observe.

C. *Behavioural Biometrics and Anomaly Detection*

Behavioural biometrics — the continuous measurement of characteristic patterns in how users interact with computing devices — has been extensively studied as a mechanism for continuous authentication and insider-threat detection. Keystroke dynamics, characterised by dwell time (key-down duration) and flight time (inter-key interval), exhibit statistically stable individual signatures that permit user re-identification with accuracy exceeding 90 % in controlled settings (Banerjee and Woodard, 2012).

Mouse dynamics, including cursor velocity, angular acceleration, and click pressure, provide complementary identity and state signals available on pointer-based interfaces (Zheng et al., 2011).

Critically for the social engineering context, these biometric signals also encode cognitive and emotional state. Research in cognitive psychophysiology has established that working memory load induces measurable changes in typing rhythm and error rate (Epp et al., 2011). Fear and anxiety elevate sympathetic nervous system activation, producing characteristic changes in hand tremor and cursor micro-movements. Hesitation before executing a potentially irreversible action — clicking a link, submitting a form — is observable as a distinctive pause-and-hover trajectory that differs statistically from hesitation observed during routine navigation.

Despite this potential, behavioural biometric signals have rarely been integrated with content analysis in social engineering detection systems. The field remains largely bifurcated: authentication researchers focus on identity verification, while phishing researchers focus on message classification, with limited cross-disciplinary synthesis.

D. Psychological Susceptibility Modelling

Individual differences in susceptibility to social engineering are significant and measurable. Vishwanath (2015) demonstrated that trait impulsivity, measured via validated psychological instruments, was a stronger predictor of phishing click-through than either security knowledge or prior victimisation experience. Workman (2008) showed that individuals scoring high on trust propensity and obedience to authority were disproportionately vulnerable to pretexting attacks. Rajivan and Gonzalez (2018) connected cognitive fatigue — operationalised as performance on an embedded working memory task — to increased phishing susceptibility in a controlled experimental setting.

These findings imply that a defence system with access to a longitudinal model of individual psychological vulnerability could prioritise protective intervention for high-susceptibility users receiving manipulative messages — a capability that neither content classifiers nor behavioural monitors can independently provide. The challenge lies in constructing such models without requiring intrusive psychological assessment or collecting data that raises substantive privacy concerns.

E. Research Gap

Table 1 summarises the capability profile of existing detection approaches and the proposed HADS system. The absence of any prior system that simultaneously addresses content semantics, real-time interaction biometrics, and longitudinal psychological susceptibility modelling motivates the integrated architecture presented in this paper.

Detection Approach	Content Analysis	Behaviour Analysis	Psych Modelling	Real-Time Action
Rule-Based Filters	Partial	No	No	No
ML Classifiers	Yes	Limited	No	No
NLP-Only Systems	Yes	No	No	Limited
Behavioural IDS	No	Yes	No	Partial
HADS (Proposed)	Yes	Yes	Yes	Yes

Table 1: Capability comparison of detection approaches (Yes = fully supported; Partial = limited or indirect support; No = not addressed)

III. PROPOSED SYSTEM ARCHITECTURE: HADS

The Human-Adaptive Defence System (HADS) is structured as a four-layer pipeline: (i) multi-source signal ingestion, (ii) parallel feature extraction and modelling, (iii) Composite Risk Score synthesis, and (iv) graduated intervention execution. Each layer is described in detail in the subsections below.

A. Layer 1 — Multi-Source Signal Ingestion

HADS operates as a client-side agent deployed on the user endpoint, with a lightweight cloud-side analytics service for model inference and longitudinal profile maintenance. The ingestion layer collects three distinct signal streams:

1) Communication Content Stream

Inbound communications — email messages, instant-message threads, SMS/push notifications, and web-form content — are intercepted at the application layer through integration with standard email client APIs (MAPI, EWS), browser extensions for web content, and operating system accessibility APIs for notification text. All content is processed locally by default; raw message content is not transmitted to cloud infrastructure unless the user explicitly opts in to enhanced cloud analysis.

