ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

Designing a Privacy-Preserving Behavioral Phishing Detection Artifact: A Design Science Framework Using Federated Differential Privacy

Alunyo, Inemesit Isaiah[1]; Aruwa, Benedict Mohammed[2]

[1]Department of Cybersecurity, Mewar International University, Kuchikau, Nasarawa State, Nigeria
[2]Department of Educational Psychology, Bingham University, Karu, Nasarawa State, Nigeria

Publication Date: 2026/05/22

Abstract: Current detection frameworks largely analyze emails as technical artifacts while overlooking the behavioral evidence generated during user evaluation. This gap is consequential in settings such as Nigeria, where rapid digital adoption and emerging data-protection obligations under the Nigeria Data Protection Act 2023 (NDPA 2023) create both an elevated threat environment and a legal requirement for privacy-conscious security design. This paper applies design science research (DSR) methodology following the process model of Peffers et al. (2007). It addresses problem identification, definition of solution objectives, and artifact design, and conducts an internal ex ante evaluation covering theoretical coherence, design requirement traceability, architectural consistency, and regulatory risk alignment. The artifact has four components: a theory-derived behavioral feature model grounded in Human Error Theory (HET), Dual-Process Theory (DPT), and Protection Motivation Theory (PMT); a client-side pipeline producing 12 privacy-minimized interaction indicators; a distributed federated learning (FL) architecture with participant-level differential privacy (DP) and a formal adversarial threat model; and a governance layer aligned with NDPA 2023 obligations. Its value lies in theory-derived behavioral feature operationalization, distributed DP with adversarial threat modeling, and an NDPA-aligned governance specification, contributing to cognitive cybersecurity, privacy-preserving machine learning, and African digital governance research.

Keywords: *Phishing Detection; Behavioral Analytics; Human Error Theory; Federated Learning; Differential Privacy; Design Science Research; Nigeria; NDPA 2023; Cognitive Cybersecurity; Privacy-Preserving Machine Learning.*

How to Cite: Alunyo, Inemesit Isaiah; Aruwa, Benedict Mohammed (2026) Designing a Privacy-Preserving Behavioral Phishing Detection Artifact: A Design Science Framework Using Federated Differential Privacy. *International Journal of Innovative Science and Research Technology*, 11(5), 1024-1039. https://doi.org/10.38124/ijisrt/26may265

I. INTRODUCTION

Phishing attacks persist as one of the most costly and operationally adaptable threats in cybersecurity. The Anti-Phishing Working Group (APWG, 2025) recorded 989,123 phishing attacks in the fourth quarter of 2024, sustained across financial, e-commerce, and social media sectors. Verizon (2025) reported that the human element features in approximately 60 percent of confirmed data breaches. These figures exist despite widespread deployment of machine learning content filters, URL reputation engines, and email authentication protocols such as SPF, DKIM, and DMARC. Phishing attacks succeed because they exploit the cognitive and motivational weaknesses of email recipients, not vulnerabilities in the email infrastructure itself [1], [2]. A phishing email succeeds not when it is received, but when a user acts upon it. That distinction is the conceptual foundation of this paper.

The challenge is especially pronounced in rapidly digitizing economies. Nigeria's mobile internet penetration and digital financial services sector have expanded considerably, and digital payment transaction volumes have grown substantially, placing financially sensitive activity in the hands of users who may have limited exposure to deceptive digital communication patterns [3]. Research consistently associates rapid digital adoption with elevated phishing susceptibility [4], [5]. The NDPA 2023 and the Nigeria Data Protection Commission (NDPC) have introduced substantive obligations for systems processing behavioral and personal data, including consent, minimization, purpose limitation, and data protection impact assessment (DPIA) [6]. Privacy is a legal requirement in this context. Nigeria represents a theoretically significant setting for artifact design because it combines high digital adoption, uneven cybersecurity literacy, substantial phishing exposure, and an emerging statutory data-protection regime.

The literature review was conducted narratively across cybersecurity, HCI security, phishing susceptibility, privacy-preserving machine learning, and Nigerian data-protection scholarship, with emphasis on peer-reviewed studies and foundational technical papers. Within the reviewed literature, no DSR artifact was identified that explicitly integrates

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

theory-derived behavioral phishing features, federated differential privacy, adversarial threat modeling, and NDPA 2023-aligned governance for the Nigerian digital context. Mainstream operational detection systems remain centered on static email, URL, and metadata artifacts; behavioral interaction data generated during email evaluation is under-standardized and rarely integrated with formal privacy-preserving learning architectures.

This paper responds to that gap through DSR, which Hevner et al. (2004) define as research that creates and evaluates information technology artifacts intended to solve identified organizational problems [7]. Following the DSR process model of Peffers et al. (2007) and Gregor and Hevner's (2013) knowledge contribution framework [8], [9], the paper proposes, justifies, and specifies a privacy-preserving behavioral phishing detection artifact and conducts an internal ex ante evaluation of its theoretical coherence, architectural consistency, and regulatory alignment. Empirical ex post evaluation is reserved for a companion study. The contribution type is Improvement, yielding also Nascent Theory through the construct-to-feature mapping and Method through the replicable evaluation protocol.

A. Artifact Definition

The artifact is a privacy-preserving behavioral phishing detection framework with four integrated design components: (1) a theory-derived behavioral feature model operationalizing human error signals from email interaction data; (2) a client-side behavioral feature extraction pipeline converting raw interaction events into 12 privacy-minimized predictive indicators; (3) a distributed FL architecture with participant-level DP for decentralized phishing-risk modeling; and (4) a governance layer aligned with NDPA 2023 obligations of consent, minimization, purpose limitation, retention limits, and DPIA. The artifact is presented as an architecturally specified, evaluation-ready framework rather than a deployed system with measured performance outcomes.

II. BACKGROUND AND RELATED WORK

A. Technical Phishing Detection

Technical phishing detection has advanced through three broad phases: rule-based blacklisting, classical machine learning on engineered features, and deep learning with contextual language representation. Blacklisting is trivially bypassed when attackers register new domains [10]. Sahingoz et al. (2019) evaluated seven classifiers on more than 73,000 URL samples and reported Random Forest accuracy exceeding 97 percent, yet static URL datasets cannot represent adversarial adaptation [2]. URL-only models miss phishing where manipulation is delivered through social engineering text rather than malicious links [11]. Natural language processing methods have extended detection to message content through TF-IDF, word embeddings, and transformer architectures [12]. Most operational detection systems remain centered on static email, URL, and metadata artifacts. Only limited work engages directly with the dynamic interaction process through which phishing achieves behavioral effect.

➢ *Gap and Design Requirement:*

Mainstream detection neglects user-side interaction signals. DR1: The artifact must capture behavioral signals generated during the email evaluation process.

B. Human Susceptibility and Cognitive Error

Phishing susceptibility follows systematic cognitive and motivational patterns. Vishwanath et al. (2011) demonstrated that heuristic processing predicts susceptibility, moderated by email involvement and habitual use patterns [5]. Canfield et al. (2016) found that self-reported security awareness does not reliably predict actual susceptibility [13]. Vishwanath et al. (2018) identified automaticity as a central mediator [1]. Wright et al. (2014) showed that authority cues, scarcity framing, and social proof significantly affect attack success [14], consistent with Cialdini's (2007) influence framework [15]. A structural limitation running through this literature is the operationalization gap: theoretical constructs explaining phishing susceptibility are rarely translated into continuous, measurable, computationally deployable features.

➢ *Gap and Design Requirement:*

Cognitive constructs lack computationally deployable operationalization. DR2: The artifact must derive behavioral features from established cognitive and motivational theories.

C. Behavioral Interaction Analytics in Security

Evidence for the diagnostic value of behavioral interaction data in phishing contexts has grown steadily. Lain et al. (2022) conducted a long-term organizational study with more than 14,000 participants and found that behavioral logs captured during simulated phishing exposures predicted susceptibility outcomes [16]. Oliveira et al. (2017) documented qualitatively different visual attention patterns between susceptible and non-susceptible users [17]. Vishwanath (2015) showed that habitual email processing predicts susceptibility [18]. Gallo et al. (2024) developed and tested a purpose-built system for collecting user behavior while reading emails, confirming the technical feasibility of behavioral data collection in phishing research [19]. Despite these advances, behavioral phishing studies predominantly use centralized data collection, and feature engineering tends to be post-hoc and atheoretical.

