

PHISHING EMAIL DETECTION USING A PAKISTAN SPECIFIC DATASET: REVISITING AN SVM BASELINE AND THE CASE FOR DATA CENTRIC AI SAFETY

Samra Riaz

Department of Computer Science, Sardar Bahadur Khan Women's University, Quetta, Pakistan

samrariaz0925@gmail.com

DOI: <https://doi.org/10.5281/zenodo.21531471>

Keywords

phishing detection; support vector machine; local context dataset; AI safety; email security; regional generalization

Article History

Received: 18 January 2026

Accepted: 03 March 2026

Published: 17 March 2026

Copyright @Author

Corresponding Author: *

Samra Riaz

Abstract

Phishing continues to be one of the most reported categories of cybercrime worldwide, with email remaining its dominant delivery channel. A structural but rarely examined feature of the phishing-detection literature is that nearly all widely used benchmark datasets, Enron, Nazario, Spam Assassin, and their derivatives are collected from North American and European sources. This creates an open and largely untested question: how representative are these corpora of phishing threats encountered in other regions, and what happens to detection systems built without regional data in view. This paper revisits an original, manually labeled phishing email corpus of 1,630 messages collected specifically from users in Pakistan over approximately eighteen months, and evaluates a linear kernel Support Vector Machine (SVM) with Bag-of-Words (BoW) feature extraction on this data, achieving 96% classification accuracy (macro-F1 = 0.95). We situate this result against the current generation of deep learning and transformer based phishing detectors, most of which report only marginal accuracy improvements over classical baselines while depending on substantially larger, Western sourced training corpora and far greater computational resources. Drawing on this comparison, we argue that dataset representativeness, not model architecture is a consequential and under addressed constraint on equitable, safe AI deployment in security relevant domains. We present the dataset's construction methodology, a full experimental evaluation, an honest accounting of the study's limitations, and a discussion of what data centric AI safety concretely requires in regional and resource constrained settings.

1 Introduction

Phishing is a social-engineering attack in which an adversary impersonates a trustworthy entity typically a bank, employer, or well-known service to induce a victim into disclosing credentials, financial details, or other sensitive information [1]. The term was first coined in 1996, and the technique has since scaled into one of the most persistent and costly categories of cybercrime. The FBI's Internet Crime Complaint Center (IC3) ranked phishing and spoofing as the single largest complaint category reported in the

United States in 2024, accounting for 193,407 individual complaints [2]. Losses attributed to cyber enabled fraud broadly reached an estimated \$16.6 billion in 2024 alone, a year-over-year increase of 33% [3].

The dominant technical countermeasure to phishing has been automated detection using machine learning, applied to email content, sender and header metadata, or embedded URLs. Classical approaches using Support Vector Machines (SVM), Naive Bayes, and decision trees achieved strong

reported results through the 2010s [4, 5, 6], and the field has since moved toward deep architectures, including convolutional and recurrent networks and fine-tuned transformer models such as BERT and RoBERTa [7, 8, 9].

A consistent, and rarely interrogated, feature of this literature, spanning both classical and deep learning approaches, is the provenance of the training data. The overwhelming majority of publicly available and widely cited phishing corpora, including Enron, Nazario, and Spam Assassin, are datasets collected from North American or European sources [11]. This concentration of data provenance raises a question that is more often assumed away than empirically tested: how well do phishing-detection models trained on these corpus generalize to deployment contexts that differ meaningfully in the specific institutions, cultural referents, and scam narratives being impersonated?

This paper does not attempt to resolve that generalization question through controlled cross corpus transfer experiments. Instead, it revisits a smaller, more foundational contribution: an original, manually labeled, Pakistan-specific phishing email dataset, built from scratch because no such corpus existed at the time of collection. We use this dataset to make the underlying data representativeness problem concrete, to report a reproducible classification baseline against it, and to argue with reference to the broader literature, that architectural sophistication in phishing detection has outpaced corpus diversity, and that this gap has direct implications for how AI safety should be understood and practiced in regionally underrepresented settings.