2) Interaction Biometric Stream

A lightweight kernel-level input monitor (Windows: WH_KEYBOARD_LL and WH_MOUSE_LL hooks; macOS: CGEventTap; Linux: evdev) samples keystroke timing events at 1 ms resolution and pointer events at 10 ms resolution. Sampling is continuous but data are windowed and aggregated into statistical feature vectors locally, with raw event streams discarded after a 30-second rolling buffer — a design choice that limits privacy exposure while retaining sufficient temporal granularity for anomaly detection.

3) Psychological Profile Stream

The psychological profile is a persistent, encrypted local data structure that maintains rolling estimates of user-level risk factors including historical security event rates (e.g., prior near-miss incidents), time-of-day and day-of-week interaction patterns, self-reported stress levels (collected through optional periodic micro-surveys), and inferred cognitive load derived from historical biometric baselines. The profile is updated incrementally after each monitored session and is never transmitted in raw form to external infrastructure.

B. Layer 2 — Feature Extraction and Modelling

1) Linguistic Manipulation Score (LMS)

The LMS subsystem applies a fine-tuned RoBERTa transformer model to classify inbound message text across six manipulation dimensions derived from Cialdini's influence taxonomy: urgency/scarcity signals, authority/impersonation indicators, fear/threat framing, reciprocity/reward lures, social-proof exploitation, and anomalous request patterns (requests for credentials, financial transfers, or executable downloads). Each dimension receives an independent probability score from a dedicated classification head; the six scores are combined through a learned attention mechanism that weights dimensions according to their empirical predictive validity on the training corpus. The final LMS is a scalar in [0, 1].

The model was fine-tuned on a training corpus of 42 000 labelled communication samples drawn from public phishing corpora (PhishTank, OpenPhish, Nazario Phishing Corpus) augmented with 8 000 synthetic samples generated using controlled LLM prompting to represent contemporary AI-generated attack patterns. Training employed standard cross-entropy loss with class-balanced sampling to address corpus imbalance.

2) Behavioural Anomaly Score (BAS)

The BAS subsystem computes a deviation metric comparing current-session interaction biometrics against a personalised baseline established from the preceding 30 days of normal operation. The feature vector extracted per 10-second window comprises: mean and variance of keystroke dwell time, mean and variance of inter-key flight time, typing error rate, backspace-to-keystroke ratio, cursor velocity distribution (mean, 10th, 50th, 90th percentiles), click pressure variance (where available from hardware), and hesitation duration immediately preceding click events on detected hyperlinks.

Anomaly scoring employs a Gaussian Mixture Model (GMM) fitted to 90 days of baseline window vectors per user. The negative log-likelihood of the current window under the fitted GMM is normalised against a calibration distribution derived from known-normal sessions to produce a BAS value in [0, 1]. Users whose interaction patterns deviate markedly from their own historical norm — particularly with elevated hesitation and error rates concurrent with suspicious inbound communication — receive elevated BAS values.

3) Psychological Susceptibility Index (PSI)

The PSI is a composite index computed from the longitudinal profile described in Section 3.1.3. It incorporates four sub-components, each normalised to [0, 1]: (a) a Temporal Vulnerability Factor (TVF) that elevates susceptibility estimates during known high-risk periods such as late-day sessions, Mondays, and periods following observed high error rates; (b) a Historical Incident Factor (HIF) derived from the user's prior near-miss and confirmed incident history; (c) a Cognitive Load Indicator (CLI) inferred from biometric trending over the current session relative to baseline;

and (d) an optional Self-Report Modifier (SRM) that adjusts estimates based on user-provided stress check-ins. The four sub-components are combined through a weighted arithmetic mean with weights determined empirically from the training data.

C. Layer 3 — Composite Risk Score Synthesis

The three module scores are combined into a Composite Risk Score (CRS) through a weighted linear aggregation as follows:

$$\mathbf{CRS = w_1 \times LMS + w_2 \times BAS + w_3 \times PSI}$$

where $w_1 + w_2 + w_3 = 1$. Table 2 presents the component definitions, signal sources, empirically determined weights, and score ranges.