➢ *Gap and Design Requirement:*

Behavioral analytics lacks theory-driven features and privacy-preserving collection. DR3: Raw behavioral logs must remain on the client device.

D. Privacy-Preserving Machine Learning for Cybersecurity

FL, introduced by McMahan et al. (2017), enables collaborative model development without centralizing raw data [20]. Kairouz et al. (2021) provide a comprehensive treatment of FL challenges including non-IID data distributions and the limits of privacy guarantees [21]. DP, formalized by Dwork and Roth (2014), bounds the probability of any inference about individual data to $\exp(\varepsilon)$ with probability 1 minus δ [22]. Abadi et al. (2016) integrated DP into stochastic gradient descent through gradient clipping and Gaussian noise addition [23]. Geyer et al. (2017) extended this to client-level DP in federated settings [24]. FL alone does not constitute a privacy

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

guarantee. Zhu et al. (2019) demonstrated gradient inversion attacks [25]. Nasr et al. (2019) showed membership inference vulnerabilities [26]. Jagielski et al. (2018) demonstrated effective poisoning attacks [27]. Bonawitz et al. (2017) partially addressed these through secure aggregation [28]. Bagdasaryan et al. (2020) showed that backdoor attacks can evade norm-based screening [29], reinforcing that robust aggregation methods requiring individual update visibility conflict with secure aggregation [30], [31].

➢ *Gap and Design Requirement:*

FL and DP have not been applied to behavioral phishing detection; adversarial threat modeling is absent. DR4: The artifact must provide formal privacy protection beyond FL, with a comprehensive adversarial threat model.

E. Nigerian and African Digital Context

Nigeria's digital environment presents phishing threat conditions not examined in the predominantly Western empirical literature. The NCC (2023) documents rapid growth in mobile internet adoption and digital financial service usage [32]. The Central Bank of Nigeria (2024) reports substantial growth in digital payment transaction volumes [3]. Kshetri (2019) argues that cybercrime in Africa is amplified by institutional capacity gaps and limited regulatory enforcement [4]. Research on African cybersecurity governance notes that rapid digital adoption in sub-Saharan Africa has not been matched by equally rapid development of national cybersecurity strategies or user-awareness programs [33]. The NDPA 2023 and the NDPC establish substantive obligations for systems processing personal data, covering consent, minimization, purpose limitation, data subject rights, and DPIA requirements [6]. No behavioral phishing study identified in the reviewed literature engages this regulatory framework or includes a Nigerian participant sample.

➢ *Gap and Design Requirement:*

No behavioral phishing artifact addresses the Nigerian regulatory environment. DR5: The artifact governance layer must align with NDPA 2023 requirements.

F. Gap Synthesis

Table 1 consolidates the five literature streams into a structured gap analysis.

Table 1 Critical Gap Analysis Across Research Streams With Derived Design Requirements

Research Stream	Prior Contributions	Unresolved Gap	Derived Design Requirement
Technical phishing detection	Strong ML performance on static URL, content, and metadata signals	No user-side behavioral interaction signals in operational detection	DR1: Capture behavioral signals during email evaluation
Human susceptibility and cognitive error	Rich theoretical accounts; dispositional and situational factors established	Cognitive constructs not operationalized as deployable ML features	DR2: Derive features from HET, DPT, and PMT
Behavioral interaction analytics	Emerging evidence; Lain et al. (2022); Gallo et al. (2024)	Centralized collection; atheoretical, post-hoc feature engineering	DR3: Ensure raw behavioral logs remain on client device
Privacy-preserving ML for cybersecurity	FL and DP established; backdoor and poisoning attacks documented	No application to behavioral phishing; no adversarial threat model	DR4: Implement distributed DP with formal adversarial threat model
Nigerian and African digital context	Growing recognition of contextual risk; NDPA 2023 enacted; digital payment adoption expanding	No empirical behavioral study; no artifact designed for NDPA 2023	DR5: Align governance layer with NDPA 2023 requirements

Note. HET = Human Error Theory; DPT = Dual-Process Theory; PMT = Protection Motivation Theory; NDPA = Nigeria Data Protection Act; FL = Federated Learning; DP = Differential Privacy; ML = Machine Learning.

III. DESIGN SCIENCE RESEARCH METHODOLOGY

The paper follows the DSR process model of Peffers et al. (2007), which specifies six sequential activities: problem identification and motivation, definition of solution objectives, design and development of the artifact, demonstration, evaluation, and communication [8]. This manuscript addresses the first three activities; demonstration, empirical evaluation, and practitioner communication are reserved for the companion study. Partial-stage DSR is accepted when the artifact specification and ex ante evaluation constitute the primary scholarly contribution [9]. Following Gregor and Hevner's (2013) knowledge contribution taxonomy, the paper contributes an Improvement artifact, Nascent Theory through the construct-to-feature mapping, and Method through the replicable evaluation protocol [9].

Table 2 Design Science Research Application in This Paper

DSR Stage (Peffers et al., 2007)	Application in This Paper	Evidence in This Manuscript
Problem identification and motivation	Mainstream detection misses behavioral interaction signals; behavioral monitoring creates privacy and regulatory risks	Section II gap synthesis (Table 1); Section I problem statement
Objectives of a solution	Detect phishing-risk behavior through theory-grounded behavioral features while protecting privacy via distributed FL and DP, in alignment with NDPA 2023	Section III design objectives (DO1 to DO5)

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

Design and development	Develop a four-layer artifact: behavioral feature model, extraction pipeline, FL and DP architecture, governance layer	Section V artifact specification (Tables 4 to 9)
Demonstration (planned)	Specify use in a browser-based email simulation with Nigerian email users at research scale	Section VI.C participant design
Evaluation (ex ante, this paper)	Internal design-rationale assessment: theoretical coherence, design requirement traceability, architectural consistency, regulatory risk alignment	Section III.C and Table 3
Evaluation (ex post, companion study)	Predictive utility, privacy guarantee testing, usability, regulatory compliance through independent expert review and empirical study	Section VI evaluation protocol
Communication	Present artifact to cybersecurity, HCI, privacy-preserving ML, and African digital governance communities	This manuscript and planned companion empirical paper

A. Design Objectives

Five design objectives govern the artifact specification. DO1 through DO5 map to design requirements DR1 through DR5 derived from Table 1. Three additional artifact-quality requirements, DR6 through DR8, address compliance readiness, empirical comparability, and interpretability, and reflect standard DSR artifact quality criteria [7].

- *DO1 (DR1):*

Incorporate user-side email interaction signals absent from traditional content-, URL-, and metadata-based detection systems.

- *DO2 (DR2):*

Derive behavioral features from established cognitive and motivational theories of human error rather than from atheoretical post-hoc feature selection.

- *DO3 (DR3 and DR4):*

Ensure raw behavioral logs remain on client devices by implementing distributed FL with participant-level DP and a formal adversarial threat model.

- *DO4 (DR4, DR6 to DR8):*

Specify an evaluation-ready architecture testable against centralized, non-private, and static-feature baselines with theoretically interpretable outputs.

- *DO5 (DR5):*

Align the artifact with NDPA 2023 principles of consent, data minimization, purpose limitation, storage limitation, data subject rights, and DPIA.

B. Knowledge Contribution Type

Following Gregor and Hevner's (2013) taxonomy, the paper contributes an Improvement artifact: a novel solution for a problem established in the literature but unresolved in design. The second contribution is Nascent Theory: the construct-to-feature mapping operationalizes cognitive vulnerability as computational behavioral indicators. The third is a Method: the evaluation protocol provides a replicable template for behavioral phishing artifact research in African digital economy contexts.

C. Ex Ante Artifact Evaluation

Although empirical deployment is reserved for the companion study, this paper conducts an internal ex ante evaluation through four mechanisms. The ex ante evaluation is a design-rationale assessment conducted by the authors; independent validation is reserved for the expert review panel, prototype testing, and empirical study specified in Section VI.