1.1 Motivation

The motivation for this work originated from a practical observation as existing phishing detection tools and published datasets, when informally tested against phishing emails actually received by users in Pakistan, performed inconsistently, and no dataset existed that reflected the specific institutions (local banks, telecom providers, government bodies) and scam narratives encountered locally. This is not a hypothetical concern. It reflects a broader and increasingly recognized pattern in AI safety research: models trained on data from one context can silently underperform, or fail in undetected ways, when

deployed in a different one [12, 13]. In security critical applications such as phishing detection, such failures are not merely an accuracy statistic they translate directly into financial loss and successful attacks against a population that generic global models were never built to protect.

1.2 Contributions

This paper makes the following contributions:

1. An original, community-sourced phishing email corpus of 1,630 manually labeled messages, reflecting institutions and scam narratives specific to Pakistan, collected over approximately eighteen months through structured community outreach.
2. A full, reproducible SVM + Bag-of-Words classification pipeline evaluated on this corpus, achieving 96% accuracy and a macro-averaged F1-score of 0.95, with a complete breakdown of precision, recall, and per class performance.
3. A comparative analysis situating this result against recent classical and deep-learning phishing detectors reported in the 2019-2025 literature, to characterize where architecture versus data tradeoffs is most consequential.
4. An explicit, self-critical limitations analysis covering dataset scale, class balance, temporal currency, label provenance, and scope, intended to inform future corpus building efforts in regionally underrepresented security domains.
5. A discussion connecting these findings to the broader practice of data centric AI safety, arguing that equitable and robust AI deployment in security relevant domains requires deliberate investment in regionally representative training data, not solely in model architecture.

1.3 Paper Organization

Section 2 reviews related work across classical machine learning, deep learning, and dataset bias. Section 3 describes the dataset and its construction. Section 4 details the methodology and experimental setup. Section 5 presents results and a comparative analysis. Section 6 discusses broader implications for AI safety in regionally underrepresented contexts. Section 7 presents limitations. Section 8 concludes and outlines future work.

2 Related Work

2.1 Classical Machine Learning Approaches to Phishing Detection

Early phishing-detection research relied on hand-engineered lexical, structural, and host-based URL features combined with classical statistical classifiers. Fette et al. [4] proposed PILFER, an SVM-based classifier using ten features derived from email structure and IP-based URL indicators, motivated by the observation that phishing emails share detectable statistical regularities with spam more broadly. Basnet et al. [5] compared multiple classifiers, including SVM and neural networks, on similar premises, finding SVM to be a consistently competitive baseline. Sahingoz et al. [6] extended feature based SVM classification specifically to URL only detection using NLP-based lexical features, reporting strong results without requiring access to email body content, a useful property when body content may be encrypted or otherwise inaccessible. Across this line of work, SVM emerges repeatedly as a strong, low compute, and comparatively data efficient baseline, which motivated its selection for the dataset examined in this paper.

2.2 Deep Learning and Transformer-Based Approaches

More recent work has moved toward learned representations rather than hand-engineered features. Altwaijry et al. [7] propose a one-dimensional convolutional neural network (1D-CNNPD) augmented with recurrent layers LSTM, Bi-LSTM, GRU, and Bi-GRU variants reporting that a bidirectional GRU augmentation yields the strongest results on the Phishing Corpus and Spam Assassin benchmark datasets. Songailaitė et al. [8] benchmark compact BERT variants, including DistilBERT and RoBERTa, for email phishing detection, reporting accuracy, precision, and recall figures all above 0.985 on an augmented test set of realistic phishing emails generated by paraphrasing genuine phishing content. Nguyen et al. [9] similarly apply BERT with transfer learning to phishing email classification, reporting comparably high performance. Thakur et al. [10] provide a systematic review of deep learning-based phishing email detection more broadly, and observe that reported accuracy gains over classical baselines are frequently modest, often

in the range of only one to three percentage points, once dataset and preprocessing conditions are held approximately constant. This observation is directly relevant to the present study's comparative analysis in Section 5.