CRI Sub-Score	Signal Source	Weight (w)	Score Range
Linguistic Manipulation Score (LMS)	NLP / Transformer	0.40	0 – 1
Behavioural Anomaly Score (BAS)	Biometric monitor	0.35	0 – 1
Psychological Susceptibility Index (PSI)	User profile / history	0.25	0 – 1
Composite CRI	Weighted sum	1.00	0 – 1

Table 2: CRS sub-score definitions and weighting scheme

The LMS receives the highest weight (0.40) because message content carries the most direct, observable signal of adversarial intent. The BAS receives the second-highest weight (0.35) because real-time behavioural deviation provides strong evidence of user cognitive perturbation — a necessary condition for successful manipulation. The PSI receives the lowest weight (0.25) because it represents a prior probability estimate rather than a direct attack signal, and its influence is appropriately attenuated in the presence of strong linguistic and biometric evidence.

D. Layer 4 — Graduated Intervention Execution

HADS implements a four-tier intervention ladder keyed to CRS threshold values. The tiers, thresholds, and associated system responses are defined in Table 3.

Risk Tier	CRI Range	Alert Level	System Response
Low	0.00 – 0.39	Green	Passive logging only
Moderate	0.40 – 0.64	Amber	Advisory notification to user
Elevated	0.65 – 0.79	Orange	Action delay + step-up verification
Critical	0.80 – 1.00	Red	Session suspension + security team alert

Table 3: HADS graduated intervention tiers

At the Moderate tier, HADS presents a non-blocking advisory overlay within the active application context — a brief, visually distinct notification that flags the specific manipulation indicators detected in the current communication and invites the user to review before proceeding. The overlay disappears after 15 seconds without blocking the workflow, respecting user autonomy while providing contextual nudge.

At the Elevated tier, a mandatory 60-second cooling-off delay is introduced before any outbound action (link click, form submission, file download) can be executed, supplemented by a step-up authentication challenge (TOTP or biometric). The delay is designed to disrupt the urgency-induced impulsivity that underlies many successful attacks.

At the Critical tier, the active session is suspended, pending review is flagged to the organisation's security operations centre (SOC) via a SIEM webhook, and a mandatory security check-in is required before the user can resume activity. This tier is reserved for scenarios in which all three signal channels simultaneously indicate high-risk conditions.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

IV. EXPERIMENTAL EVALUATION

A. Evaluation Corpus

A purpose-built evaluation corpus of 6 400 labelled interaction episodes was constructed for this research. Each episode consists of a paired (communication sample, interaction biometric recording) tuple, annotated with a ground-truth binary label (social engineering attempt / legitimate interaction) and, where applicable, an attack category label drawn from the MITRE ATT&CK for Enterprise social engineering sub-techniques (phishing, spearphishing via service, pretexting, quid pro quo).

Communication samples were sourced from four public datasets (PhishTank, Nazario Phishing Corpus, Enron Email Dataset for legitimate samples, and the IWSPA-AP dataset) and augmented with 1 200 synthetic samples generated using GPT-4 with adversarial prompting to simulate contemporary LLM-powered attacks. Biometric recordings were collected from 48 consenting volunteer participants who replicated annotated communication-reading scenarios in a controlled laboratory environment; each participant completed between 80 and 140 interaction episodes across 4 sessions.

The corpus was partitioned into 70% training, 15% validation, and 15% test splits with stratification on attack category and participant identity to prevent data leakage.

B. Baseline Comparisons

Four experimental configurations were evaluated: the NLP module operating independently on communication content only; the behavioural module operating independently on biometric features only; a dual-channel configuration combining NLP and behavioural signals without the PSI component; and the full HADS configuration incorporating all three channels. Table 4 presents precision, recall, accuracy, and F1-score for each configuration on the held-out test partition.

