The Role of Human Behavior in Phishing Attacks: A Behavioral Cybersecurity Approach

Ankur Sharma
Bourns, Inc. / IT & Security
Western Governors University, Millcreek, USA

ABSTRACT

Even with powerful tools to detect and prevent phishing attacks, the threat remains one of the most prevalent cybersecurity issues. Although cybersecurity tools increasingly integrate artificial intelligence (AI) and machine learning (ML), one area remains persistently exploited by cybercriminals: human behavior, targeted through social engineering. This study analyzes the role of human behavior in phishing attacks from a behavioral cybersecurity perspective, combining a conceptual, literature-based synthesis with an empirical evaluation of five machine learning classifiers on two public phishing datasets comprising 69,700 labeled instances in total. The key behavioral dimensions analyzed are trust, perceived urgency, fear, curiosity, digital literacy, cybersecurity awareness, and decision-making biases. In the empirical evaluation, the Random Forest classifier achieved 97.11% and 95.58% accuracy on the two datasets, confirming the maturity of technical detection. A complementary scenario analysis restricted to deception cues that are, in principle, visible to end users retained approximately 91% accuracy on both datasets, indicating that the information required to recognize most phishing attacks is present in what users can observe, and that victimization arises primarily from cognitive and emotional manipulation rather than from an absence of information. The results indicate that technological solutions alone are insufficient in the absence of human-centered cybersecurity strategies and regular behavioral interventions. By integrating recent findings with quantitative evidence, the study advances the field of behavioral cybersecurity and presents a case for adaptive behavioral approaches to minimize phishing risks.

General Terms

Security, Human Factors, Experimentation, Measurement.

Keywords

Behavioral cybersecurity; phishing attacks; human factors; social engineering; cybersecurity awareness; phishing susceptibility; machine learning; human-centered security strategy.

1. INTRODUCTION

Digital communication technologies and online services have had a profound impact on how organizations and individuals communicate today. This technological development has, however, increased exposure to cyber threats, especially phishing attacks, one of the most prevalent and effective types of cybercrime. Phishing is a manipulative communication technique designed to trick people into sharing personal information, downloading malicious software, or granting unnecessary access to systems. Even though there have been significant advancements in AI and ML for detecting phishing attacks and securing cyber systems, ongoing human-factors challenges keep phishing attacks highly successful [1], [2].

In recent years, researchers have recognized that technical security infrastructure alone is not enough to defeat new phishing tactics. Attackers manipulate emotions of fear, urgency, trust, authority, and curiosity to override rational thinking and elicit impulsive user actions [3]. This shift has driven the emergence of behavioral cybersecurity as an interdisciplinary field that combines cybersecurity, psychology, behavioral science, and human-computer interaction (HCI) to understand cyber risk behaviors [4], [5].

Human behavior has thus emerged as a key factor in managing cybersecurity risk. Existing studies indicate that phishing susceptibility is significantly influenced by cybersecurity awareness, digital literacy, stress, cognitive overload, and individual perceptions of phishing risk [6], [7]. In addition, phishing tactics have become more sophisticated, particularly spear phishing, cryptocurrency phishing, and other techniques that target user vulnerabilities with the aid of artificial intelligence [8]. While many ML- and AI-based models have proven successful in detecting phishing emails and fraudulent websites, attacks constantly evolve to exploit behavioral weaknesses that technical systems may not catch [9], [10].

However, much remains to be done in research on the behavioral aspects of phishing. Previous cybersecurity research has focused mainly on technical detection approaches, including machine learning classifiers, URL filtering systems, and automated threat identification, with comparatively little attention to the psychological and behavioral factors that shape phishing susceptibility [11], [12]. Existing cybersecurity models also tend to underweight the impact of cognitive biases, emotional manipulation, trust perceptions, and decision-making on phishing victimization, and incorporating behavioral science into phishing prevention remains a relatively new research direction [13], [4]. The need for human-centric cybersecurity solutions that complement technical defense mechanisms through adaptive behavioral interventions and awareness strategies is therefore increasing.

Recent developments have also highlighted the value of behavioral interventions and cybersecurity training frameworks in reducing phishing susceptibility. Awareness programs can shift digital practices and improve resistance to social engineering, but they are mainly geared toward raising knowledge and leave the deeper cognitive and emotional mechanisms of victimization unaddressed [14], [13]. This imbalance underscores the need for a combined behavioral and technological strategy to mitigate phishing attacks.

The purpose of this study is to explore phishing attacks from a behavioral cybersecurity perspective and to examine the role human behavior plays in them. The study makes four contributions:

- It synthesizes the psychological, cognitive, and behavioral factors associated with phishing victimization and the social engineering techniques

that target human cognitive processes.

- It provides an empirical, multi-dataset evaluation of five machine learning classifiers on two public phishing corpora (69,700 instances in total), grounding the conceptual analysis in reproducible quantitative evidence.
- It introduces a scenario analysis that separates user-observable deception cues from technical lookup-based signals, quantifying how much phishing evidence is, in principle, available to end users at the moment of decision.
- It proposes a human-centered behavioral cybersecurity framework that integrates technical detection with behavioral and psychological defense layers.

The rest of this paper is organized as follows. Section 2 reviews the literature on phishing attacks, behavioral cybersecurity, and human susceptibility factors. Section 3 presents the methodological approach, including the design of the experimental evaluation. Section 4 reports the thematic and experimental results. Section 5 discusses the implications of the findings, and Section 6 concludes the study with recommendations for future behavioral cybersecurity research and practice.

2. LITERATURE REVIEW

2.1 Concept and Evolution of Phishing Attacks

With the evolution of the internet and digitalization, phishing has become one of the most pervasive and sophisticated cyber threats. Phishing was once limited to malicious emails prompting users to enter sensitive credentials such as passwords, financial details, or personal information. The rapid development of communication technologies, online financial systems, and cloud-based platforms has since transformed phishing into a multidimensional threat involving technical trickery, psychological manipulation, and social engineering [3].

Modern phishing extends beyond traditional email to spear phishing, clone phishing, smishing, vishing, social media phishing, and cryptocurrency phishing. Spear phishing is particularly effective because it is personalized and tailored to individual targets or contexts [8]. Attackers use contextual familiarity, institutional impersonation, and emotional persuasion to make messages appear legitimate and to prompt compliance.

Advanced technologies, including AI and automated content creation, are making phishing more adaptive. AI-enhanced attacks can imitate authentic communication patterns, personalize phishing content, and complicate detection [12]. Phishing has consequently evolved from a technical cyber threat into a socio-technical challenge that exploits both system vulnerabilities and human behavior.