2.3 Dataset Bias, Generalization, and the Reproducibility Problem

A smaller but growing body of work has begun to interrogate phishing detection datasets themselves, rather than the models trained on them. De Nardin et al. [11] caution that researchers must carefully manage sampling bias across the relatively small number of benchmark datasets that dominate the field, since repeated reuse of the same corpora risks overfitting to corpus-specific artifacts rather than learning generalizable phishing indicators. This concern is not merely theoretical: a related study on context-aware phishing detection reports that models can learn spurious, corpus specific tokens, for example, entity names unique to the Enron email corpus, such as employee or place names, as legitimacy indicators, which is a direct symptom of narrow and non-representative training data [12]. Kapoor and Narayanan's broader work on data leakage in machine learning based science similarly warns that reported performance figures can substantially overstate real world generalization when evaluation data is not meaningfully independent of training data in its distributional characteristics [14]. A separate line of work on low resource NLP settings more broadly has shown that systems trained predominantly on data from high resource contexts exhibit systematically different behavior when applied outside those contexts [13], a pattern consistent with the concerns motivating the present study.

2.4 Positioning of the Present Study

To our knowledge, no widely cited phishing email dataset in the published literature originates from natively collected Pakistani sources. The dataset examined in this paper was built specifically to begin addressing that gap, using natively collected, community sourced messages rather than adaptation of an existing corpus.

3 Dataset

3.1 Construction and Collection

Methodology

The dataset was constructed from scratch, motivated by the observation that no phishing email dataset reflecting the Pakistani context was available at the time of collection (2018–2019). Two complementary collection channels were used:

1. **Awareness-flyer distribution:** Flyers explaining what phishing emails are and how to recognize them were distributed at universities and public and government offices. Recipients were invited to forward suspected phishing emails they had personally received to a dedicated collection email address.

2. **Direct community collaboration:** Students, faculty members, and personal and professional networks across a range of institutional backgrounds contributed emails directly, often following brief in-person awareness sessions explaining the purpose of the collection.

Collection spanned approximately eighteen months in order to reach a sufficient volume of samples and to capture some diversity of sources, senders, and scam types, rather than relying on a single institutional source that might narrowly reflect one organization's threat profile.

3.2 Ethical Considerations

Participants voluntarily forwarded emails they had already received and identified (or suspected) as phishing attempts; no deceptive collection methods were used, and no new phishing attacks were generated or sent to participants as part of data collection. Sender addresses in the resulting dataset reflect the original phishing or legitimate senders as received, not the identities of the participants who forwarded them.

3.3 Composition

The final corpus comprises 1,630 labeled emails, structured around four attributes:

- **Sender:** the originating email address, often used informally by recipients to make an initial judgment of legitimacy.
- **Subject:** the email's subject line.
- **Content:** the full email body, including any embedded links, redirect targets, and persuasive or urgency-based language.

- **Polarity:** a binary label (1 = phishing, 0 = legitimate), assigned at the point of collection based on the forwarding recipient email's content identification.

Data was initially assembled in tabular spreadsheet form, then converted to CSV format and passed through UTF-8 normalization to remove nonstandard or corrupted characters prior to computational processing.

3.4 Class Distribution

The dataset's overall class balance skews toward phishing examples, consistent with the collection method (participants were specifically asked to forward suspected phishing emails, rather than a random sample of all received email). This design choice, and its implications for the reported evaluation metrics, is discussed further in Sections 5 and 7.

4 Methodology

4.1 Preprocessing Pipeline

Preprocessing followed a standard text-classification pipeline consisting of the following stages:

1. **Corpus formatting.** Raw tabular data was converted to CSV and UTF-8 encoded.
2. **Tokenization.** Email content was tokenized on sentence boundaries (full stops) and whitespace, converting each message into a list of discrete word-level tokens.
3. **Bag-of-Words (BoW) feature extraction.** Each message was represented as a fixed-length numeric vector using scikit-learn's Count Vectorizer, with a maximum vocabulary size of 5,000 features, counting the frequency of each vocabulary word's occurrence in the message.
4. **Label encoding.** Polarity labels were encoded numerically (0 = legitimate, 1 = phishing) for use in supervised training.

