Beyond Generic Phishing Detection: Explainable AI for Finance-Adapted Models in Banking and Fintech

Istiaque Bhuiyan and Tanvir Bhuiyan *

Posted Date: 14 May 2026

doi: 10.20944/preprints202605.0920.v1

Keywords: phishing; fintech; AI; cyber risk; SHAP; ML

 doi:10.20944/preprints202605.0920.v1

Article

Beyond Generic Phishing Detection: Explainable AI for Finance-Adapted Models in Banking and Fintech

Istiaque Bhuiyan [1] and Tanvir Bhuiyan [2,*]

[1] La Trobe University, Melbourne, Australia
[2] Murdoch University, Perth, Australia
* Correspondence: tanvir.bhuiyan@murdoch.edu.au

Abstract

Purpose: This study examines whether finance-adapted (FA) phishing detection models improve detection of finance-themed (FT) attacks, whether improvements differ across email and webpage modalities, and whether finance adaptation creates a specialisation–generalisation trade-off. **Design/Methodology/Approach:** A domain-aware framework is developed using email (164,972 instances) and webpage (11,430 instances) datasets. FT and non-finance-themed (NFT) instances are identified using weighted lexicon-based labelling. Generic models are compared with FA models across Logistic Regression, Linear SVC, and Random Forest using F1-score, MCC, balanced accuracy, ROC-AUC, and PR-AUC. Statistical validation employs bootstrap confidence intervals and McNemar's test, while SHAP and permutation importance interpret webpage model behaviour. **Findings:** FA models outperform generic models in FT email classification, confirming that finance-specific semantic cues improve detection. However, gains are weaker and less consistent in webpage classification, where models rely mainly on structural indicators (page rank, Google index, hyperlinks). Results reveal a specialisation–generalisation trade-off: FA models improve in-domain detection but do not consistently outperform generic models on NFT instances, with F1-score declines of -0.0057 to -0.0151 on non-finance subsets. **Practical Implications:** Financial institutions and fintech platforms should deploy domain-adapted detection for email-based threats, where finance-specific linguistic cues yield measurable gains, while maintaining generic or ensemble models for broader webpage phishing coverage. **Originality/Value:** This study introduces a finance-themed, multi-modal, explainable AI framework for phishing detection, demonstrating that domain adaptation depends critically on data modality and feature representation. It provides a novel systematic comparison of generic versus FA phishing detection across both modalities with statistical validation and explainability analysis.

Keywords: phishing; fintech; AI; cyber risk; SHAP; ML

1. Introduction

1.1. Background and Research Problem

Phishing remains one of the most prevalent and financially damaging forms of cyber-enabled deception. By impersonating trusted institutions and exploiting urgency, fear, routine transactions, and compliance behaviour, phishing attacks attempt to induce users to disclose credentials, authorise payments, or interact with malicious links and webpages (Sturman et al., 2024; Almutairi et al., 2025). This threat is particularly important for financial institutions and fintech platforms, where users increasingly conduct banking, payments, lending, investment, and digital wallet transactions through online and mobile interfaces (Jafri et al., 2024; Javaheri et al., 2024).

While phishing is often studied as a general binary classification problem, many real-world attacks are finance-themed (FT). These attacks imitate banks, fintech applications, payment gateways, tax authorities, payroll systems, billing processes, digital wallets, and account verification workflows

 doi:10.20944/preprints202605.0920.v1

(Javaheri et al., 2024; Almutairi et al., 2025). In fintech settings, such attacks are especially consequential because compromised credentials or fraudulent interactions can directly lead to unauthorised payments, account takeover, identity theft, and broader financial cyber risk (Jafri et al., 2024; Javaheri et al., 2024).

The financial and fintech context changes both the language and intent of phishing attacks. FT phishing emails commonly use transactional language, such as payment reminders, invoice notices, refund claims, failed payment alerts, account suspension warnings, and verification requests (Sturman et al., 2024; Almutairi et al., 2025). Similarly, phishing webpages often mimic banking portals, payment interfaces, fintech dashboards, customer service pages, and digital wallet login screens (Javaheri et al., 2024). These recurring patterns suggest that FT phishing is not simply a subset of general phishing, but a domain-specific form of digital deception with distinctive semantic, contextual, and structural characteristics (Javaheri et al., 2024; Jafri et al., 2024).