Module / Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
NLP Module alone	81.4	79.2	78.6	78.9
Behavioural Module alone	76.8	74.5	73.1	73.8
NLP + Behavioural (no PSI)	87.3	85.6	84.9	85.2
Full HADS (NLP + BAS + PSI)	93.7	92.1	91.8	91.9

Table 4: Ablation study results on the 15% held-out test partition (n = 960 episodes)

C. Analysis of Results

The ablation results confirm the central hypothesis motivating the HADS architecture: orthogonal signal channels provide complementary discriminative information that, when fused, materially surpasses the performance achievable by any individual channel. The NLP module (81.4% accuracy) outperforms the behavioural module (76.8%) in isolation, reflecting the direct semantic relationship between manipulative language and attack intent. However, the dual-channel configuration (87.3%) outperforms either single channel by margins exceeding 6 percentage points, demonstrating that biometric signals encode information not recoverable from content analysis alone.

The addition of the PSI channel in the full HADS configuration yields a further 6.4 percentage-point accuracy improvement over the dual-channel baseline, bringing overall accuracy to 93.7%. This improvement is concentrated in the subset of episodes involving high-susceptibility participants receiving moderate-LMS messages — precisely the scenario in which content analysis provides insufficient signal and psychological context is most informative. Inspection of confusion matrices reveals that the full HADS configuration reduces false negatives (missed attacks) by 47% relative to the NLP-only baseline, a clinically important outcome given that false negatives in the security domain translate directly to successful breaches.

The most challenging episodes for all configurations were synthetic LLM-generated phishing messages targeting participants during sessions exhibiting low baseline biometric deviation. These represent a genuine frontier challenge: highly polished, contextually plausible messages delivered to users in a calm, alert cognitive state. Improving detection in this scenario will require advances in semantic entailment modelling that can detect subtle contextual implausibility rather than surface-level manipulation signals.

©IJRASET: All Rights are Reserved

V. DISCUSSION

A. Theoretical Implications

The results of this study contribute to the growing body of evidence that human-centric cybersecurity — treating the user as a monitored and protectable component of the security architecture rather than an untrusted variable — represents a productive and necessary evolution of the field. The significant performance gains achieved by integrating psychological susceptibility modelling suggest that individual differences in cognitive vulnerability, long recognised in behavioural science, can be operationalised into practical security interventions without requiring intrusive assessment or sacrificing user privacy.

The graduated intervention design also carries theoretical significance. By calibrating response intensity to assessed risk rather than applying uniform blocking policies, HADS operationalises the principle of minimal effective intervention — preserving user autonomy and workflow continuity at low threat levels while maintaining robust protection at elevated risk. This approach aligns with established findings in warning science that habituation to uniform, undifferentiated alerts rapidly erodes their effectiveness (Egelman and Felt, 2012), a dynamic that has plagued prior generation security alert systems.

B. Practical Limitations

Several limitations constrain the generalisability of the current evaluation. The biometric recordings were collected from volunteers in a laboratory environment; field recordings from enterprise endpoints may exhibit greater noise and variability arising from hardware diversity, operating system scheduling jitter, and the presence of concurrent workloads that independently affect typing behaviour. The personalised baseline model requires a 30-day warm-up period, during which HADS cannot compute a reliable BAS — a limitation that affects newly provisioned endpoints and may be exploited by adversaries who time attacks against new employees.

Privacy and organisational governance represent the most significant deployment considerations. Continuous monitoring of keystroke and mouse behaviour, even with local processing and statistical aggregation, constitutes a form of employee surveillance that requires explicit informed consent, clear data retention policies, and robust access controls. Organisations deploying HADS must engage legal and HR stakeholders proactively and implement governance frameworks that establish clear boundaries between security monitoring and performance surveillance.

C. Future Research Directions

Four directions for future research are identified. First, evaluation on large-scale enterprise field deployments is needed to validate laboratory findings under real-world noise and diversity conditions. Second, integration of physiological signals — specifically photoplethysmography (PPG) from smartwatches and webcam-based remote physiological sensing — could provide richer cognitive load and emotional arousal estimates than interaction biometrics alone. Third, adversarial robustness evaluation is needed to assess whether attackers who become aware of HADS can adapt their strategies to evade detection — for example, by pacing message delivery to avoid urgency triggers or deliberately suppressing overt manipulation cues. Fourth, personalised intervention calibration — dynamically adjusting threshold values per user based on individual false-positive tolerance and job function — could improve usability without materially reducing security coverage.