4.2 Model: Support Vector Machine

A Support Vector Machine (SVM) is a supervised learning algorithm that seeks the maximum margin hyperplane separating classes in feature space. For a linear kernel, given training vectors $x_i \in R^n$ with labels $y_i \in \{-1, 1\}$, the SVM solves:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{s.t.} \quad y_i (w^T x_i + b) \geq 1 \quad \forall i \quad (1)$$

where w is the weight vector defining the separating

hyperplane and b is the bias term. The linear kernel was chosen given its consistently strong reported performance in prior phishing and spam classification literature [4, 6], its comparatively low computational cost, and its suitability for the vocabulary-sized (5,000 -dimensional), moderately sized feature space produced by BoW extraction on a dataset of this scale.

4.3 Experimental Setup

The dataset was split into training and test sets using an 80/20 split via scikit-learn's train test split function (test size = 0.20). A linear kernel Support

Vector Classifier (SVC) was trained with de-fault regularization ($C = 1.0$). The implementation used Python 3, scikit-learn, pandas, and NumPy, within a Jupyter Notebook environment running on a standard 64-bit Windows 10 machine with 8GB RAM deliberately modest hardware, chosen to reflect resource constraints realistic for researchers and practitioners in similarly resourced settings.

4.4 Evaluation Metrics

Model performance was evaluated using standard classification metrics computed from the confusion matrix:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives respectively. In the phishing-detection context, recall on the phishing class is of particular operational importance, since a false negative (a phishing email misclassified as legitimate) represents a missed attack, while a false positive (a legitimate email misclassified as phishing) typically

carries a lower cost, such as a delayed or manually reviewed message.

5 Results and Comparative Analysis

5.1 Classification Performance

Table 1 reports precision, recall, and F1-score for both classes on the held-out test set ($n = 325$), along with overall and macro/weighted averaged accuracy.

Table 1: SVM Classification Report on Test Set ($n = 325$)

Class	Precision	Recall	F1	Support
Legitimate (0)	0.98	0.89	0.93	102
Phishing (1)	0.95	0.99	0.97	223
Accuracy		0.96		325
Macro avg	0.97	0.94	0.95	325
Weighted avg	0.96	0.96	0.96	325

The model achieves 0.99 recall on the phishing class, meaning only a small fraction of phishing emails in the test set were misclassified as legitimate. Precision on the phishing class (0.95) indicates a modest false positive rate, some legitimate emails were flagged as phishing, which, in a deployed system,

would need to be weighed against the cost of manual review versus the cost of a missed attack. Recall on the legitimate class is comparatively lower (0.89), meaning a somewhat larger proportion of legitimate emails were misclassified as phishing than vice versa; this asymmetry is consistent with, and likely partly explained by, the

class imbalance in the test set (223 phishing vs. 102 legitimate examples), discussed further in Section 7. A closer reading of the confusion matrix suggests that most legitimate class errors involve messages containing marketing style language, discount offers, urgency phrasing, and embedded links, that overlaps lexically with genuine phishing content, a known source of ambiguity in Bag-of-Words-based classification more broadly, since BoW representations are insensitive to the broader structural and contextual cues (such as sender domain reputation) that a human reader would typically use to

disambiguate such cases.

5.2 Comparative Context: English-Language Benchmarks

Table 2 situates this result against recent published phishing detection systems evaluated on established benchmark datasets. Direct numerical comparison across studies should be interpreted cautiously, since these systems are trained and evaluated on different datasets and class distributions; the comparison is included to characterize broad performance trends across approaches, not to claim a controlled, apples-to-apples benchmark result.

Table 2: Reported Accuracy Across Recent Phishing Email Detection Studies

Study	Approach	Accuracy
This work (2020)	Linear SVM + BoW	0.96
Fette et al. [4]	SVM (PILFER)	~0.96
Altwaijry et al. [7]	1D-CNN + Bi-GRU	~0.97-0.98
Songaila'ite et al. [8]	DistilBERT	/ >0.985
	RoBERTa	
Nguyen et al. [9]	BERT (transfer)	~0.97

The linear SVM baseline evaluated on the local, natively collected corpus performs competitively with, and in some cases within only a few percentage points of, considerably more complex and data hungry architectures reported on established benchmarks. This is consistent with Thakur et al.'s observation [10] that architectural gains over classical baselines in phishing detection are frequently modest, and it suggests that for regionally underrepresented settings in particular, investment in representative data collection may offer a better return on effort than investment in model complexity alone.