The first mechanism is theoretical coherence assessment: each behavioral feature in Table 7 is derived from at least one theoretically motivated construct and verified against Table 5. The second is design requirement traceability: all artifact components are mapped against DR1 through DR8 in Tables 4 and 5, confirming coverage. The third is architectural consistency analysis: the threat model, DP mechanism, and secure aggregation design are evaluated for internal consistency, including identification and resolution of the secure aggregation and anomaly detection conflict. The fourth is regulatory risk alignment: the governance layer in Table 9 is assessed against NDPA 2023 provisions covering consent, minimization, purpose limitation, data subject rights, DPIA, and cross-border transfer.

Table 3 presents the ex ante evaluation outcomes. The evaluation is positive but conditional: every requirement is traceable, every feature is theoretically justified, and the governance layer addresses the principal NDPA 2023 safeguards. Residual risks in prototype performance, legal sufficiency, privacy-utility trade-off, and user acceptability are appropriately deferred to ex post evaluation.

Table 3 Summary of Ex Ante Artifact Evaluation Outcomes

Ex Ante Criterion	Evidence Used	Outcome	Residual Issue (Deferred to Ex Post)
Theoretical coherence	Table 5 construct-to-feature matrix; Figure 1 logic model	Conditionally passed: every feature maps to at least one HET, DPT, PMT, or HCI construct	Requires independent expert review by five-rater panel (Section VI.A)
Design requirement traceability	Tables 1, 4, 5; DR1 to DR8 coverage	Passed: each DR maps to at least one artifact component; DR6 to DR8 justified as quality criteria	Requires independent validation and prototype implementation

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

Architectural consistency	Sections V.D to V.E; deployment mode specification	Conditionally passed: DP privacy unit defined; secure aggregation and robust aggregation conflict identified and resolved through two deployment modes	Requires prototype implementation and FL configuration testing
Regulatory risk alignment	Table 9; NDPA 2023 provisions	Conditionally passed: NDPA-aligned safeguards specified for all major processing activities	Requires Nigerian data-protection counsel verification before any data collection

Note. HET = Human Error Theory; DPT = Dual-Process Theory; PMT = Protection Motivation Theory; HCI = Human-Computer Interaction; NDPA = Nigeria Data Protection Act; DP = Differential Privacy; DR = Design Requirement; FL = Federated Learning. 'Conditionally passed' indicates the design satisfies the criterion as specified but confirmation requires independent evaluation.

IV. THEORETICAL FOUNDATION FOR ARTIFACT DESIGN

Three theoretical traditions underpin the artifact's behavioral feature layer. Each theory contributes specific, construct-level predictions that are translated into measurable behavioral indicators in Table 5. Features are selected not merely for predictive potential but for their capacity to represent theoretically meaningful behavioral evidence of error type, cognitive processing mode, or protective appraisal.

A. Human Error Theory

Reason's (1990) taxonomy identifies three error types with direct relevance to phishing susceptibility [34]. Slips are attentional failures at the execution stage: the user intends a safe action but an automatic, habitual response triggered by a contextual cue overrides that intention. Phishing attacks engineer slips through interface mimicry. Lapses are memory failures: the user forgets to perform a security-relevant verification step. Mistakes are knowledge-based failures where the user applies an incorrect schema, such as assuming that a visually familiar institutional email is legitimate. Each error type predicts a distinct interaction signature: slip-driven susceptibility produces short latency and minimal pre-click inspection; lapse-driven susceptibility produces incomplete navigation sequences; mistake-driven susceptibility produces confident, non-hesitant misdirected action.

B. Dual-Process Theory

Kahneman's (2011) dual-process framework contrasts two processing modes [35]. System 1 is fast, automatic, and pattern-driven; it operates with minimal cognitive effort and is readily manipulated through familiarity, emotional priming, and salience. System 2 is deliberate, effortful, and rule-based; it can detect deceptive cues when sufficiently engaged. Phishing attacks trigger System 1 and suppress System 2 through urgency framing, authority cues, and visual mimicry. System 2 activity produces longer hover durations, higher link inspection ratios, and more deliberate navigation; System 1 dominance produces rapid, low-inspection clicks with low hover entropy.

C. Protection Motivation Theory

Rogers's (1975) Protection Motivation Theory specifies two appraisal processes governing protective behavior [36]. Threat appraisal is the user's judgment of how serious and probable a threat is. Coping appraisal is the user's judgment of whether an effective protective response is available and personally feasible. Users with high threat awareness and high coping self-efficacy engage in deliberate verification behavior. Users who do not perceive the threat or who lack confidence in their evaluation ability perform minimal inspection regardless of actual threat level. PMT adds a motivational explanation that neither HET nor DPT provides, accounting for why users with equivalent cognitive capacity show dramatically different levels of protective inspection [37].

D. Theoretical Complementarity and Non-Redundancy

Each theory serves a distinct analytical function. HET *classifies the failure mechanism*: it identifies whether the behavioral signature reflects a slip, a lapse, or a knowledge-based mistake. DPT *explains the cognitive processing pathway*: it predicts whether the user is likely to produce a rapid automatic response or a slower deliberate one. PMT *accounts for the motivational appraisal* governing whether protective verification is even attempted. No single theory provides all three functions, and together they form a causal chain: a user who lacks threat awareness (PMT) does not engage System 2 (DPT), which increases the probability of attentional slips (HET).

E. Theory-to-Artifact Logic Model

Figure 1 presents the theory-to-artifact logic model, showing the derivation chain from theoretical constructs through behavioral signals, privacy-preserving learning, and governance to the risk score output.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

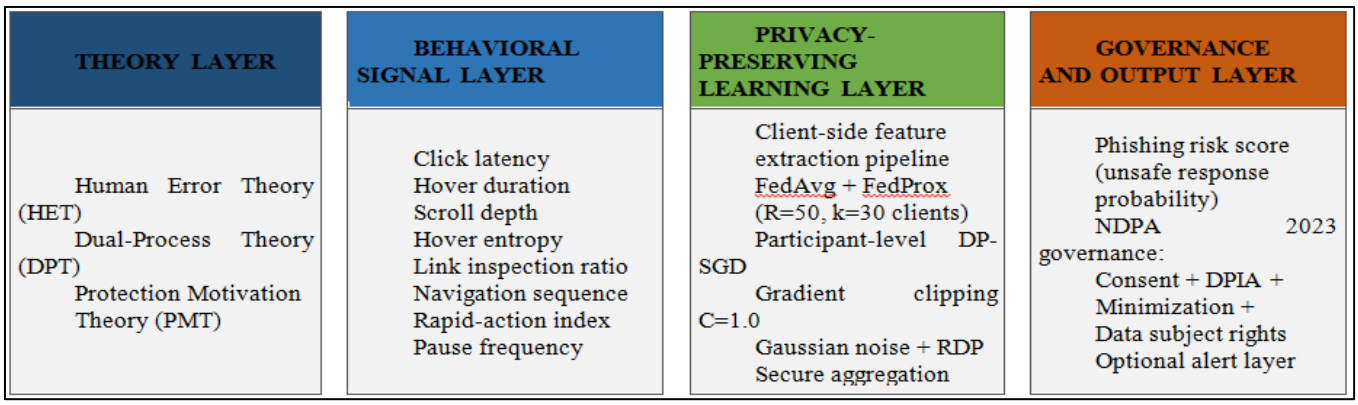

Fig 1 Theory-to-Artifact Logic Model for the Privacy-Preserving Behavioral Phishing Detection Framework

- Note. The alert/intervention layer is an optional extension of the detection artifact evaluated separately from detection performance. Governance (NDPA 2023) regulates all pipeline stages from data collection through model output. All privacy statements are design-level specifications until verified by the implemented RDP accountant and training code.

F. Construct-to-Feature Mapping

Table 4 presents the systematic derivation of behavioral features from theoretical constructs.