Despite this, much of the phishing detection literature remains generic in orientation. Existing models are often trained on aggregated phishing datasets without explicitly distinguishing finance-related attacks from broader phishing content. As a result, generic models may capture broad malicious patterns but underemphasise finance-specific and fintech-specific cues that are critical for detecting financially oriented attacks (Sturman et al., 2024; Javaheri et al., 2024). This creates an important research gap: whether finance-adapted (FA) phishing detection models can provide measurable advantages over generic models in FT phishing contexts (Javaheri et al., 2024; Jafri et al., 2024).

A second limitation concerns data modality. Phishing attacks are often multi-stage, beginning with an email lure and ending with a malicious webpage (Almutairi et al., 2025). However, email and webpage data contain different types of information. Email data captures linguistic and behavioural cues, such as urgency, persuasion, and financial impersonation, whereas webpage data captures structural and technical indicators, such as URL composition, domain reputation, hyperlink behaviour, and page configuration (Sturman et al., 2024; Javaheri et al., 2024). Therefore, finance adaptation may not work equally well across both modalities (Javaheri et al., 2024).

A third issue concerns the specialisation–generalisation trade-off. A FA model may become more sensitive to finance-related cues and perform better in FT contexts, but it may also become less effective on non-finance-themed phishing instances. This trade-off is important for financial institutions and fintech platforms because cybersecurity systems must detect both finance-specific attacks and broader phishing threats (Javaheri et al., 2024; Jafri et al., 2024).

Accordingly, this study is guided by the following central research question:

Can finance-adapted phishing detection models improve the detection of finance-themed phishing attacks, and do these improvements differ across email and webpage modalities?

A related question is whether finance adaptation creates a specialisation–generalisation trade-off, where improved performance in FT contexts comes at the cost of weaker performance on non-finance phishing instances (Javaheri et al., 2024; Sturman et al., 2024).

1.2. Research Context

This study examines phishing detection across two complementary data environments: email-based phishing and webpage-based phishing. These two modalities represent different stages of the phishing attack lifecycle and are highly relevant to financial and fintech platforms. A phishing attack may begin with an email that imitates a bank, payment provider, or fintech service, and then direct the user to a fraudulent webpage designed to collect login credentials, payment details, or account recovery information (Ribeiro et al., 2026; Baltuttis & Teubner, 2024).

The email context is particularly suitable for analysing FT phishing because financial and fintech cues are often explicitly embedded in textual content, including references to payments, accounts, invoices, refunds, banking services, digital wallets, and financial institutions. This creates a rich semantic environment in which domain-specific phishing patterns can be learned effectively (Ribeiro et al., 2026; Baltuttis & Teubner, 2024).

 doi:10.20944/preprints202605.0920.v1

In contrast, the webpage context is characterised by structured and technical features, such as URL composition, domain reputation, hyperlink behaviour, page configuration, login forms, redirections, and external resource loading (Shafin, 2025). These indicators are highly relevant to fintech platforms because many phishing webpages attempt to reproduce the interface structure of legitimate digital finance services. However, such features may capture general webpage suspiciousness more strongly than explicit finance-related semantics, making finance adaptation potentially less impactful in webpage classification (Shafin, 2025).

1.3. Research Objectives

Based on the theoretical arguments developed in Section 2, this study pursues three primary research objectives:

- RO1: To examine whether finance-adapted (FA) phishing models outperform generic models in detecting finance-themed phishing attacks.
- RO2: To assess whether the performance improvement associated with finance adaptation differs between email and webpage modalities.
- RO3: To evaluate whether finance adaptation introduces a trade-off between in-domain performance and generalisation to non-finance phishing instances.