6 Discussion: Toward Data Centric AI Safety

The results in Section 5 support a broader argument relevant to AI safety practice beyond phishing detection specifically. Much of the public and technical discourse around AI safety centers on model level interventions, alignment techniques, red teaming, interpretability, and evaluation of frontier model capabilities. These are necessary areas of work. However, the comparative results in this paper, together with the dataset bias literature reviewed in Section 2.3, point to a complementary and comparatively under-resourced safety concern: *data*

centric safety, in which the representativeness of training data relative to the deployment population is treated as a first order safety property, not an afterthought to be addressed once model architecture is finalized.

This distinction has concrete consequences in security relevant domains. A phishing detection system trained exclusively on Enron derived or other Western sourced corpora may perform well on held out data drawn from the same distribution, while silently underperforming, in ways that are not visible from its reported benchmark accuracy, against phishing emails that reference local banks, government programs, or culturally specific scam narratives common in Pakistan or other underrepresented regions. Because such underperformance would not appear in the model's published evaluation metrics (which are computed on the same narrow distribution the model was trained on), it constitutes exactly the kind of silent, hard to detect failure mode that AI safety research aims to surface and address. Users in regions whose data is absent from training corpora are not protected by the same safety guarantees implied by a model's reported accuracy, even when that accuracy figure is fully honest and correctly computed.

We do not claim that the dataset or SVM baseline

presented in this paper resolves this problem; at 1,630 examples, it is a modest, foundational contribution rather than a solution at scale (see Section 7). What this paper does argue is that the problem is empirically real, under discussed relative to model centric safety work, and tractable: it requires deliberate, often unglamorous data collection effort of the kind described in Section 3, rather than a fundamentally new algorithmic breakthrough. Extending this kind of regionally grounded data collection to other underrepresented threat domains, and treating that extension as a genuine AI safety contribution, not merely a data engineering exercise, is, in our view, one of the more immediately actionable paths toward safer, more equitable AI deployment in security critical applications.

7 Limitations

We identify the following limitations. Several were not addressed in the original study underlying this paper and are presented here in the interest of transparency and to inform future work:

- **Dataset scale.** At 1,630 messages, the corpus is small relative to contemporary benchmark datasets, some of which now exceed tens of thousands of samples [15]. This limits the statistical confidence of the reported metrics and constrains the model's exposure to the diversity of phishing techniques that exist even within a single regional context.
- **Class imbalance.** The test set contains 223 phishing versus 102 legitimate emails (roughly 2:1), a direct consequence of the collection method (participants were asked specifically to forward suspected phishing emails). This imbalance likely inflates recall on the majority (phishing) class and depresses recall on the minority (legitimate) class; a class balanced re-evaluation, or evaluation using metrics less sensitive to class imbalance (e.g., Matthews's correlation coefficient), was not performed in the original study and would strengthen future work.
- **No controlled deep-learning comparison on the same data.** The comparison in Table 2 is cross dataset rather than controlled; a CNN or transformer model was not trained directly on this corpus. The claim that classical methods are competitive *on this specific dataset* (rather than only in aggregate across the broader literature) therefore remains untested and is an important direction for future work.

- **Temporal drift.** Phishing tactics evolve quickly, and large language models have measurably changed the phishing threat landscape since this dataset was collected in 2018–2020, enabling more sophisticated, context-aware, and human like phishing content generation [16]. A corpus collected before this shift will not reflect current generation, AI-generated phishing content, and detection systems trained on it may generalize poorly to such content without retraining or augmentation.

- **Language scope.** The dataset consists entirely of English language phishing emails collected within Pakistan. It does not include Urdu language or code-mixed content, which remains an open and unexplored direction for future regional data collection work.

- **Label provenance and potential annotation noise.** Labels were assigned based on the forwarding recipient's own identification of an email as phishing or legitimate, rather than independent verification through, for example, URL sandboxing, domain reputation lookups, or expert adjudication. This may introduce labeling noise, particularly for borderline cases such as aggressive marketing emails that share surface features with phishing content.