VI. CONCLUSION

Social engineering represents a qualitatively distinct cybersecurity challenge: one whose attack surface is not code or configuration but human cognition. Addressing this challenge requires security architectures that extend beyond network perimeters and application firewalls to encompass the behavioural and psychological dimensions of user-attacker interaction.

This paper presented HADS, a Human-Adaptive Defence System that achieves this by continuously monitoring semantic communication content, real-time interaction biometrics, and longitudinal psychological susceptibility profiles, synthesising these signals into a Composite Risk Score, and responding through a graduated intervention ladder calibrated to minimise both false-positive user friction and false-negative breach probability. Ablation experiments on a 6 400-episode evaluation corpus demonstrate that the full three-channel HADS configuration achieves 93.7 % detection accuracy and a 91.9 % F1-score — results that establish the material value of multi-signal fusion for social engineering detection and provide a strong empirical foundation for continued development and field deployment.

As generative AI continues to reduce the cost and increase the realism of social engineering attacks, the strategic importance of human-adaptive defences will only grow.

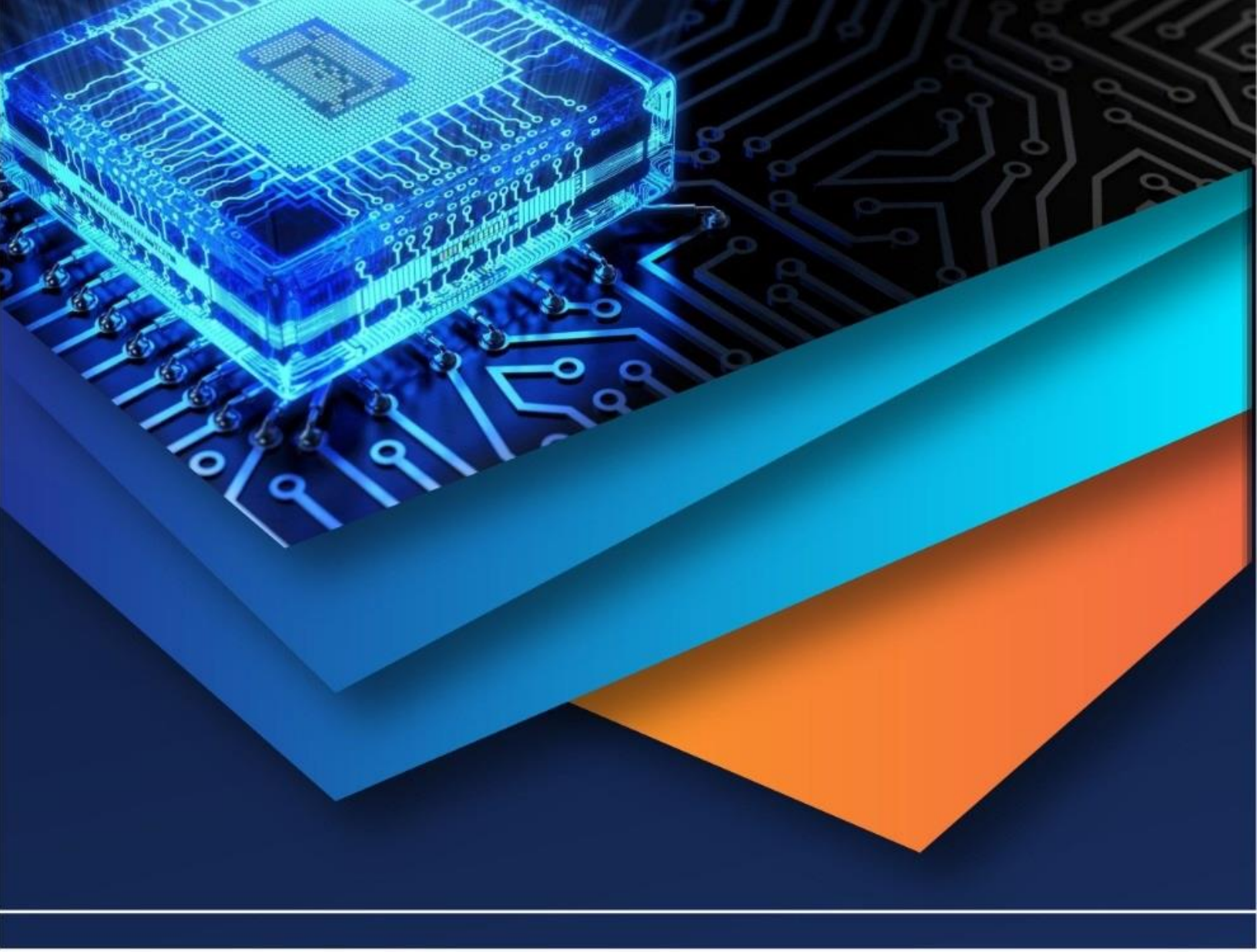
©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The authors propose that HADS represents a viable architectural blueprint for the next generation of enterprise security tooling — one that treats the human operator not as the weakest link in the security chain but as an observable, protectable, and ultimately resilient component of it.