2.2 Behavioral Cybersecurity Perspective

Behavioral cybersecurity is an interdisciplinary field concerned with the connection between human behavior and cybersecurity risk management. Distinct from traditional, technology-oriented strategies that focus on technical infrastructure, it emphasizes that human cognitive processes, decision-making, emotions, and patterns of interaction with digital environments shape cybersecurity outcomes [4].

Recent literature identifies human beings as one of the most vulnerable components of any cybersecurity system. Siponen et al. [5] showed that psychological rationalization processes and behavioral discounting mechanisms significantly affect cybersecurity hygiene and secure practices. Mersinas et al. [13] argued that ethical and behavioral factors must be considered to shape positive attitudes toward cybersecurity without provoking resistance or behavioral fatigue. Complementing these user-level perspectives, Jungebloud et al. [15] demonstrated that behavioral considerations can be embedded directly into system-level security engineering through model-based structural and behavioral risk assessment, while Joshi et al. [16] used prospect theory to show that security decision-makers systematically deviate from expected-utility rationality when allocating resources to controls, confirming that bounded and biased decision-making operates at organizational as well as individual levels.

Recognizing behavioral cybersecurity represents an important paradigm shift. Success in phishing is related less to technical sophistication than to users' responses to deceptive communication. A deeper understanding of behavioral susceptibility is therefore crucial for enhancing cyber and cyber-physical resilience in the current digital landscape.

2.3 Human Factors and Phishing Susceptibility

Numerous human behavioral factors have been identified as key contributors to phishing vulnerability. The literature documents how attackers exploit human weaknesses, emotions, and routine online activities to influence user behavior and undermine security best practices.

Cybersecurity awareness and digital literacy remain the most widely accepted predictors of phishing resistance. Khan et al. [6] found that structured awareness training helps users identify phishing and other suspicious online activity more readily, although the efficacy of training is uneven and depends on behavioral engagement, environmental factors, and retention.

Several studies also show that emotional and cognitive states significantly affect susceptibility. Stress, fatigue, time pressure, multitasking, and information overload reduce users' capacity for critical evaluation and increase impulsive decision-making. Gahletia [7] found that social media environments amplify vulnerability through rapid information uptake, reduced verification practices, and increased exposure to deceptive communication patterns.

Trust perception is another key factor in victimization. Attackers frequently impersonate trusted institutions such as employers, financial organizations, or government agencies to create an illusion of legitimacy and suppress suspicion. Carroll et al. [3] highlighted that phishing communications have become markedly more realistic and increasingly difficult for end users to distinguish from genuine messages.

Susceptibility also varies with demographics and behavior: professional role, cybersecurity experience, digital dependence, and habitual online behaviors all contribute. Together, these findings indicate that phishing susceptibility is not primarily a technological problem but a complex behavioral phenomenon shaped by cognitive, emotional, and social factors.

2.4 Social Engineering and Psychological Manipulation

Social engineering is one of the basic methods of phishing. Rather than exploiting vulnerabilities in a system, social

engineering exploits human psychology to induce actions that compromise cybersecurity defenses. The reviewed literature shows that phishing campaigns rely heavily on psychological persuasion techniques that evoke emotional and cognitive reactions.

Authority manipulation is among the most frequently used techniques: posing as a trusted entity or authority figure makes recipients more likely to follow instructions. Fear-based messaging operates similarly, as seen in campaigns built on account-suspension alerts, financial threats, or security-compromise notifications.

Urgency and scarcity are commonly used to dissuade users from rational verification. When users feel compelled to act immediately, they are more likely to click fraudulent links or accept fraudulent requests before screening for deception. Curiosity-driven manipulation, using provocative headlines or intriguing content, further contributes to impulsive behavior in digital environments.

The literature also indicates that social engineering is becoming more sophisticated through behavior-based profiling and personalization. Personalized spear-phishing attacks combine message content, organizational structure, and familiar behavioral patterns to increase psychological impact and attack success. Contemporary phishing attacks are therefore better characterized as psychologically adaptive manipulation systems than as traditional technical attacks. Figure 1 illustrates the behavioral phishing manipulation cycle identified in the reviewed literature.

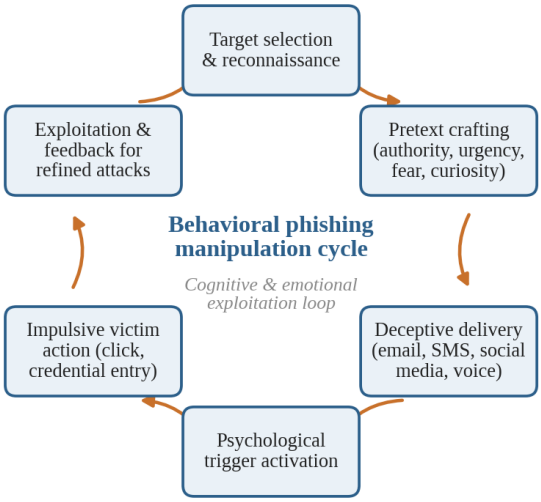

Fig 1: Behavioral cycle of phishing attack manipulation

2.5 Artificial Intelligence and Machine Learning in Phishing Detection

Significant advances have been made in AI and ML for phishing detection. Many intelligent frameworks have been proposed to detect suspicious communication patterns, deceptive websites, abnormal digital behavior, and malicious URLs.

Asiri et al. [1] reported that phishing identification systems are becoming increasingly effective through URL analysis, HTML inspection, and machine learning classification. Kapan and Sora Gunal [2] showed that detection performance improves when classifiers are combined with optimized feature extraction. Tin et al. [17] further demonstrated that predictive models trained on behavioral data can anticipate cybersecurity threats, connecting the detection literature directly to behavioral signals. Transformer-based models and clustering techniques have also demonstrated strong potential [12], [18], and complementary prevention strategies have emerged, including HTTP header analysis, CAPTCHA keystroke dynamics, and ensemble learning methods [10], [19], [11].

The literature is nevertheless consistent in stating that such technologies will not fully solve phishing. As techniques evolve, they are designed not only to evade automated systems but to exploit human weaknesses that lie outside the reach of technical controls. This constraint underscores the need for strategies that combine intelligent technical defenses with behavioral cybersecurity solutions.