To achieve these objectives, the study develops and compares generic and FA models across multiple classifiers and evaluates performance across three subsets: overall, FT, and non-finance-themed (NFT) data.

1.4. Summary of Findings

The empirical analysis yields three main findings.

First, finance adaptation improves performance in the finance-themed email setting. FA models outperform generic models on the FT subset, with statistically significant improvements observed for both Logistic Regression and Linear SVC. This confirms that domain-specific training enhances the detection of finance-related phishing patterns.

Second, the results show a clear modality effect. Finance adaptation produces stronger and more consistent gains in email classification than in webpage classification. This reflects the fact that email data contains richer and more explicit finance-related semantic cues, whereas webpage models rely primarily on general structural indicators.

Third, the study identifies a clear specialisation–generalisation trade-off. While FA models improve performance on FT instances, they do not consistently outperform generic models on NFT data and may even underperform in broader contexts. This demonstrates that domain adaptation enhances in-domain sensitivity but may reduce out-of-domain generalisation.

1.5. Contributions and Paper Structure

This study contributes to the literature in three main ways.

The first contribution is conceptual. It introduces a domain-aware perspective to phishing detection by explicitly distinguishing finance-themed phishing from general phishing. This challenges the conventional assumption that phishing is a homogeneous classification problem and highlights the need to treat financial and fintech-related phishing as a distinct cyber risk category.

The second contribution is empirical. It provides systematic evidence on the effectiveness of FA models, demonstrating that domain adaptation improves performance in targeted FT contexts but does not generalise uniformly across all phishing scenarios.

The third contribution is methodological and practical. It integrates multi-modal analysis and explainable AI (XAI) techniques, including SHAP and permutation importance, to show not only whether finance adaptation improves performance, but also how it changes model behaviour and feature reliance. These insights are particularly relevant for financial institutions, fintech platforms, and cybersecurity teams seeking interpretable and domain-sensitive phishing detection systems.

 doi:10.20944/preprints202605.0920.v1

The remainder of the paper is organised as follows:

Section 2 develops the theoretical framework and hypotheses, focusing on domain-specific phishing behaviour, modality differences, and the specialisation–generalisation trade-off.

Section 3 describes the data sources, preprocessing steps, finance-theme labelling methodology, model development, and evaluation framework.

Section 4 presents the empirical results, including model performance comparisons, cross-modality analysis, trade-off evaluation, and explainability findings.

Section 5 discusses the implications of the findings for theory and practice, followed by limitations and future research directions.

Section 6 concludes the study by summarising key insights and contributions.

2. Literature Review and Hypotheses Development

2.1. Finance-Themed Phishing as a Distinct Form of Digital Deception

Phishing has long been recognised as a major form of cyber-enabled deception in which attackers impersonate trusted entities to manipulate users into disclosing credentials, clicking malicious links, or performing unauthorised actions (Gallo et al., 2024; Butavicius et al., 2022). Existing research shows that phishing succeeds not only because of technical spoofing, but also because it exploits human trust, routine behaviour, urgency, and perceived institutional legitimacy (Naqvi et al., 2023; Schmitt & Flechais, 2024). In many real-world settings, these deceptive strategies are embedded in financially consequential contexts such as banking alerts, payment confirmations, invoice notices, payroll messages, refund claims, tax notifications, and account verification prompts (AlBenJasim et al., 2024; Laxman et al., 2024). This financial orientation is important because it gives phishing attacks a more specific semantic and behavioural structure than generic phishing attempts (Loggen et al., 2024). FT phishing messages often rely on transactional vocabulary, references to monetary loss or delay, warnings about account suspension, and requests for immediate action tied to payments or account access (Căciulescu et al., 2024). These cues are not merely incidental features of the message; rather, they form part of the attacker's persuasion strategy by imitating familiar financial workflows and exploiting users' sensitivity to financial risk and time pressure (Jabir et al., 2025; Ribeiro et al., 2024). As a result, FT phishing can be viewed not simply as a random subset of phishing data, but as a domain-specific manifestation of phishing with recurring linguistic, contextual, and institutional patterns (Alawida et al., 2022). However, much of the phishing detection literature has treated phishing as a broad binary classification problem, typically distinguishing phishing from legitimate content without paying close attention to thematic subdomains (Safi & Singh, 2023; Kavya & Sumathi, 2024). While this generic framing has helped establish strong baseline models, it also risks overlooking the possibility that different phishing domains may exhibit distinct predictive signatures (Opara et al., 2024). In particular, a model trained on broad phishing corpora may learn general malicious indicators while underemphasising finance-specific cues that are especially relevant in financially themed attacks (Doshi et al., 2023; Asiri et al., 2024).