- **Absence of adversarial or robustness evaluation.** The evaluation reported here measures performance on the original test distribution only. No adversarial perturbation testing (e.g., synonym substitution, homoglyph attacks, or paraphrased phishing content) was conducted, which limits confidence in the model's robustness against adaptive attackers.

8 Conclusion and Future Work

This paper revisited an original, Pakistan specific phishing email corpus and a linear kernel SVM classifier, achieving 96% accuracy and a macro-F1 score of 0.95 on a held-out test set. Using this result, together with a comparative reading of recent phishing and spam detection literature, we argued for a shift in emphasis within phishing detection research and, more broadly, within AI safety practice: from continued architectural refinement evaluated on a small set of well-worn benchmark datasets, toward deliberate, sustained investment in representative, regionally diverse training data. The comparative

analysis in Section 5 suggests that a carefully constructed classical baseline trained on natively collected local data can approach the performance of substantially more complex deep learning systems trained on Western sourced corpora, reinforcing the view that data representativeness, not model sophistication alone, is frequently the more binding constraint on safe and equitable AI deployment in security critical domains.

Future work arising from this study should prioritize: (1) substantially expanding the corpus in size and temporal currency, including recent large language model generated phishing samples, to address the temporal drift limitation identified in Section 7; (2) training and evaluating deep learning and transformer based models directly on this corpus under controlled, class balanced conditions, in order to isolate the effect of data locality from model architecture rather than relying on cross dataset comparison; (3) independent label verification using domain reputation and URL analysis tools, rather than recipient self-report alone, to reduce potential annotation noise; and (4) adversarial robustness testing against paraphrased or LLM generated phishing variants, to assess whether the reported performance holds under adaptive attack conditions. We view this combination of expanded, natively collected regional data and rigorous, controlled evaluation as a concrete and tractable research agenda for advancing data centric AI safety in regionally underrepresented security domains.

REFERENCES

- Anti-Phishing Working Group, "Phishing Activity Trends Report," APWG, 2014.
- Federal Bureau of Investigation, Internet Crime Complaint Center, "2024 Internet Crime Report," IC3 Annual Report, 2025.
- Anonymous, "A feature-engineered dataset of benign and phishing URLs for machine learning and large language models evaluation," *ScienceDirect*, 2025.
- I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in *Proc. 16th Int. Conf. World Wide Web (WWW)*, 2007, pp. 649–656.
- R. Basnet, S. Mukkamala, and A. H. Sung, "Detection of phishing attacks: A machine learning approach," in *Studies in Fuzziness and Soft Computing*, vol. 226, 2008, pp. 373–383.
- 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.
- N. Altwaijry et al., "Advancing phishing email detection: A comparative study of deep learning models," *Sensors*, vol. 24, no. 7, p. 2077, 2024.
- M. Songailaite, E. Kankevičiūtė, B. Zhyhun, and J. Mandravickaitė, "BERT-based models for phishing detection," in *Proc. 28th Conf. Information Society and University Studies (IVUS)*, CEUR Workshop Proceedings, 2023.
- T. Nguyen, H. Nguyen, and Q. Nguyen, "Email phishing detection using BERT and transfer learning," *IEEE Access*, vol. 10, pp. 105421–105432, 2022.
- K. Thakur, M. L. Ali, M. A. Obaidat, and A. Kamruzzaman, "A systematic review on deep-learning-based phishing email detection," *Electronics*, vol. 12, no. 21, p. 4545, 2023.
- F. De Nardin et al., "Deep learning-based intrusion detection systems for phishing email detection: A systematic literature review," in *Proc. ICCV Workshops (VisionDocs)*, 2025.
- Anonymous, "Context aware phishing email detection using machine learning and NLP," *arXiv preprint arXiv:2603.27326*, 2024.
- Anonymous, "Bias beyond English: Evaluating social bias and debiasing methods in a low resource setting," *arXiv preprint arXiv:2504.11183*, 2024.
- S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in ML based science," *arXiv preprint arXiv:2207.07048*, 2022.
- Anonymous, "PhreshPhish: A real world, high quality, largescale phishing website dataset and benchmark," *arXiv preprint arXiv:2507.10854*, 2025.

Anonymous, "A systematic literature review of large language models in phishing attack generation and detection," *Computers & Security*, ScienceDirect, 2026.