Table 4 Theory-to-Feature Operationalization Matrix

Theory	Construct	Behavioral Manifestation	Behavioral Feature	Susceptibility Direction
Dual-Process Theory [35]	System 1 processing speed	Rapid click response with minimal deliberation before action	Click latency (ms)	Lower value, higher susceptibility
Human Error Theory [34]	Attentional slip	Premature action before completing link review; failure to inspect sender address field	Time to first click (ms); sender hover duration (ms)	Shorter duration, higher susceptibility
Human Error Theory [34]	Memory lapse	Incomplete interaction sequence; failure to return and verify email sections	Navigation completeness score; backtracking count	Lower value, higher susceptibility
PMT [36]	Threat appraisal engagement	Extended reading of email body; cursor activity over threat-relevant content	Scroll depth (%); pause frequency (count per minute)	Lower value, higher susceptibility
PMT [36]	Coping appraisal and self-efficacy	Active inspection of embedded links and sender address before action	Link inspection ratio [0,1]; link preview activation (binary)	Lower value, higher susceptibility
Dual-Process Theory [35]	Analytical (System 2) engagement	Deliberate cursor movement patterns; revisitation of suspicious email elements	Hover entropy (bits); cursor path irregularity index	Lower value, higher susceptibility
Human Error Theory [34]	Knowledge-based mistake	Confident, non-hesitant misdirected action; no observable uncertainty signals	Rapid-action index [0,1]; hesitation event count	Higher RAI, higher susceptibility
HCI decision behavior	Attention allocation	Cursor fixation distribution across header, body, link, and footer regions	Hover entropy across email regions	Concentration on non-informative regions, higher susceptibility

Note. HET = Human Error Theory; DPT = Dual-Process Theory; PMT = Protection Motivation Theory; HCI = Human-Computer Interaction; RAI = Rapid-Action Index. Full mathematical definitions are in Section V.B. Susceptibility direction states the theoretically hypothesized relationship between feature value and unsafe phishing response.

The features in Table 4 are probabilistic proxies, not direct measurements of cognitive states. Click latency may reflect device response lag, motor ability differences, or habitual interaction style. Hover entropy and scroll depth may be influenced by screen size, input device type, accessibility technology use, or network-induced rendering delays. These confounds call for empirical sensitivity analysis and subgroup comparisons rather than abandonment of the features.

V. ARTIFACT DESIGN SPECIFICATION

A. Artifact Overview and Design Layers

The artifact is organized as a layered architecture in which each layer addresses one or more design requirements. Table 5 presents the layers, contents, design requirement mappings, and specification status. All layers are described as *architecturally specified*, indicating design is complete and evaluation-ready but prototype implementation is pending.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

Table 5 Artifact Design Layers, Contents and Specification Status

Artifact Layer	Contents	Design Requirement	Specification Status	Section Reference
Conceptual layer	Theory-to-feature mapping (Table 4) and logic model (Figure 1)	DR2	Architecturally specified	Sections IV.E to IV.F
Data layer	Behavioral event logging; client-side feature extraction pipeline; 12-feature taxonomy (Table 7)	DR1, DR3	Architecturally specified	Section V.B
Learning layer	ML model configurations; FL setup; DP parameterization; baselines; prediction target hierarchy (Table 8)	DR4, DR7	Architecturally specified	Sections V.C to V.D
Security layer	Adversarial threat model; distributed DP; secure aggregation; robust aggregation; out-of-scope boundary	DR4	Architecturally specified	Section V.E
Governance layer	Consent architecture; DPIA; pseudonymization; minimization; retention; data subject rights; controller and processor roles (Table 9)	DR5, DR6	Architecturally specified	Section V.F
Evaluation layer	Success criteria; six evaluation types; pre-specified validation protocol (Tables 10, 11)	DR7, DR8	Specified for empirical validation	Section VI

Note. DR = Design Requirement; DPIA = Data Protection Impact Assessment.

Table 6 Design Requirements for the Privacy-Preserving Behavioral Phishing Detection Artifact

ID	Design Requirement	Source	Artifact Response
DR1	Capture user-side behavioral signals preceding phishing compromise	Literature gap (Table 1, Row 1)	12 behavioral features derived from email client interaction logs (Table 7)
DR2	Ground behavioral features in established cognitive theory	Literature gap (Table 1, Row 2)	HET, DPT, and PMT construct-to-feature mapping (Table 4)
DR3	Avoid centralization of raw behavioral logs	Literature gap (Table 1, Row 3)	Client-side feature extraction; only feature vectors transmitted via secure aggregation
DR4	Provide formal privacy protection beyond FL alone	Literature gap (Table 1, Row 4)	Distributed DP with participant-level Gaussian noise, gradient clipping, RDP accounting, and secure aggregation; adversarial threat model in Section V.E
DR5	Align with NDPA 2023 regulatory obligations	Literature gap (Table 1, Row 5)	Governance layer covering consent, minimization, purpose limitation, DPIA, and data subject rights (Table 9)
DR6	Support regulatory compliance documentation	DSR artifact quality criterion	DPIA, consent form, retention schedule, and access-control plan documented before empirical study begins
DR7	Enable empirical comparison against baselines	DSR artifact quality criterion	Five comparison configurations; participant-level train-test split enforced
DR8	Preserve interpretability for security practitioners	DSR artifact quality criterion	SHAP-based feature attribution linked to theory-derived constructs from Table 4

Note. DR = Design Requirement; HET = Human Error Theory; DPT = Dual-Process Theory; PMT = Protection Motivation Theory; FL = Federated Learning; DP = Differential Privacy; DPIA = Data Protection Impact Assessment; NDPA = Nigeria Data Protection Act; SHAP = SHapley Additive exPlanations; RDP = Renyi Differential Privacy.

B. Behavioral Feature Layer

The feature extraction pipeline operates entirely on the client device, capturing interaction events at 50 ms temporal resolution using browser-native APIs including MouseEvent, PointerEvent, ScrollEvent, and FocusEvent. All 12 features are computed on the client immediately after the interaction session ends. Raw event logs are deleted within the same session, satisfying DR3. Standardization parameters for all continuous features are estimated from the training partition only and applied to validation and test partitions to prevent information leakage across splits.

➢ *Normative Verification Sequence.*

The sequence deviation index relies on a normative verification pathway established through expert consensus involving at least five raters. The normative pathway specifies four steps: (1) sender address inspection, considered complete when hover or focus on the sender field is sustained for at least 300 ms; (2) link preview activation, considered complete when a preview event is logged or hover over an embedded link is sustained for at least 500 ms; (3) email body reading, considered complete when scroll depth reaches at least 60 percent or dwell time in the body region is at least 10 seconds; and (4) action decision, recorded as a click, delete, forward, or report event. These thresholds are not theoretical constants but operational starting points subject to pilot calibration.

➢ *Rapid-Action Index (RAI).*

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

Let z(H) denote the standardized hesitation event count and z(D) denote the standardized decision time, both computed from training-partition statistics. The RAI is defined as: RAI = sigmoid[−z(H) − z(D)], where sigmoid(x) = 1 / (1 + e^−x). Higher RAI values indicate fewer hesitations and shorter decision time, both theoretically associated with higher susceptibility. Equal weighting of hesitation count and decision time is an initial design assumption; alternative formulations will be tested during sensitivity analysis.

- *Hover Entropy.*

Hover entropy is computed as $H = -\sum p_i \log_2(p_i)$, where p_i is the proportion of total cursor hover dwell time spent in email region i. Regions are defined as: sender and header area, email body text, embedded link areas, and footer. Higher entropy indicates more evenly distributed attention; lower entropy indicates concentration on a subset of regions, which is theoretically associated with superficial evaluation.

Table 7 Behavioral Feature Taxonomy: Operational Definitions, Theoretical Linkage and Units