This concern is important because the objective of classification is not only to distinguish malicious from benign content in the abstract, but also to capture the most informative patterns within the target decision environment (Korkmaz, 2026). The logic for domain adaptation follows directly from this point. In supervised classification, when a target subdomain contains recurrent semantic or structural regularities, training models with greater emphasis on that subdomain can improve sensitivity to the most diagnostically relevant signals (Li et al., 2022). Applied to phishing, this suggests that a FA model may be better positioned than a generic model to recognise vocabulary, message framing, and contextual triggers associated with financial impersonation and transactional deception (Zhang et al., 2022). In other words, if FT phishing systematically differs from generic phishing in the way it is written or structured, then a classifier exposed more directly to such content should be better able to detect it (Tanveer et al., 2023; Li et al., 2022). This expectation is also consistent with the broader literature on domain-specific analytics, which argues that specialised learning can

 doi:10.20944/preprints202605.0920.v1

improve predictive performance when the target class is characterised by distinctive cues that are diluted in more heterogeneous training data (Li et al., 2022). In cybersecurity settings, such an argument implies that financially themed attacks may benefit from detection models that are tuned to the semantics and contexts of financial deception rather than relying solely on generic maliciousness patterns (Zhang et al., 2022; Tanveer et al., 2023). Therefore, the first expectation of this study is that FA models will achieve stronger performance on FT phishing tasks than models trained more generically.

H1: *Finance-adapted phishing models outperform generic phishing models on finance themed phishing classification tasks.*

2.2. Why Finance Adaptation May Differ Across Email and Webpage Modalities

Although phishing is often discussed as a single attack category, its operational manifestation is typically multi-stage and multi-modal (Bustio-Martínez et al., 2025; Opara et al., 2024). A phishing attack may begin with an email designed to attract attention and create urgency, then direct the victim to a webpage that imitates a legitimate institution and attempts to harvest credentials or other sensitive information (Asiri et al., 2024; Schmitt & Flechais, 2024). These two components belong to the same deceptive process, but they contain different forms of information and therefore may not benefit equally from domain adaptation (Kavya & Sumathi, 2024). Email phishing primarily operates through language and communication framing (Gallo et al., 2024). The email body and subject line can directly encode urgency, financial threat, refund opportunities, account verification demands, missed payment warnings, or institution-specific service language (Butavicius et al., 2022; Bustio-Martínez et al., 2024).

These verbal and contextual signals are especially important in FT phishing because the attacker often seeks to mimic the exact tone and structure of routine financial communication (Doshi et al., 2023). Thus, in raw email data, finance-specific cues are likely to be relatively visible in the text itself, allowing FA models to learn highly relevant lexical and semantic patterns (Gallo et al., 2024; Butavicius et al., 2022).

Webpage phishing, by contrast, often relies on a different set of indicators (Safi & Singh, 2023). Rather than communicating primarily through extended persuasive text, phishing webpages frequently reveal malicious intent through URL composition, domain anomalies, login forms, external redirects, hyperlink structures, suspicious branding placements, page authority signals, and other structural webpage characteristics (Opara et al., 2024).

These features are highly useful for phishing detection, but many of them are also broadly applicable across different types of phishing rather than being exclusively tied to financial impersonation (Choudhary et al., 2023). As a result, finance specificity may be less directly expressed in webpage representations than in raw email text, particularly when structured webpage features capture general webpage suspiciousness rather than explicit financial semantics (Opara et al., 2024; Safi & Singh, 2023).