2.6 Cybersecurity Awareness and Behavioral Interventions

Educating users is a common approach to countering phishing and encouraging safe digital practices. Prior research indicates that awareness training positively affects phishing detection and promotes more cautious online behavior [14]. At the level of workforce development, Almeida [20] showed that structured cybersecurity skills frameworks are increasingly used to define and cultivate the human competencies that underpin organizational security capability.

Recent literature, however, casts doubt on the long-term effectiveness of traditional awareness strategies. Informational training yields short-term improvements in phishing recognition, but relaxed or cognitively demanding situations prompt users to revert to less secure behavioral patterns. Knowledge acquisition alone is insufficient to sustain cybersecurity compliance.

Behavioral intervention frameworks have therefore become an increasingly important part of cybersecurity risk management. Mersinas et al. [13] emphasized ethical behavioral interventions aimed at influencing secure behaviors, behavioral consistency, and decision-making processes. Behavioral reinforcement measures, including simulated phishing exercises, adaptive awareness systems, and repeated engagement mechanisms, show potential for enhancing long-term resilience.

Current research supports a shift toward human-centered cybersecurity cultures that favor iterative, behavior-change-oriented training over compliance-driven instruction, acknowledging that organizational culture, emotional states, cognitive load, and digital-environment interactions all shape cybersecurity behavior. Behavioral interventions may therefore prove more effective at preventing phishing than strictly awareness-based approaches.

2.7 Research Gap

Despite significant strides in phishing detection and awareness, key research gaps remain in the behavioral aspects of phishing prevention. The literature offers abundant technical solutions, including machine learning classifiers, URL filtering mechanisms, and automated detection architectures, but comparatively little research addresses the intricate psychological and behavioral factors underlying susceptibility.

Most existing frameworks retain a technology-centric view and fail to incorporate behavioral science, cognitive psychology, and social engineering analysis into phishing defenses. Limited agreement also exists on the long-term success of behavioral interventions and adaptive awareness models. Moreover, few studies attempt to connect the two strands quantitatively, for example by asking how much of the evidence exploited by machine learning detectors is, in principle, available to human

users at the moment of decision.

Concurrently, AI-powered phishing techniques make a comprehensive, human-centric approach more important than ever, integrating intelligent technical defenses with behavioral analysis and psychologically informed responses. Table 1 summarizes selected studies and their major contributions to behavioral cybersecurity research.

Table 1. Summary of existing studies on behavioral cybersecurity and phishing attacks

Study	Research focus	Method / approach	Key findings
Prümmer et al. [14]	Cybersecurity awareness training	Systematic review	Training improves phishing awareness, but long-term behavioral adaptation remains limited
Baltuttis et al. [4]	Cybersecurity behavior typology	Behavioral classification	Human behavior significantly influences cybersecurity compliance
Carroll et al. [3]	Phishing email deception	Analytical investigation	Social engineering and emotional manipulation increase phishing success
Asiri et al. [1]	AI phishing detection	Machine learning survey	Intelligent systems improve phishing detection accuracy
Mersinas et al. [13]	Behavioral cybersecurity interventions	Conceptual framework	Ethical behavioral interventions enhance cybersecurity resilience
Khan et al. [6]	Awareness training effectiveness	Kirkpatrick evaluation model	Structured awareness programs improve phishing recognition

3. METHODOLOGY

3.1 Research Design

This study employed a two-component design. The first component is a conceptual, literature-based analysis of the role of human behavior in phishing attacks within behavioral cybersecurity, using narrative review and thematic synthesis to assess scholarly work on phishing susceptibility, social engineering techniques, cybersecurity awareness, and behavioral interventions. The second component is an empirical evaluation of machine learning phishing detection on two public benchmark datasets, designed to triangulate the conceptual synthesis with reproducible quantitative evidence and to examine the relationship between machine-detectable and human-perceptible deception signals.

The conceptual component incorporates perspectives from cybersecurity, behavioral science, psychology, and human-computer interaction to provide a holistic view of how behavioral factors affect phishing success. The empirical component differs from purely technical detection studies by explicitly interpreting classifier behavior through a behavioral lens, asking not only how accurately phishing can be detected but where the discriminative evidence resides relative to what users can observe.

3.2 Data Sources and Literature Selection

The literature component draws on secondary data from peer-reviewed journal articles, conference papers, and scholarly cybersecurity publications, with an emphasis on recent work from 2022 to 2025 to reflect current threats and behavioral developments. Sources were selected from respected academic databases and journals in cybersecurity, information systems, artificial intelligence, and the behavioral sciences. Studies were included if they addressed at least one of the following domains:

- Human behavior and cybersecurity
- Phishing attack mechanisms
- Social engineering techniques
- Cybersecurity awareness and training
- Behavioral intervention strategies
- AI-based phishing prevention systems

Studies not directly relevant to phishing behavior or cybersecurity behavior were omitted to ensure consistency of analysis and thematization.

3.3 Analytical Framework

The selected literature was analyzed using a thematic analytical framework, iteratively examining common themes, patterns, and concepts across studies. Thematic categorization organized the findings into the main behavioral dimensions of cybersecurity:

- Psychological manipulation techniques
- Emotional and cognitive susceptibility factors
- Cybersecurity awareness and digital literacy
- Behavioral intervention strategies
- Human-centered cybersecurity frameworks
- Technical versus behavioral cybersecurity approaches

This structure supported the identification of relationships between successful phishing attacks and behavioral tendencies, and enabled comparison of technological and behavioral countermeasures.

3.4 Experimental Evaluation Design

To address the need for evaluation across multiple datasets and scenarios, the empirical component assesses five widely used classifiers on two public phishing datasets that differ in era, scale, and feature granularity. Table 2 in Section 4.4 compares defensive paradigms conceptually; Table 3 summarizes the datasets used here.

Table 3. Characteristics of the two evaluation datasets

Property	Dataset A	Dataset B
Source	UCI Phishing	Vrbančič et al.,

	Websites, Mohammad et al. [21]	Data in Brief [22]
Instances	11,055	58,645
Phishing / legitimate	6,157 / 4,898	30,647 / 27,998
Features	30 categorical indicators (−1, 0, 1)	111 numeric attributes
Feature families	Address-bar, abnormality, HTML/JavaScript, and domain-based indicators	URL-lexical attributes per URL component, plus domain and external lookup signals

Dataset A is the UCI Phishing Websites dataset of Mohammad et al. [21], comprising 11,055 instances described by 30 categorical indicators spanning address-bar characteristics (e.g., IP address in URL, URL length, shortening services, '@' symbols, prefix-suffix hyphenation, subdomain depth), abnormality signals, HTML and JavaScript behaviors (e.g., status-bar tampering, right-click blocking, pop-up credential prompts), and domain-based signals (e.g., domain age, DNS records, web traffic rank). Dataset B is the corpus of Vrbančič et al. [22], comprising 58,645 instances described by 111 numeric attributes, including fine-grained lexical counts computed separately for each URL component and a set of external signals obtained through network lookups (e.g., WHOIS activation and expiration times, resolved IP counts, nameserver and MX counts, TTL, ASN, SPF, TLS certificate validity, redirect counts, and search-engine indexing). Using two independently collected datasets with different feature philosophies supports the generalizability of conclusions beyond a single corpus.