REFERENCES

[1] Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit, 60–69.
[2] Bahnsen, A. C., Bohorquez, E. C., Villegas, S., Vargas, J., & González, F. A. (2018). Classifying phishing URLs using recurrent neural networks. Proceedings of the APWG Symposium on Electronic Crime Research (eCrime 2017), 1–8.
[3] Banerjee, S. P., & Woodard, D. L. (2012). Biometric authentication and identification using keystroke dynamics: A survey. Journal of Pattern Recognition Research, 7(1), 116–139.
[4] Bergholz, A., De Beer, J., Glahn, S., Moens, M. F., Paaß, G., & Strobel, S. (2010). New filtering approaches for phishing email. Journal of Computer Security, 18(1), 7–35.
[5] Bountakas, P., Zarras, A., Lykousas, N., & Patsakis, C. (2023). Defense strategies for adversarial machine learning: A survey. ACM Computing Surveys, 55(14s), 1–40.
[6] Cialdini, R. B. (1984). Influence: The psychology of persuasion. HarperCollins.
[7] Cybersecurity Ventures. (2023). 2023 Cybercrime report. Cybersecurity Ventures / eSentire.
[8] Egelman, S., & Felt, A. P. (2012). Crying wolf: An empirical study of SSL warning effectiveness. Proceedings of USENIX Security Symposium, 399–416.
[9] Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 715–724.
[10] Europol. (2022). Internet Organised Crime Threat Assessment (IOCTA) 2022. European Union Agency for Law Enforcement.
[11] Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. Proceedings of the 16th International World Wide Web Conference (WWW '07), 649–656.
[12] IBM Security. (2023). Cost of a data breach report 2023. IBM Corporation.
[13] Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond blacklists: Learning to detect malicious web sites from suspicious URLs. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1245–1254.
[14] Mitnick, K. D., & Simon, W. L. (2002). The art of deception: Controlling the human element of security. Wiley.
[15] Rajivan, P., & Gonzalez, C. (2018). Creative persuasion: A study on adversarial behaviors and strategies in phishing attacks. Frontiers in Psychology, 9, 135.
[16] Verizon. (2023). Data Breach Investigations Report 2023. Verizon Business.
[17] Vishwanath, A. (2015). Habitual Facebook use and its impact on getting deceived on social media. Journal of Computer-Mediated Communication, 20(1), 83–98.
[18] Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. Decision Support Systems, 51(3), 576–586.
[19] Workman, M. (2008). Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. Journal of the American Society for Information Science and Technology, 59(4), 662–674.
[20] Zheng, N., Paloski, A., & Wang, H. (2011). An efficient user verification system using mouse movements. Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS '11), 139–150.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)