This difference matters for theory development. If FT signals are more explicitly embedded in email text than in webpage structure, then finance adaptation should produce stronger gains in email-based phishing classification than in webpage-based classification (Bustio-Martínez et al., 2025). Put differently, email models may benefit more from finance-focused training because they can directly exploit domain-specific vocabulary and contextual phrasing, whereas webpage models may continue to depend largely on general phishing indicators that are useful across thematic domains (Khadka et al., 2026; Opara et al., 2024). The expected benefit of finance adaptation is therefore unlikely to be uniform across modalities (Asiri et al., 2024).

This modality-based argument is supported by the broader literature on multi-modal cybersecurity analytics, which suggests that different feature spaces capture different types of risk signals and that domain adaptation may interact with the informational density of each modality (Ofusori et al., 2024).

 doi:10.20944/preprints202605.0920.v1

In raw textual settings, domain cues are often rich, explicit, and contextually meaningful (Asiri et al., 2024). In structured webpage settings, domain cues may be present but partially mediated by broader technical indicators that are less uniquely tied to the financial theme (Khadka et al., 2026; Opara et al., 2024). Therefore, while finance adaptation may still improve webpage detection, its relative effect is expected to be stronger in email classification than in webpage classification.

H2: *The performance improvement associated with finance adaptation is stronger in raw email phishing classification than in raw webpage phishing classification.*

2.3. Specialisation Benefits and the Generalisation Trade-Off

Although domain adaptation can improve classification within a targeted subdomain, specialised learning often involves a trade-off between precision within the focal domain and breadth outside it (Haider Rizvi et al., 2025; Li et al., 2022). Models trained on broader datasets benefit from exposure to a wider variety of patterns, allowing them to capture more general signals that remain useful across diverse contexts (Zhang et al., 2022). By contrast, models adapted to a narrower theme may become more sensitive to the cues most relevant within that theme, but less responsive to patterns that fall outside it (Tanveer et al., 2023; Zhang et al., 2022). This trade-off is highly relevant in phishing detection because phishing attacks are heterogeneous in topic, style, tone, and institutional impersonation strategy (Tjingaete & Juremi, 2023). Some phishing messages imitate banks and payment providers, whereas others impersonate delivery services, social media platforms, government agencies, technical support channels, or workplace systems. A FA model may therefore become especially effective at identifying messages framed around accounts, payments, invoices, refunds, and transaction risk, yet not necessarily outperform more generic models when confronted with non-finance phishing cues (Kavya & Sumathi, 2024; Doshi et al., 2023). Theoretically, this reflects a standard bias-variance type tension in domain-focused classification (Korkmaz, 2026). Greater adaptation to a specialised target environment can improve fit to the target distribution, but it can also narrow the classifier's effective scope by overweighting the patterns most common in that environment (Li et al., 2022; Tanveer et al., 2023). In the present context, a FA phishing model is expected to improve in-domain detection performance because it is tuned to financially relevant content, but this same specialisation may reduce its comparative advantage on broader phishing instances that lack finance-related features (Rashid et al., 2024; Hannousse & Yahiouche, 2021). This argument is also important from an applied perspective. Security teams may prefer specialised detectors when protecting highly sensitive environments such as banking, fintech, online payments, payroll, or digital account systems (AlBenJasim et al., 2024). However, a model optimised for FT phishing should not automatically be assumed to dominate a generic phishing model in all settings (Laxman et al., 2024; Loggen et al., 2024).

Instead, its main strength should lie in more accurate detection of the target domain, while performance outside that domain may remain unchanged or even decline relative to a broadly trained classifier (Căciulescu et al., 2024). Accordingly, the expected contribution of finance adaptation is domain-specific rather than universal (Rashid et al., 2024). The relevant question is not whether FA models are always better, but whether they provide measurable gains where FT cues are present and decision relevance is highest (Rashid et al., 2024; Hannousse & Yahiouche, 2021). This leads to the final hypothesis.