Feature	Operational Definition	Unit	Theoretical Construct	Susceptibility Direction	Primary Target
Click latency	Time from email open to first link click; z-scored from training partition	ms	System 1 speed (DPT); attentional slip (HET)	Lower, higher susceptibility	Unsafe response
Sender hover duration	Total cursor dwell time over sender name and address field; z-scored from training partition	ms	Threat appraisal (PMT); System 2 (DPT)	Lower, higher susceptibility	Unsafe response
Link inspection ratio	Unique hover events over embedded links divided by total actionable links in the email	Ratio [0,1]	Coping appraisal self-efficacy (PMT)	Lower, higher susceptibility	Unsafe response
Scroll depth	Maximum percentage of email body scrolled before final action or exit	Percent [0,100]	Threat appraisal engagement (PMT)	Lower, higher susceptibility	Unsafe response
Backtracking count	Number of upward scroll reversals and re-hover events on previously inspected elements	Count	Memory lapse recovery (HET)	Lower, higher susceptibility	Unsafe response
Hover entropy	H = -sum(p_i * log2(p_i)) of cursor dwell time distribution across header, body, link, and footer regions	Bits	Attention allocation (HCI and DPT)	Lower, higher susceptibility	Unsafe response
Sequence deviation index	Edit distance between observed and normative sequence, normalized by maximum possible distance; expert-weighted variant also computed	Normalized [0,1]	Knowledge-based mistake (HET)	Higher, higher susceptibility	Unsafe response
Pause frequency	Count of cursor-stationary events (velocity below 5 px/s, duration above 500 ms) per minute; thresholds are pilot-calibrated design values	Count/min	Deliberative inhibition (DPT)	Lower, higher susceptibility	Unsafe response
Decision time	Total time from email open to final action: click, delete, forward, or report; z-scored from training partition	Seconds	Processing depth (DPT and HET)	Lower, higher susceptibility	Unsafe response
Rapid-action index	sigmoid[-z(H) - z(D)]; z-scores from training partition only; equal weighting is initial design assumption	Score [0,1]	Speed of misdirected action (HET)	Higher, higher susceptibility	Unsafe response
Link preview activation	Equals 1 if browser link preview was activated before clicking; equals 0 otherwise; environment-dependent feature	Binary [0,1]	Coping self-efficacy (PMT)	0, higher susceptibility	Unsafe response
Navigation completeness	Proportion of four expert-defined verification steps completed before final action; step-completion thresholds are pilot-calibrated values	Proportion [0,1]	Memory lapse avoidance (HET)	Lower, higher susceptibility	Unsafe response

Note. DPT = Dual-Process Theory; HET = Human Error Theory; PMT = Protection Motivation Theory; HCI = Human-Computer Interaction; RAI = Rapid-Action Index. z-scores are computed from training-partition statistics only and applied to validation and test partitions to prevent leakage.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

Table 8 specifies the prediction target hierarchy. The behavioral artifact is primarily a user-action risk detector, not an email artifact classifier. Email-level phishing labels are used to construct baselines and contextualize interaction outcomes; the behavioral model targets the user decision process. All behavioral and multimodal models will be evaluated at the participant-email interaction level.

Table 8 Prediction Target Hierarchy for the Empirical Companion Study

Prediction Target	Label	Unit of Analysis	Notes
Email artifact classification	Phishing vs. legitimate	Email stimulus	Used for content-only and metadata-only baseline models only
Behavioral risk classification (primary)	Safe vs. unsafe user response	Participant by email interaction	Primary target for behavioral and multimodal models; resolves unit-of-analysis ambiguity
Response severity	No click / click / credential entry / report	Participant by email interaction	Ordinal extension for severity analysis in companion study
Time to unsafe action	Seconds from email open to first unsafe click event	Participant by email interaction	Continuous regression sub-analysis for latency-focused questions

Note. The primary prediction target is unsafe phishing response at the participant-email interaction level. The behavioral artifact is a user-action risk detector, not a mail classifier.

C. Machine Learning Architecture

Model development follows a comparison architecture isolating the contribution of behavioral features over established feature sets. Five configurations are specified: (1) majority class baseline; (2) content-only model with TF-IDF and n-gram features using Logistic Regression, Random Forest, and XGBoost; (3) metadata-only model with email header features using the same classifiers; (4) behavioral-only model with the 12 features from Table 7 using the same classifiers plus a Gradient Boosted model with SHAP attribution; and (5) multimodal combined model concatenating all three feature sets. A two-hidden-layer neural network is included as an exploratory comparator. Class imbalance will be addressed through class-weighted loss functions, threshold tuning on the validation set, and PR-AUC reporting alongside ROC-AUC. Participant-level train-test splitting prevents behavioral signature leakage across splits.

D. Federated Learning and Differential Privacy

The FL framework is configured as a cross-device system in which each participant's browser session acts as an independent client. The planned empirical study simulates a cross-device FL configuration at research scale; production-scale cross-device FL would require substantially larger client populations and additional communication-efficiency evaluation. Aggregation uses FedAvg [20] with FedProx regularization [38] to manage non-IID behavioral data heterogeneity. In each of R = 50 communication rounds, the server broadcasts the current global model to k = 30 randomly sampled clients (10 percent sampling rate). Each client performs E = 3 local epochs with batch size B = 16 and learning rate 0.01, then transmits clipped and noise-perturbed gradient updates. FL hyperparameters are initial design values subject to pilot evaluation.

The artifact adopts a distributed differential privacy design: client updates are clipped locally at gradient clipping threshold C = 1.0 (L2-norm), calibrated Gaussian noise is added at the client before transmission, secure aggregation [28] prevents the server from seeing individual client updates, and privacy loss is tracked at the participant/client level using a Renyi Differential Privacy (RDP) accountant. All privacy statements are design-level specifications until verified by the implemented RDP accountant and training code. Noise multiplier values sigma in {0.5, 1.0, 2.0, 4.0} correspond to provisional design target privacy budgets epsilon in {16, 8, 4, 1} at delta = 10^(−5), following common practice that delta should be smaller than the inverse of the population size. Epsilon = 8 is treated as a mid-range evaluation point, not a deployment recommendation. Because participant-level DP with approximately 300 clients may impose substantial utility costs, the evaluation will report sensitivity analyses across privacy budgets, sampling rates, and aggregation mechanisms.

E. Privacy Threat Model

FL alone does not constitute a privacy guarantee [25]. Three threat types involving model and gradient exploitation are worth distinguishing: gradient leakage concerns reconstruction of training data from intermediate gradient updates; model inversion concerns reconstruction or inference from the trained final model; and membership inference concerns determining whether a particular participant contributed to the training set.

- *Honest-But-Curious Server.*

Defense: distributed DP-SGD ensures transmitted updates are (ε, δ)-differentially private at the participant level; secure aggregation prevents the server from accessing individual updates.

- *Gradient Leakage Attacks [25].*

Defense: gradient clipping bounds sensitivity; Gaussian DP noise provides formal protection; secure aggregation removes direct access to individual update vectors.

- *Membership Inference Attacks [26].*

Defense: participant-level DP bounds the maximum information gain about any individual's participation to exp(ε) with probability 1 minus δ.

- *Poisoning and Backdoor Attacks [27], [29].*

Bagdasaryan et al. (2020) showed that backdoor attacks can evade norm-based screening [29]; norm-based filtering is a preliminary screen only. The evaluation must compare

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

FedAvg against robust aggregation alternatives including trimmed mean, coordinate-wise median, and Krum-style selection [30], [31]. Two deployment modes address the conflict between robust aggregation and secure aggregation: (a) high-security mode using full secure aggregation without individual-update inspection, and (b) diagnostic mode without secure aggregation, used during controlled evaluation phases.

- *Model Inversion Attacks.*

Defense: DP training limits any individual participant's influence on the model; model inversion resistance should be empirically tested during prototype evaluation.

- *Out-of-Scope Threats.*

The artifact does not claim protection against compromised client devices, malicious browser extensions, endpoint malware, or adversaries with direct physical access to user screens or credentials. Operational deployment would require signed browser extensions, least-privilege permissions, enterprise extension management, and endpoint protection.

F. Governance and Regulatory Compliance Layer

The governance layer ensures data collection, model training, and deployment procedures align with NDPA 2023 [6]. The artifact uses pseudonymization, not anonymization: behavioral interaction signatures may remain re-identifying even after direct personal identifiers are removed. The dataset is never treated as fully anonymous; public reporting uses aggregate statistics only. The empirical study must formally define whether the university, partner organizations, and FL coordinator act as controllers, joint controllers, or processors under NDPA 2023. Consent records, DPIA documentation, and audit logs may require separate, longer retention periods for accountability and will be documented independently from behavioral feature retention schedules.