Five classifiers were evaluated: Logistic Regression (a linear baseline with feature standardization), Gaussian Naive Bayes (a probabilistic baseline assuming conditional feature independence), Decision Tree (an interpretable non-linear model), Random Forest (a bagged ensemble of 200 trees), and histogram-based Gradient Boosting (a boosted ensemble). All models were implemented in scikit-learn [23]. Each dataset was partitioned with a stratified 70/30 train-test split under a fixed random seed (42), and five-fold cross-validation on the training partition was used to verify stability. Performance is reported as accuracy, precision, recall, F1-score, and area under the ROC curve (ROC-AUC), with phishing as the positive class.

In addition to the standard evaluation, a scenario analysis was designed to operationalize the study's central behavioral question: is phishing victimization caused by an absence of information, or by a failure to attend to information that is present? For each dataset, features were partitioned into (a) user-observable deception cues, defined as signals that are, in principle, perceptible to an end user in the browser interface, comprising address-bar composition, the security indicator, and page interaction behaviors, and (b) technical signals requiring lookups a user cannot perform at a glance, such as WHOIS registration data, DNS records, traffic statistics, and search-engine indexing. For Dataset A this yields 12 observable cues out of 30 features; for Dataset B, 98 URL-lexical attributes out of 111, with 13 external lookup signals excluded. The Random Forest classifier was then retrained under each scenario, and the performance difference quantifies the discriminative value of evidence visible to users versus evidence available only to machines.

3.5 Conceptual Behavioral Cybersecurity Model

The study additionally employs a conceptual, behavior-based view of cybersecurity to explain how phishing attacks use psychological and cognitive processes to circumvent traditional security measures. The model assumes that behavior, emotions, cognition, and environment interact multiplicatively in shaping susceptibility, including trust levels, urgency responses, fear-based decision-making, information overload, and cybersecurity awareness. It further acknowledges that technological infrastructure alone may be insufficient without complementary adaptive behavioral interventions, positioning behavioral cybersecurity as a supportive component of technical countermeasures.

3.6 Reliability and Validity Considerations

To increase validity and reliability, the synthesis draws on recent scholarly resources from reputable cybersecurity and information systems journals, and uses cross-study comparison and thematic consistency to limit interpretational bias. For the empirical component, reliability is supported by the use of public benchmark datasets, fixed random seeds, stratified splitting, and cross-validated stability checks, allowing the full experimental pipeline to be reproduced.

3.7 Ethical Considerations

The research involved no human participants, no collection of personal data, and no live phishing activity. The empirical evaluation used publicly available, de-identified benchmark datasets released for research purposes. Ethical approval was therefore not required. The study nevertheless adhered to principles of academic integrity by citing sources correctly, interpreting prior literature responsibly, and avoiding overgeneralization of findings.

4. RESULTS

4.1 Overview of the Reviewed Literature

The literature review indicates that phishing attacks remain highly effective despite extensive detection efforts and AI-powered security infrastructure. Across the analyzed studies, susceptibility was consistently linked to behavioral, psychological, and cognitive weaknesses rather than purely technical flaws: attackers increasingly leverage human-centric factors such as trust, perceived urgency, emotional response, curiosity, and cognitive overload to induce insecure decisions.

The reviewed studies also show that phishing has expanded from mass email campaigns to highly targeted, psychologically adaptive social engineering operations. Modern attacks combine personalized communication styles, customized interfaces, urgency, and authority spoofing to compel compliance while evading conventional awareness measures.

4.2 Behavioral Factors Influencing Phishing Susceptibility

A common theme across the research is that phishing vulnerability depends heavily on human behavior. Awareness and digital literacy were repeatedly found to be critical determinants of resistance: phishing lures remain more effective against individuals with limited knowledge of phishing indicators, suspicious URLs, or deceptive communication patterns.

Cognitive and emotional states also significantly affect phishing decision-making. Stress, workload pressure, fatigue, fear, and perceived urgency were frequently associated with behavioral impulsiveness that interferes with critical assessment of suspicious communications; people were often prompted by emotion to act before verifying authenticity.

Furthermore, over-trust in institutional communications, overconfidence in one's own cybersecurity ability, and habitual digital behavior patterns increased susceptibility. These results confirm that phishing susceptibility is shaped by general psychological and behavioral factors, not technology alone.

4.3 Social Engineering and Psychological Manipulation Patterns

All reviewed studies identified social engineering as key to phishing success. Attackers commonly combine psychological tactics of authority, fear, scarcity, curiosity, and urgency, and impersonate trusted entities such as employers, financial institutions, or government agencies to create legitimacy and induce compliance.

Messages conveying urgency proved especially effective at suppressing rational decision-making: threats of account suspension, financial loss, or security compromise often prompted immediate responses before any verification. Fear-laden and emotionally charged language likewise increased engagement with malicious links and attachments at the point of interaction.

The analysis also suggests that attackers tune communication tactics and delivery environments to specific user groups. Spear-phishing campaigns targeting professionals, organizational staff, and cryptocurrency investors achieved higher success rates through personalization and contextual familiarity.

4.4 Evaluation of Technical Detection Approaches

The reviewed literature documents significant progress in machine learning, clustering algorithms, transformer architectures, and URL-based detection, with intelligent classification of malicious URLs, suspicious HTTP headers, and deceptive web structures showing strong capability. Despite these advances, the literature is unanimous that technical detection alone is insufficient: phishing constantly evolves to outmaneuver automated defenses using human tactics rather than technical ones. Table 2 contrasts the two defensive paradigms identified in the reviewed literature.