H3: *Finance-adapted phishing models outperform generic phishing models on finance-themed instances but do not consistently outperform them on non-finance instances.*

2.4. Conceptual Framing of the Study

Taken together, the preceding arguments suggest that phishing detection should not always be approached as a fully generic classification problem (Gallo et al., 2024). When phishing content is embedded in a specific institutional and transactional environment, such as finance, the attack may

 doi:10.20944/preprints202605.0920.v1

contain recurring domain cues that justify a more targeted modelling strategy (AlBenJasim et al., 2024; Li et al., 2022). At the same time, the value of this strategy is expected to depend on the modality through which the attack is observed and on the trade-off between specialised sensitivity and broader coverage (Bustio-Martínez et al., 2025). The present study builds on this reasoning by comparing generic and FA phishing classifiers across two complementary data settings: raw email data and raw webpage data. This design allows the study to examine three linked questions: whether finance adaptation improves performance on FT phishing content, whether the strength of that improvement differs between email and webpage modalities, and whether specialisation is associated with weaker relative performance outside the target finance domain. In this way, the study contributes to phishing research by introducing a domain-aware perspective that is both theoretically grounded and practically relevant to financial cybersecurity contexts.

3. Data and Methodology

3.1. Research Design

This study adopted a supervised machine learning design to investigate phishing detection across two attack surfaces: phishing emails and phishing webpages. The methodology was designed to compare generic phishing detection models with FA phishing detection models. The central aim was to examine whether models trained with additional emphasis on FT phishing and finance-related legitimate samples perform better in financial phishing contexts than models trained on general phishing data only. The workflow consisted of six main stages: dataset preparation, preprocessing, FT labelling, construction of generic and FA training sets, model training and evaluation, and SHapley Additive exPlanations (SHAP)-based explainability analysis. Figure 1 below illustrates the workflow diagram of this study.