Table 9 Governance Requirements and NDPA 2023 Alignment

Governance Issue	NDPA 2023 Provision	Design Response	Implementation Note
Lawful basis	Section 25: consent or legitimate interest required	Explicit written informed consent as the primary lawful basis; potential legitimate cybersecurity research interest assessed as supplementary basis by legal counsel	Lawful basis confirmed by Nigerian data-protection counsel; legitimate interest not assumed automatically
Data minimization	Section 26: collect only data necessary for the stated purpose	Only 12 specified behavioral features extracted; raw interaction logs deleted within the session	Client-side pipeline only; no raw log transmission to any server
Purpose limitation	Section 26: data may not be used beyond the stated purpose	Behavioral data used exclusively for phishing detection model training and evaluation	Contractual prohibition on reuse for productivity monitoring, employee surveillance, or any other purpose
Storage limitation	Section 26: personal data must not be retained beyond the necessary period	Feature vectors retained only for the model training period and deleted after model convergence; retention period not to exceed six months after study completion	Consent records, DPIA documentation, and audit logs may require longer separate retention for accountability and will be documented independently
Pseudonymization	Section 26: appropriate technical safeguards required	Participant identifiers replaced with randomly generated IDs before feature extraction; no linking information retained thereafter	Dataset is pseudonymized, not anonymous; aggregate-level reporting only in any publication
Participant withdrawal	Sections 34 to 40: data subject rights include the right to withdraw	Participants may withdraw before model aggregation; post-aggregation withdrawal is technically limited because the FL update has already been incorporated into the global model	Consent form must explain this limitation; machine unlearning feasibility assessed [39]; full model retraining offered if unlearning is infeasible
Data subject rights	Sections 34 to 40: access, correction, deletion, objection, and portability	Rights documented in consent form; deletion request before aggregation triggers immediate feature-vector removal	A rights officer is appointed; all procedures documented in the DPIA
DPIA	Section 43: required for high-risk processing activities	DPIA conducted and fully documented before IRB submission and data collection	DPIA documents processing risks, mitigation measures, and residual risks; reviewed by an independent assessor
Cross-border transfer and cloud hosting	Section 43: transfers require adequate protection	Model updates remain within Nigeria during the study; any FL coordinator on cloud infrastructure must be region-locked to Nigerian or legally approved infrastructure with a transfer-impact assessment	Architecture reviewed before any cloud-based deployment

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

Controller and processor roles	Section 26: accountability for data processing	The empirical study must formally define whether the university, partner organizations, and FL coordinator act as controllers, joint controllers, or processors under NDPA 2023	Controller and processor role definition documented in the DPIA and organizational agreements before data collection
Function creep prevention	Section 26: purpose limitation	Behavioral data contractually prohibited from use beyond phishing detection research	Prohibition documented in consent forms and organizational partnership agreements
Access control	Section 26: appropriate security measures required	Feature data accessible only to named research team members; all access events logged	Role-based access control; audit logs reviewed at regular intervals
Trained model as derivative personal data	Section 26: personal data definition encompasses derived data	The trained FL model may encode residual individual influence despite DP; model release requires a formal privacy audit	Model not publicly released without documented DP guarantees and independent review

Note. NDPA = Nigeria Data Protection Act; DPIA = Data Protection Impact Assessment; FL = Federated Learning; DP = Differential Privacy; IRB = Institutional Review Board. Section references must be verified by Nigerian data-protection counsel before submission or deployment.

VI. PROPOSED EVALUATION PROTOCOL

A. Artifact Evaluation Strategy

Following Peffers et al. (2007), artifact evaluation assesses the extent to which the designed solution addresses the identified problem [8]. Table 10 presents six evaluation types, each targeting a distinct quality dimension. Detection performance and alert intervention effectiveness are reported separately to avoid conflating the artifact's risk-scoring capacity with the behavioral effect of warning messages.

Table 10 Artifact Evaluation Strategy: Types, Evaluative Questions and Planned Evidence

Evaluation Type	Evaluative Question	Method	Planned Evidence
Theoretical evaluation	Is each behavioral feature derived from at least one defensible cognitive or HCI construct?	Expert review of Table 4 by at least five raters: two cybersecurity researchers, one HCI researcher, one cognitive psychologist, and one privacy specialist; interrater agreement calculated	Construct-validity agreement rate; documented resolution of any disagreements in the revised mapping
Technical evaluation	Can the artifact be implemented without materially degrading email-client usability on desktop and laptop environments?	Prototype implementation; runtime profiling; interaction-logging reliability testing	CPU utilization, memory consumption, browser latency, and event-logging accuracy; usability thresholds adjusted from pilot profiling and accepted HCI latency norms
Predictive evaluation	Does behavioral data provide incremental phishing detection value over content-only and metadata-only baselines?	Controlled experiment with at least 300 Nigerian email users; five-configuration model comparison; participant-level train-test split; class imbalance handled through weighted loss and threshold tuning	F1-score, ROC-AUC, PR-AUC, FPR, and FNR with 95% bootstrap confidence intervals; McNemar test and DeLong test with Bonferroni correction; SHAP attribution linked to Table 4
Privacy evaluation	Does distributed FL with participant-level DP reduce exposure of raw behavioral logs relative to centralized collection?	RDP accounting; simulated gradient inversion, membership inference, and backdoor attacks; FedAvg comparison against robust aggregation alternatives under diagnostic mode	Final epsilon and delta from RDP accountant; attack success rates; privacy-utility trade-off curve across epsilon values 1, 4, 8, and 16
Regulatory evaluation	Does the artifact design support alignment with NDPA 2023 obligations?	DPIA completion; controller and processor role definition; consent flow review; legal review by a Nigerian data-protection specialist	DPIA documentation; legal opinion on NDPA 2023 alignment; consent form approval before data collection
Usability	Do real-time behavioral risk	User study on alert	Warning comprehension rate; false-

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

evaluation	alerts improve user phishing decisions without creating warning fatigue?	comprehension and false-positive tolerance; alert frequency sensitivity testing; warning habituation measurement [40]	positive annoyance threshold; alert fatigue measure; safe-action improvement rate. Alert delivery is an optional intervention layer evaluated separately from detection performance.

Note. HCI = Human-Computer Interaction; RDP = Renyi Differential Privacy; DPIA = Data Protection Impact Assessment; NDPA = Nigeria Data Protection Act; SHAP = SHapley Additive exPlanations; FL = Federated Learning; DP = Differential Privacy; PR-AUC = Precision-Recall Area Under the Curve; ROC-AUC = Receiver Operating Characteristic Area Under the Curve.

Table 11 Pre-Specified Artifact Success Criteria for the Empirical Companion Study

Criterion	Minimum Expectation	Assessment Method
Theoretical coherence	Each behavioral feature maps to at least one defensible cognitive or HCI construct with documented justification; no feature included without a derivation in Table 4	Expert review by five-rater panel; interrater agreement on construct-to-feature matrix; consistency with Figure 1 logic model
Privacy preservation	Raw behavioral logs remain on the client device in both deployment modes; no raw logs transmitted to any server	Architecture review; code audit of all data transmission paths in both high-security and diagnostic modes
Formal privacy guarantee	DP configuration specifies privacy unit, epsilon, delta, gradient clipping threshold, noise mechanism, accountant type, and composition details; final epsilon computed from RDP accountant	RDP accounting verification before model training; sensitivity analysis across epsilon values 1, 4, 8, and 16
Technical feasibility	Prototype operates on desktop and laptop environments without causing statistically significant degradation in user-perceived interface responsiveness relative to the baseline simulation	Pilot profiling on target hardware; user-perceived latency comparison; thresholds established from pilot data
Predictive utility	Behavioral model F1-score, ROC-AUC, and PR-AUC are statistically compared against all four baseline configurations; McNemar test and DeLong test with Bonferroni correction; evaluation at participant-email interaction level	Peer-reviewed statistical analysis; 95% bootstrap confidence intervals; participant-level splits strictly enforced; PR-AUC reported alongside ROC-AUC
Interpretability	SHAP feature importance rankings linked to theory-derived constructs from Table 4; results interpreted in terms of HET, DPT, and PMT constructs	SHAP summary plots; theory-linked narrative in companion paper
Compliance readiness	DPIA, consent architecture, minimization procedures, retention schedule, controller and processor role definitions, data subject rights procedures, and machine unlearning feasibility assessment fully documented and reviewed by legal counsel before IRB submission	DPIA completed before IRB submission; legal opinion documented; controller and processor roles defined
Robustness and fairness	Artifact tested under non-IID behavioral data, simulated backdoor and poisoning attacks, device heterogeneity, accessibility technology conditions, and minority-group performance; fairness evaluation by device type, demographic subgroup, and interaction style	Ablation analysis; attack simulation; subgroup F1-score and ROC-AUC; fairness metrics; accessibility technology mode tested separately

Note. HCI = Human-Computer Interaction; DP = Differential Privacy; DPIA = Data Protection Impact Assessment; RDP = Renyi Differential Privacy; CI = Confidence Interval; SHAP = SHapley Additive exPlanations; IRB = Institutional Review Board; NDPA = Nigeria Data Protection Act; PR-AUC = Precision-Recall Area Under the Curve.