Table 2. Comparative analysis of technical and behavioral cybersecurity approaches

Dimension	Technical cybersecurity approach	Behavioral cybersecurity approach
Primary focus	System protection and threat detection	Human behavior and decision-making
Key techniques	AI models, URL filtering, classifiers	Awareness training, behavioral interventions
Target vulnerability	Technical system weaknesses	Cognitive and emotional vulnerabilities
Main strength	High-speed automated detection	Improved human cybersecurity resilience
Main limitation	Limited understanding of user behavior	Difficult behavioral standardization
Adaptability	Algorithm-driven updates	Human-centered adaptive learning
Long-term effectiveness	Strong technical defense	Sustainable behavioral compliance
Role in phishing prevention	Detects malicious activities	Reduces phishing susceptibility

The evidence therefore indicates that detection systems can be algorithmically strong yet unable to address the behavioral factors that lead to victimization, reinforcing the need to integrate behavioral defenses with technical measures. The following two subsections test and quantify this claim empirically.

4.5 Experimental Results: Classifier Performance

Table 4 reports the test-set performance of the five classifiers on both datasets, and Figure 3 visualizes the F1-scores. Random Forest performed best on both corpora, achieving 97.11% accuracy (F1 = 0.9742, ROC-AUC = 0.997) on Dataset A and 95.58% accuracy (F1 = 0.9580, ROC-AUC = 0.991) on Dataset B, with Gradient Boosting a close second. Five-fold cross-validation on the training partitions confirmed stability, with Random Forest F1 of 0.9724 ± 0.0021 on Dataset A and 0.9568 ± 0.0023 on Dataset B.

Table 4. Test-set performance of the five classifiers on both datasets

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Dataset A — UCI Phishing Websites (11,055 instances)					
Logistic Regression	0.9276	0.9272	0.9442	0.9356	0.9790
Naive Bayes	0.6063	0.9945	0.2945	0.4545	0.9678
Decision Tree	0.9611	0.9623	0.9681	0.9652	0.9737
Random Forest	0.9711	0.9669	0.9816	0.9742	0.9972
Gradient Boosting	0.9692	0.9683	0.9767	0.9725	0.9965
Dataset B — Vrbančič et al. (58,645 instances)					
Logistic Regression	0.8987	0.8912	0.9183	0.9045	0.9612
Naive Bayes	0.7339	0.9193	0.5380	0.6787	0.9193
Decision Tree	0.9307	0.9341	0.9331	0.9336	0.9305
Random Forest	0.9558	0.9518	0.9642	0.9580	0.9913
Gradient Boosting	0.9490	0.9488	0.9539	0.9513	0.9891

Three aspects of these results merit detailed interpretation. First, the ranking of classifiers is identical across the two datasets despite their different eras, scales, and feature philosophies, indicating that the findings are properties of the phishing detection problem rather than artifacts of a particular

corpus. Second, the linear baseline remains competitive (92.76% and 89.87% accuracy), showing that much phishing evidence is linearly separable; ensemble methods add most of their value by capturing interactions between deception cues. Third, Gaussian Naive Bayes collapses in recall (29.5% and 53.8% of phishing instances detected) while retaining high precision. This failure is behaviorally informative: the independence assumption is violated precisely because deception cues co-occur by design. Attackers compose manipulation strategies, bundling, for example, a spoofed subdomain with an urgency-laden pretext and a deceptive anchor structure, so the statistical dependence among features is itself a signature of deliberate psychological staging rather than random noise.

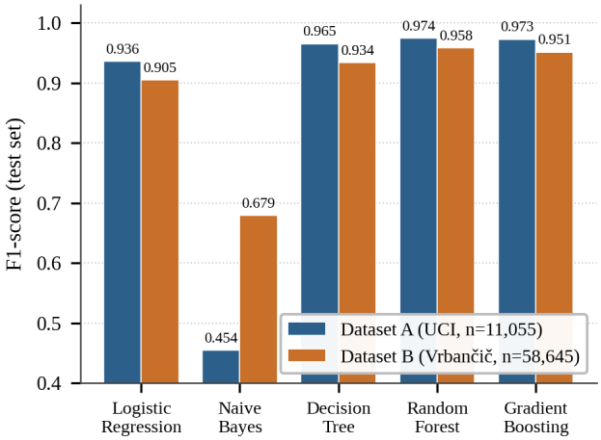

Fig 3: Test-set F1-scores of the five classifiers on both datasets

Figure 4 presents the ROC curves. The ensemble methods dominate across the entire operating range on both datasets, which matters operationally because email-security gateways must run at very low false-positive rates to remain usable; the ensembles retain high true-positive rates even in that regime.

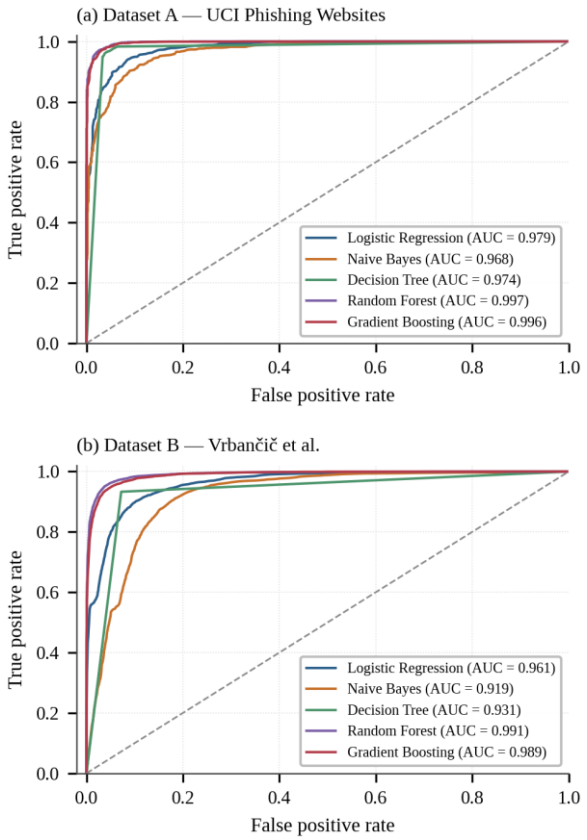

Fig 4: ROC curves for all classifiers on (a) Dataset A and (b) Dataset B

4.6 Experimental Results: Behavioral Interpretation

Figure 5 shows the fifteen most important features learned by the Random Forest on Dataset A, where feature semantics permit direct behavioral reading. Importance is heavily concentrated: the top two features, the SSL certificate state and the proportion of deceptive anchor URLs, jointly account for 57.0% of total importance. The single most decisive signal, SSL final state (importance 0.312), is precisely the padlock heuristic that security awareness programs have taught users for two decades. Its dominance is double-edged: the cue is user-observable in principle, yet modern attackers increasingly obtain valid certificates for phishing domains, degrading the padlock's diagnostic value for humans while machines can still exploit certificate issuer and age nuances. Three of the top ten features are user-observable cues (SSL state, subdomain depth, and prefix-suffix hyphenation), and these are exactly the address-bar spoofing patterns, such as legitimate-brand-name.malicious-domain.com constructions, that social engineering must render inconspicuous for an attack to succeed.

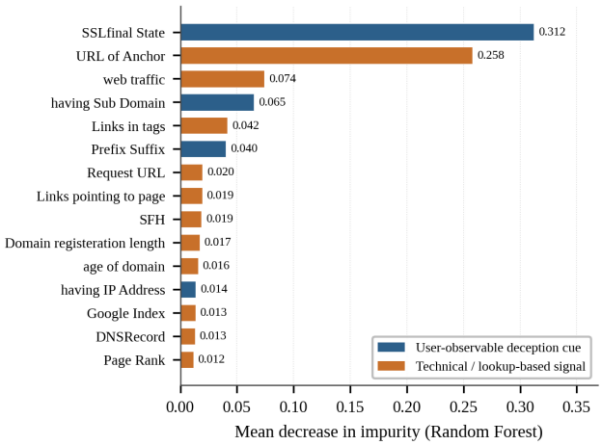

Fig 5: Top-15 Random Forest feature importances on Dataset A, colored by user observability

Table 5 and Figure 6 report the scenario analysis. Restricted to user-observable cues alone, the Random Forest retains 90.77% accuracy on Dataset A (12 of 30 features) and 90.57% on Dataset B (98 of 111 features), reductions of 6.3 and 5.0 percentage points respectively.

Table 5. Scenario analysis: all features versus user-observable cues (Random Forest)

Dataset	Feature scenario	k	Accuracy	F1-score	ROC-AUC
A (UCI)	All features	30	0.9711	0.9742	0.9972
A (UCI)	User-observable cues only	12	0.9077	0.9192	0.9609
B (Vrbančič)	All features	111	0.9558	0.9580	0.9913
B (Vrbančič)	User-observable cues only	98	0.9057	0.9087	0.9646

This result carries a two-sided behavioral interpretation. On one side, roughly 91% of phishing instances are machine-separable from evidence that is present in what a user can see at the moment of decision. The information deficit explanation of phishing victimization is therefore untenable as a primary account: in most cases the deception is perceptible, but emotional manipulation, urgency, and cognitive load prevent the evidence from being attended to and integrated, which is consistent with the susceptibility mechanisms identified in the thematic synthesis. On the other side, the residual gap of five to six percentage points represents discriminative evidence that no amount of user vigilance can recover, because it resides in WHOIS registration ages, DNS characteristics, traffic statistics, and indexing signals invisible in the browser. Neither layer is sufficient alone, which is the quantitative core of the socio-technical argument developed in Section 5.

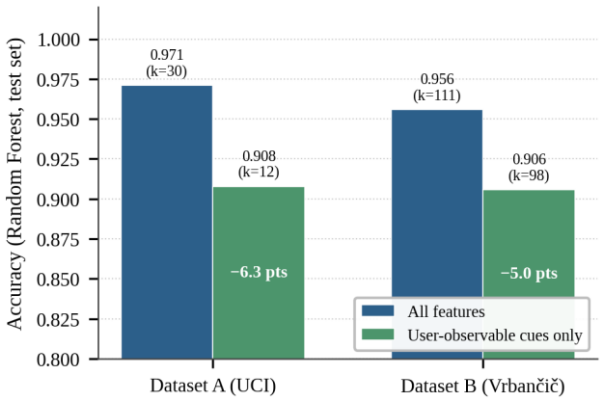

Fig 6: Accuracy under the full-feature and user-observable-cue scenarios

A further practical observation concerns precision. Under the observable-cue scenario on Dataset A, precision falls from 0.9669 to 0.8970, meaning that judgments based solely on visible cues generate materially more false alarms. Translated to human terms, even a perfectly attentive user relying on individual visible indicators will misjudge legitimate messages more often than an enriched technical system; awareness training should therefore teach probabilistic combinations of cues and verification routines rather than single-indicator rules.

4.7 Cybersecurity Awareness and Behavioral Intervention Outcomes

The reviewed literature showed strong benefits of awareness training and behavioral interventions for secure digital behavior and phishing resistance: structured programs improved threat awareness, reinforced cautious behavior, and increased users' ability to detect suspicious communication patterns.

However, traditional awareness training exhibits variable behavioral effects over time. In multiple studies, individuals knowledgeable about cybersecurity principles continued to engage in risky behavior under stress, time constraints, or cognitive overload, confirming that knowledge acquisition alone does not secure behavioral practice. Interventions built on repeated reinforcement, adaptive learning, simulated phishing, and psychologically informed awareness were more effective in fostering sustainable secure behavior, and ethical behavioral interventions targeting ingrained habits and decision processes were identified as an emerging field of practice.

4.8 Emerging Behavioral Threat Trends

A key trend in the literature is the consolidation of behavioral cybersecurity as a research and operations field. Recent work emphasizes that cybersecurity is not solely a technological issue but a human-behavioral one requiring interdisciplinary solutions. The analysis indicates that human-centric cybersecurity is maturing, with growing integration of psychological and behavioral science, risk perception analysis, and adaptive intervention strategies into defense models. Future phishing defenses are likely to combine AI-based technical detection with behavioral analytics that identify risky user activity, cognitive vulnerability, and unusual interaction tendencies in real time.

4.9 Summary of Key Findings

Overall, the results show that phishing attacks succeed largely because of human behavioral weaknesses, not technological gaps alone. Psychological manipulation, cognitive bias

exploitation, emotional triggering, and social engineering remain principal avenues for evading technical controls. The experimental evaluation confirms that intelligent detection is mature, with ensemble methods exceeding 95% accuracy on two independent corpora, while the observable-cue analysis demonstrates that most deception evidence is visible to users yet fails to prevent victimization, and that a irreducible share of evidence is available only to technical systems. There is accordingly a clear case for human-centered solutions that integrate technological protection with adaptive behavioral interventions, awareness, and psychological risk mitigation.