B. Design Propositions

The following five design propositions replace empirical hypotheses. Each makes a theoretically justified directional claim while explicitly deferring empirical confirmation to the companion study.

- *P1 (DO1, DR1):*

Behavioral interaction signals are expected to provide incremental phishing detection value over content-only and metadata-only feature sets because they encode user decision-process information unavailable in static email artifacts. The direction and magnitude of that contribution is an empirical question to be determined in the companion study.

- *P2 (DO2, DR2):*

Features associated with rapid action, minimal inspection, and limited deliberation are theoretically associated with heuristic processing, attentional slips, and reduced coping appraisal, and are expected to be positively associated with unsafe phishing response. The relative importance of specific features must be established empirically through SHAP attribution analysis.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

- *P3 (DO3, DR4):*

A distributed FL architecture with participant-level DP is expected to reduce exposure of raw behavioral logs relative to centralized collection while incurring a privacy-utility trade-off in detection performance. Whether that trade-off is operationally acceptable at specific epsilon values is an empirical question.

- *P4 (DO5, DR5, DR6):*

The governance layer, as specified in Table 9 and designed for alignment with the NDPA 2023, addresses the lawful basis, consent, minimization, purpose limitation, DPIA, and data subject rights requirements. Whether the design achieves legal compliance must be confirmed through review by qualified Nigerian data-protection counsel before any deployment.

- *P5 (DO3, DR3, DR6):*

A behavioral phishing detection artifact combining local feature extraction, participant-level DP, secure aggregation, and explicit function-creep restrictions is expected to have greater institutional acceptability than centralized behavioral monitoring architectures. Empirical confirmation requires measurement of user trust and perceived surveillance risk in the usability evaluation.

C. Participant Design for Empirical Validation

The empirical validation involves adult Nigerian email users recruited from three sectors: university students, banking and financial services professionals, and civil service employees. Recruitment is planned through partnerships with two Nigerian universities, one commercial bank, and two government agencies. A minimum of N = 300 participants is specified based on a power analysis for logistic regression with 12 behavioral predictors, assuming medium effect size ($f^2 = 0.15$), power = 0.80, and $\alpha = .05$ [41]. A 20 percent attrition buffer raises the target to N = 360.

Inclusion criteria require participants to be aged 18 years or older, use email actively (at least five messages per day), have normal or corrected-to-normal vision, and pass all attention check items. Exclusion criteria include prior participation in phishing simulation research within the past 12 months, self-reported cybersecurity training within the past six months, and evidence of bot behavior in interaction timestamps. The initial validation is restricted to desktop and laptop computer environments; mobile and touchscreen environments will be evaluated in a later extension study. Participants who use assistive input devices will be analyzed separately rather than pooled with the main sample.

The email stimulus set comprises 40 emails per participant (20 phishing, 20 legitimate), calibrated across three sophistication levels and five phishing categories. The corpus includes emails adapted to Nigerian English and Nigerian Pidgin English communication conventions, with subgroup analysis testing whether behavioral features remain stable across language and literacy contexts. The normative verification sequence and stimulus set will be piloted with 20 participants before the main experiment.

VII. DISCUSSION

A. Theoretical Contribution

The construct-to-feature mapping in Table 4 is the paper's primary theoretical contribution. It establishes, within the reviewed literature, a systematic derivation chain from established cognitive and motivational theories to computationally deployable ML features for phishing detection, strengthening the theoretical grounding that remains underdeveloped in much prior behavioral phishing work. By deriving features from HET, DPT, and PMT constructs rather than selecting them post-hoc, the mapping enables model outputs to carry theoretical meaning. If SHAP analysis in the companion study assigns high importance to click latency and lower importance to navigation completeness, that result could be interpreted as preliminary evidence that attentional slip plays a stronger role than memory lapse in the study sample.

B. Methodological Contribution

The artifact design specification contributes a methodological template for the behavioral phishing detection field. The participant-level train-test split requirement, the 12 theoretically grounded and mathematically defined features in Table 7, the five-configuration ML comparison architecture with class-imbalance handling, and the four-layer FL and DP architecture with formal threat model together provide a replicable template. The specification of the RAI formula, hover entropy computation, step-completion thresholds, and sensitivity-testing requirements for pause thresholds and RAI weighting address methodological gaps that prior behavioral security studies have left unresolved.

C. Practical and Regulatory Contribution

The governance layer in Table 9 illustrates how privacy-by-design behavioral monitoring could be structured for alignment with NDPA 2023 obligations, subject to legal review. Whether the design achieves full legal compliance must be confirmed by qualified counsel before any operational deployment. The artifact is conceptually compatible with browser-based and enterprise email environments. Concrete deployment requires evaluation of browser-extension permissions, enterprise email-client APIs, organizational policy controls, and computational overhead, as well as signed browser extensions, least-privilege permissions, enterprise extension management, and endpoint protection.

D. Implications for the Nigerian and African Context

The behavioral feature set, anchored in universal cognitive theory, is designed to support cross-contextual adaptation, although transferability to other cultural, linguistic, and device environments must be empirically tested. Future research should extend the framework to other African jurisdictions with distinct regulatory contexts, including South Africa's Protection of Personal Information Act and Kenya's Data Protection Act 2019, to assess both governance transferability and behavioral feature validity across settings.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

VIII. LIMITATIONS AND RISK CONTROLS

Table 12 presents the artifact's acknowledged limitations alongside specified risk controls and future research directions.

Table 12 Artifact Limitations and Specified Risk Controls

Limitation	Risk Level	Risk Control and Future Research Direction
Simulated environment ecological validity: behavior in a known research context may differ from naturalistic inbox settings due to Hawthorne effect and reduced task stakes	High	Future research should replicate findings using longitudinal naturalistic monitoring with appropriate NDPA 2023 consent; partial-deception design with full post-study debriefing reduces awareness bias
Partial disclosure priming: recruitment disclosing email and cybersecurity context may elevate security-relevant cognition	Medium	Full deception with IRB-approved post-study debriefing eliminates this; DPIA must assess whether full deception is ethically permissible under NDPA 2023
Sample representativeness: sample limited to three sectors, excluding informal economy workers, rural populations, and feature-phone users	Medium	Future studies should extend to underrepresented populations; stratified recruitment across geographic and socioeconomic segments
Behavioral confounds: interaction features may reflect device type, motor ability, screen size, input device, or network latency rather than cognitive vulnerability	High	Sensitivity analyses controlling for device type and user demographics are required; device-stratified subgroup reporting should be standard
Accessibility technologies: screen readers, keyboard-only navigation, and assistive input devices may alter or eliminate hover- and cursor-based features	High	Accessibility-specific interaction modes must be evaluated separately; users whose patterns arise from assistive technology use must not be penalized
Adversarial behavioral mimicry: attackers aware of behavioral detection may mimic cautious interaction patterns to evade detection	Medium	Robustness to adversarial mimicry should be empirically tested; the artifact should be positioned as one detection layer in a multi-signal system
Non-IID FL heterogeneity: individual behavioral profiles create significant distribution heterogeneity that FedProx may not fully resolve	Medium	Comparison of FedProx against personalized FL approaches, including MAML-based meta-learning and per-FedAvg, should be included in the companion study
DP fairness disparity: noise injection may disproportionately degrade detection performance for users with underrepresented behavioral patterns	High	Fairness evaluation by device type, demographic group, and interaction style is required; disparate impact analysis across participant subgroups must be reported
Residual re-identification risk: behavioral interaction signatures may remain re-identifying after direct identifier removal	High	Dataset treated as pseudonymized, not anonymous; public release requires additional anonymization risk assessment and governance approval
Model drift: phishing tactics and user behavior change over time, reducing model effectiveness post-deployment	Medium	Periodic model recalibration, drift monitoring, and scheduled re-evaluation under privacy-preserving FL update rules should be part of any operational deployment plan
Warning and alert fatigue: real-time risk alerts may become ineffective when triggered too frequently [40]	Medium	Alert throttling, severity-based warning hierarchies, and false-positive threshold optimization required in usability evaluation
Withdrawal after aggregation: participant withdrawal after model aggregation may not fully remove individual data influence from the trained FL model	High	Consent forms must explain this technical limitation; machine unlearning feasibility assessed [39]; full model retraining offered if unlearning is infeasible
Language and localization: phishing susceptibility may depend on Nigerian English, Pidgin English, and local-language communication patterns and regional trust markers	Medium	Stimulus corpus should include localized variants; subgroup analysis should test whether behavioral features remain stable across language and literacy contexts

IX. CONCLUSION

This paper has proposed, justified, and specified a design science artifact for privacy-preserving behavioral phishing detection using email interaction signals. Three challenges converge in the problem the artifact addresses: human cognitive vulnerability continues to drive phishing success despite technical countermeasures; fine-grained behavioral interaction data carries information about that vulnerability that static email signals cannot capture; and collecting behavioral data at scale creates privacy risks requiring formal, legally grounded mitigation.