5. DISCUSSION

The results confirm the growing trend of viewing phishing as both a technical threat and a behavioral-psychological phenomenon. Although newer security architectures embrace AI/ML-based detection, attackers continue to succeed by exploiting human cognitive and emotional gaps, in line with human-centric and behavioral cybersecurity research.

A central finding is the persistent imbalance between technological defense development and behavioral risk mitigation. Most existing ecosystems concentrate on technical solutions such as URL filtering, anomaly detection, clustering, and AI-based systems [1], [9]. While these technologies markedly improve detection accuracy, a result reproduced directly in Section 4.5, the analysis indicates that attackers increasingly exploit behavioral tendencies rather than system vulnerabilities, a pattern sometimes described as behavioral phishing. Phishing effectiveness is thus tied closely to human decision-making and social engineering manipulation.

The role of psychological exploitation deserves emphasis. As the reviewed studies revealed, phishing's central goal is to inflame emotion in order to disarm critical thinking and prompt hasty action, operating on principles of psychological persuasion and behavioral conditioning. The experimental results sharpen this point quantitatively: because approximately 91% of phishing instances are separable from user-visible evidence alone, the attacker's practical problem is not to hide the evidence, which is largely impossible, but to prevent the victim from processing it, which is exactly what urgency, fear, and authority framing accomplish. Behavioral science is therefore not an accessory to phishing defense but structurally necessary to it.

Another key takeaway concerns the relationship between awareness and compliance. Awareness training enhances phishing detection, yet risk awareness does not automatically translate into secure action: individuals with prior cybersecurity experience remained vulnerable under stress, time pressure, emotional manipulation, and cognitive overload. Vulnerability thus extends beyond informational deficiency to a complex interplay of cognition, emotion, situation, and automatic online habits. The precision degradation observed in the observable-cue scenario adds a complementary caution: even attentive users relying on single visible indicators will produce misjudgments, so training should build compound verification routines rather than one-rule heuristics.

These observations challenge assumptions underlying traditional awareness initiatives, which focus on knowledge transfer and presume protective outcomes. Human behavior in cybersecurity contexts is neither always rational nor entirely intentional; heuristic and emotional decision-making and routinized interaction patterns dominate precisely in the cognitively demanding, high-pressure situations that attackers engineer. Strategies relying solely on informational awareness will therefore have short-lived impact unless supported by iterative behavior change and ongoing engagement.

The findings also contribute to behavioral cybersecurity as an interdisciplinary field integrating cybersecurity with psychology, behavioral economics, and human-computer interaction. As attackers adopt increasingly sophisticated psychological approaches, frameworks must account for patterns of human vulnerability alongside technical vectors. Behavioral interventions offer substantial complementary value: ethical behavioral interventions, simulated phishing exercises, adaptive awareness systems, and behavior-driven education all show potential to strengthen resistance, with adaptive methods appearing more effective for long-term change than static programs [13].

The discussion also yields lessons for organizational cybersecurity governance. Many organizations still treat phishing primarily as a technical infrastructure problem addressed through mail filtering, authentication, and endpoint security. The findings suggest that fostering a secure behavioral culture may be equally important to resilience, favoring governance practices that emphasize behavioral responsibility, ongoing training, emotional recognition, and phishing-simulation exposure. The two-sided result of the scenario analysis provides the design principle: user-facing programs should exploit the roughly 91% of evidence users can in principle perceive, while technical systems must cover the residual evidence that users cannot access, such as registration age and DNS characteristics, so that each layer compensates for the other's structural blind spots.

The advent of AI-powered phishing further raises complexity. Generative technologies can personalize content at scale, increasing realism, relevance, and persuasion, and making malicious messages harder to distinguish from authentic ones. Next-generation defenses will likely need to integrate intelligent technical detection, real-time behavioral analytics, and adaptive user support. Figure 2 presents the human-centered behavioral cybersecurity framework proposed on the basis of these findings, integrating a technical defense layer, a behavioral intervention layer, and a psychological support layer through human-centered governance and adaptive user support.

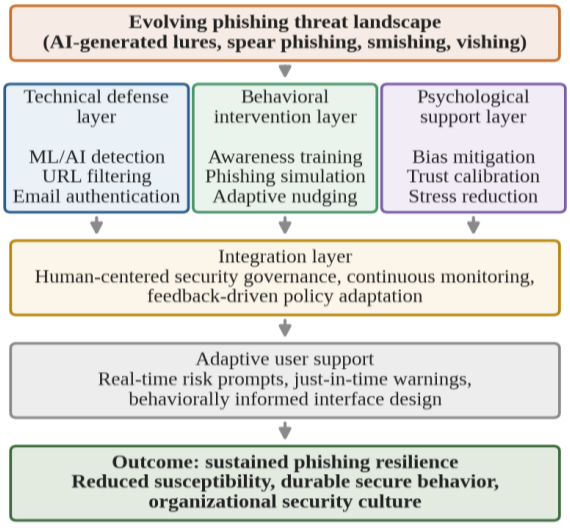

Fig 2: Proposed human-centered behavioral cybersecurity framework

Theoretically, this research supports the shift from technology-centric paradigms toward integrated socio-technical cybersecurity models, indicating that resilience must be

considered across technological, behavioral, cognitive, and organizational dimensions to reflect the real dynamics of phishing in today's digital environment.

The study has limitations. The behavioral component is conceptual and literature-based rather than experimental with human participants; no phishing simulation or behavioral observation was conducted, and differences in samples, methods, and settings across the reviewed studies may affect generalizability. The empirical component addresses the detectability of phishing artifacts, not user behavior directly, and the observability partition reflects in-principle perceptibility rather than measured perception; eye-tracking or simulation studies would be required to establish what users actually attend to. Dataset A also predates the most recent AI-generated phishing wave, although the consistency of results across the two corpora mitigates era effects. These limitations chart a concrete agenda for future behavioral cybersecurity research.

6. CONCLUSION

This study shows that phishing attacks continue to succeed largely because they exploit human vulnerabilities and psychological biases. Beyond technical factors, susceptibility is shaped by trust, perceived urgency, fear, cognitive bias, and limited cybersecurity awareness, and social engineering increasingly enables attackers to bypass or weaken traditional technical controls. The empirical evaluation across two public datasets confirms both the maturity of technical detection, with Random Forest reaching 97.11% and 95.58% accuracy, and the behavioral core of the problem: approximately 91% of phishing instances are distinguishable from cues visible to users, so victimization predominantly reflects manipulated attention and decision-making rather than missing information, while a residual share of evidence remains accessible only to technical systems.

Effective phishing prevention therefore requires an integrated approach combining technical cybersecurity measures with behavior-focused interventions. Cybersecurity awareness programs, adaptive behavioral training, and psychologically informed security frameworks can strengthen resilience across organizations and individual users. Integrating behavioral cybersecurity principles into broader security strategies, as captured in the proposed framework, is essential for building sustainable long-term defenses against phishing in an increasingly complex digital environment. Future work should extend the present analysis with human-subjects phishing simulations, perception measurement, and evaluation on post-2024 AI-generated phishing corpora.

7. REFERENCES

[1] Asiri, S., Xiao, Y., Alzahrani, S., Li, S., and Li, T. 2023. A survey of intelligent detection designs of HTML URL phishing attacks. IEEE Access 11, 6421–6443. DOI: https://doi.org/10.1109/ACCESS.2023.3237798

[2] Kapan, S. and Sora Gunal, E. 2023. Improved phishing attack detection with machine learning: a comprehensive evaluation of classifiers and features. Applied Sciences 13, 24, 13269. DOI: https://doi.org/10.3390/app132413269

[3] Carroll, F., Adejobi, J. A., and Montasari, R. 2022. How good are we at detecting a phishing attack? Investigating the evolving phishing attack email and why it continues to successfully deceive society. SN Computer Science 3, 2, 170. DOI: https://doi.org/10.1007/s42979-022-01069-1

[4] Baltuttis, D., Teubner, T., and Adam, M. T. P. 2024. A typology of cybersecurity behavior among knowledge workers. Computers and Security 140, 103741. DOI: https://doi.org/10.1016/j.cose.2024.103741

[5] Siponen, M., Topalli, V., Soliman, W., and Vestman, T. 2025. Reconsidering neutralization techniques in behavioral cybersecurity as a form of cybersecurity hygiene discounting. Computers and Security 150, 104306. DOI: https://doi.org/10.1016/j.cose.2024.104306

[6] Khan, N. F., Ikram, N., Murtaza, H., and Javed, M. 2023. Evaluating protection motivation based cybersecurity awareness training on Kirkpatrick's model. Computers and Security 125, 103049. DOI: https://doi.org/10.1016/j.cose.2022.103049

[7] Gahletia, K. 2025. Cybersecurity awareness in the age of social media: a behavioral study. International Journal for Research in Applied Science and Engineering Technology 13, 7, 46–49. DOI: https://doi.org/10.22214/ijraset.2025.72931

[8] Alyami, M., Alhotaylah, R., Alshehri, S., and Alghamdi, A. 2023. Phishing attacks on cryptocurrency investors in the Arab states of the Gulf. Journal of Risk and Financial Management 16, 5, 271. DOI: https://doi.org/10.3390/jrfm16050271

[9] Mosa, D. T., Shams, M. Y., Abohany, A. A., El-Kenawy, E. S. M., and Thabet, M. 2023. Machine learning techniques for detecting phishing URL attacks. Computers, Materials and Continua 75, 1, 1271–1290. DOI: https://doi.org/10.32604/cmc.2023.036422

[10] Shukla, S., Misra, M., and Varshney, G. 2024. HTTP header-based phishing attack detection using machine learning. Transactions on Emerging Telecommunications Technologies 35, 1, e4872. DOI: https://doi.org/10.1002/ett.4872

[11] Alsariera, Y. A., Alanazi, M. H., Said, Y., and Allan, F. 2024. An investigation of AI-based ensemble methods for the detection of phishing attacks. Engineering, Technology and Applied Science Research 14, 3, 14266–14274. DOI: https://doi.org/10.48084/etasr.7267

[12] Asiri, S., Xiao, Y., and Li, T. 2024. PhishTransformer: a novel approach to detect phishing attacks using URL collection and transformer. Electronics 13, 1, 30. DOI: https://doi.org/10.3390/electronics13010030

[13] Mersinas, K., Bada, M., and Furnell, S. 2025. Cybersecurity behavior change: a conceptualization of ethical principles for behavioral interventions. Computers and Security 148, 104025. DOI: https://doi.org/10.1016/j.cose.2024.104025

[14] Prümmer, J., van Steen, T., and van den Berg, B. 2024. A systematic review of current cybersecurity training methods. Computers and Security 136, 103585. DOI: https://doi.org/10.1016/j.cose.2023.103585

[15] Jungebloud, T., Nguyen, N. H., Kim, D. D., and Zimmermann, A. 2025. Model-based structural and behavioral cybersecurity risk assessment in system designs. Computers and Security 157, 104543. DOI: https://doi.org/10.1016/j.cose.2025.104543

[16] Joshi, C., Slapničar, S., Yang, J., and Ko, R. K. L. 2025. Contrasting the optimal resource allocation to cybersecurity controls and cyber insurance using prospect

theory versus expected utility theory. Computers and Security 154, 104450. DOI: https://doi.org/10.1016/j.cose.2025.104450

[17] Tin, T. T., Xin, K. J., Aitizaz, A., Tiung, L. K., Keat, T. C., and Sarwar, H. 2023. Machine learning-based predictive modeling of cybersecurity threats utilizing behavioral data. International Journal of Advanced Computer Science and Applications 14, 9, 832–840. DOI: https://doi.org/10.14569/IJACSA.2023.0140987

[18] Al-Sabbagh, A., Hamze, K., Khan, S., and Elkhodr, M. 2024. An enhanced K-means clustering algorithm for phishing attack detections. Electronics 13, 18, 3677. DOI: https://doi.org/10.3390/electronics13183677

[19] Alamri, E. K., Alnajim, A. M., and Alsuhibany, S. A. 2022. Investigation of using CAPTCHA keystroke dynamics to enhance the prevention of phishing attacks. Future Internet 14, 3, 82. DOI: https://doi.org/10.3390/fi14030082

[20] Almeida, F. 2025. Comparative analysis of EU-based cybersecurity skills frameworks. Computers and Security 151, 104329. DOI: https://doi.org/10.1016/j.cose.2025.104329

[21] Mohammad, R. M., Thabtah, F., and McCluskey, L. 2012. An assessment of features related to phishing websites using an automated technique. In Proceedings of the International Conference for Internet Technology and Secured Transactions (ICITST 2012). IEEE, 492–497.

[22] Vrbančič, G., Fister, I., Jr., and Podgorelec, V. 2020. Datasets for phishing websites detection. Data in Brief 33, 106438. DOI: https://doi.org/10.1016/j.dib.2020.106438.

[23] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.