The paper makes three contributions to the reviewed literature. The first is the construct-to-feature mapping in Table 4 and Figure 1, which derives 12 behavioral ML

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

features from HET, DPT, and PMT, providing theoretical grounding that remains underdeveloped in much prior behavioral phishing work. The second is the distributed DP architecture in Sections V.D and V.E, which specifies the participant as the privacy unit, provides precise mathematical definitions for hover entropy and the RAI, acknowledges the secure aggregation and robust aggregation composition challenge, specifies two deployment modes, clarifies that all privacy statements are design-level until verified by the implemented RDP accountant, and acknowledges the utility costs that participant-level DP may impose with a modest client population. The third is the governance layer in Table 9, which covers NDPA 2023-aligned consent, minimization, purpose limitation, DPIA, data subject rights, controller and processor role definition, lawful basis caution, the technical limits of post-aggregation withdrawal, and the status of the trained model as potentially derivative personal data.

The artifact is architecturally specified and evaluation-ready. It does not claim specific detection performance, confirmed privacy thresholds, demonstrated legal compliance, or transferability to other cultural or device contexts. Those claims require the empirical evaluation, expert review, legal assessment, and prototype profiling specified in Section VI. That separation between artifact design and empirical validation reflects the design science research approach adopted here and forms the foundation of the paper's scholarly positioning.

REFERENCES

[1]. A. Vishwanath, B. Harrison, and Y. J. Ng, "Suspicion, cognition, and automaticity model of phishing susceptibility," Communication Research, vol. 45, no. 8, pp. 1146-1166, 2018.

[2]. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Systems with Applications, vol. 117, pp. 345-357, 2019.

[3]. Central Bank of Nigeria, Annual Report on Payment Systems Statistics 2023, CBN, 2024. [Online]. Available: https://www.cbn.gov.ng

[4]. N. Kshetri, "Cybercrime and cybersecurity in Africa," Journal of Global Information Technology Management, vol. 22, no. 2, pp. 77-81, 2019.

[5]. A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," Decision Support Systems, vol. 51, no. 3, pp. 576-586, 2011.

[6]. Nigeria Data Protection Commission (NDPC), Nigeria Data Protection Act 2023, Federal Republic of Nigeria, 2023. [Online]. Available: https://ndpc.gov.ng/ndp-act-2023/

[7]. A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quarterly, vol. 28, no. 1, pp. 75-105, 2004.

[8]. K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," Journal of Management Information Systems, vol. 24, no. 3, pp. 45-77, 2007.

[9]. S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," MIS Quarterly, vol. 37, no. 2, pp. 337-355, 2013.

[10]. J. Hong, "The state of phishing attacks," Communications of the ACM, vol. 55, no. 1, pp. 74-81, 2012.

[11]. K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," Expert Systems with Applications, vol. 106, pp. 1-20, 2018.

[12]. T. Peng, I. Harris, and Y. Sawa, "Detecting phishing attacks using natural language processing and machine learning," in Proc. 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp. 300-301, 2018.

[13]. C. I. Canfield, B. Fischhoff, and A. Davis, "Quantifying phishing susceptibility for detection and behavior decisions," Human Factors, vol. 58, no. 8, pp. 1158-1172, 2016.

[14]. R. T. Wright, M. L. Jensen, J. B. Thatcher, M. Dinger, and K. Marett, "Influence techniques in phishing attacks: An examination of vulnerability and resistance," Information Systems Research, vol. 25, no. 2, pp. 385-400, 2014.

[15]. R. B. Cialdini, Influence: The Psychology of Persuasion, Rev. ed. HarperCollins, 2007.

[16]. D. Lain, K. Kostiainen, and S. Capkun, "Phishing in organizations: Findings from a large-scale and long-term study," in Proc. 43rd IEEE Symposium on Security and Privacy, pp. 842-859, 2022.

[17]. D. Oliveira et al., "Dissecting spear phishing emails for older vs young adults," in Proc. 2017 CHI Conference on Human Factors in Computing Systems, pp. 6412-6424, 2017.

[18]. A. Vishwanath, "Examining the distinct antecedents of e-mail habits and its influence on the outcomes of a phishing attack," Journal of Computer-Mediated Communication, vol. 20, no. 5, pp. 570-584, 2015.

[19]. L. Gallo, D. Gentile, S. Ruggiero, A. Botta, and G. Ventre, "The human factor in phishing: Collecting and analyzing user behavior when reading emails," Computers & Security, vol. 139, 103671, 2024.

[20]. H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), PMLR 54, pp. 1273-1282, 2017.

[21]. P. Kairouz et al., "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol. 14, no. 1-2, pp. 1-210, 2021.

[22]. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3-4, pp. 211-407, 2014.

[23]. M. Abadi et al., "Deep learning with differential privacy," in Proc. 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308-318, 2016.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26may265

[24]. R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," arXiv preprint arXiv:1712.07557, 2017.
[25]. L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in Advances in Neural Information Processing Systems 32 (NeurIPS 2019), pp. 14774-14784, 2019.
[26]. M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning," in Proc. 2019 IEEE Symposium on Security and Privacy, pp. 739-753, 2019.
[27]. M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in Proc. 2018 IEEE Symposium on Security and Privacy, pp. 19-35, 2018.
[28]. K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc. 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175-1191, 2017.
[29]. E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in Proc. 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020), PMLR 108, pp. 2938-2948, 2020.
[30]. P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Proc. 31st Conference on Neural Information Processing Systems (NeurIPS 2017), pp. 119-129, 2017.
[31]. D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in Proc. 35th International Conference on Machine Learning (ICML 2018), PMLR 80, pp. 5650-5659, 2018.
[32]. Nigerian Communications Commission (NCC), Annual Report on Telecommunications Statistics 2023, NCC, 2023. [Online]. Available: https://www.ncc.gov.ng
[33]. U. J. Orji, "Cybersecurity law and regulation in Nigeria," Journal of Cybersecurity and Privacy, vol. 1, no. 1, pp. 1-14, 2018. [Source requires author verification]
[34]. J. Reason, Human Error, Cambridge University Press, 1990.
[35]. D. Kahneman, Thinking, Fast and Slow, Farrar, Straus and Giroux, 2011.
[36]. R. W. Rogers, "A protection motivation theory of fear appeals and attitude change," Journal of Psychology, vol. 91, no. 1, pp. 93-114, 1975.
[37]. T. Herath and H. R. Rao, "Protection motivation and deterrence: A framework for security policy compliance in organisations," European Journal of Information Systems, vol. 18, no. 2, pp. 106-125, 2009.
[38]. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Smola, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. Machine Learning and Systems 2020 (MLSys 2020), vol. 2, pp. 429-450, 2020.
[39]. L. Bourtoule et al., "Machine unlearning," in Proc. 2021 IEEE Symposium on Security and Privacy, pp. 141-159, 2021.
[40]. A. P. Felt, R. W. Reeder, E. Ha, and N. Malkin, "Improving SSL warnings: Comprehension and adherence," in Proc. 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015), pp. 2893-2902, 2015.
[41]. F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," Behavior Research Methods, vol. 39, no. 2, pp. 175-191, 2007.
[42]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Conference on Neural Information Processing Systems (NeurIPS 2017), pp. 4765-4774, 2017.
[43]. L. F. Cranor, "A framework for reasoning about the human in the loop," in Proc. 1st Conference on Usability, Psychology, and Security (UPSEC 2008), USENIX, 2008.