 doi:10.20944/preprints202605.0920.v1

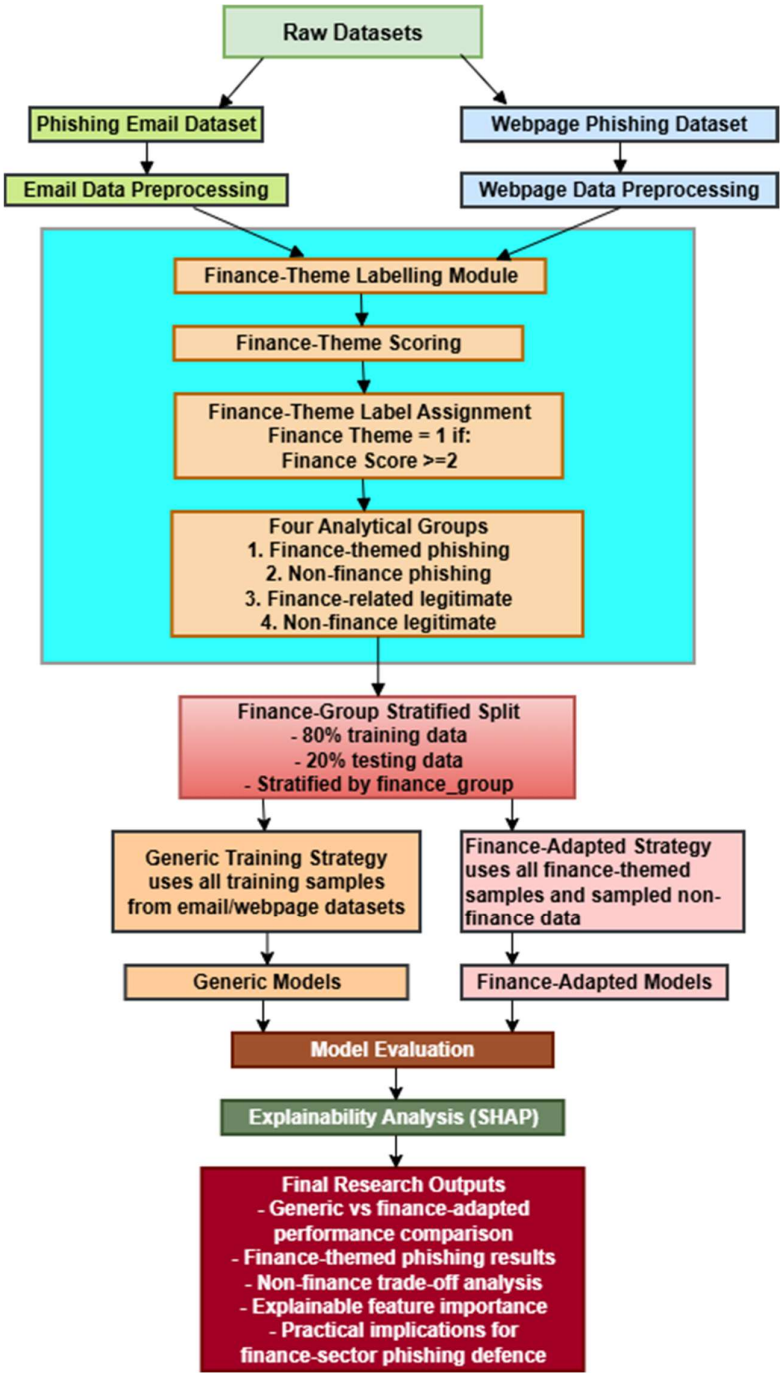

Figure 1. Research Workflow Diagram.

3.2. Datasets

Two phishing datasets were used in this study: a phishing email dataset and a webpage phishing detection dataset. Together, these datasets allowed phishing detection to be examined from both email-content and webpage-structure perspectives.

3.2.1. Phishing Email Dataset

The phishing email dataset (see Table 1 below) was obtained from Kaggle and is available under the CC BY-SA 4.0 licence. The dataset was compiled from multiple public email sources, including Enron, Ling, CEAS, Nazario, Nigerian Fraud, and SpamAssassin. It contains both email text and, for

 doi:10.20944/preprints202605.0920.v1

some subsets, contextual metadata such as sender, receiver, date, and URLs. The dataset has been associated with recent work on interpretable and robust phishing email detection by Al-Subaiey et al. (2024).

Table 1. Phishing Email Dataset.

Email Source	Number of Emails	Main Available Fields
CEAS_08	39,154	Sender, receiver, date, subject, body, label, URLs
Enron	29,767	Subject, body, label
Ling	2,859	Subject, body, label
Nazario	1,565	Sender, receiver, date, subject, body, URLs, label
Nigerian Fraud	3,332	Sender, receiver, date, subject, body, URLs, label
SpamAssassin	5,809	Sender, receiver, date, subject, body, label, URLs
Total	**82,486**	**Combined email dataset**

The phishing email dataset contains approximately 82,486 emails, consisting of 42,891 spam/phishing emails and 39,595 legitimate emails. This gives a relatively balanced distribution (Table 2 below) of approximately 52.00% phishing/spam and 48.00% legitimate emails. This dataset was suitable for training and evaluating text-based phishing detection models because it contains email subject lines, body text, and binary phishing/legitimate labels across multiple sources.

Table 2. Class Distribution of the Phishing Email Dataset.

Class	Number of Emails	Percentage
Phishing/spam	42,891	52.00%
Legitimate	39,595	48.00%
Total	**82,486**	**100.00%**

3.2.2. Webpage Phishing Detection Dataset

The second dataset (see Table 3 below) was the Webpage Phishing Detection Dataset, originally published by Hannousse and Yahiouche (2021) on Mendeley Data. It contains 11,430 URLs and 87 extracted features designed for machine learning-based phishing webpage detection. The dataset is balanced (see Table 4 below), containing 5,715 phishing URLs and 5,715 legitimate URLs. The 87 features are grouped into three categories: URL structure and syntax features, webpage content features, and external-service-based features. This dataset was suitable for structured-feature phishing detection because it provides engineered tabular variables rather than raw webpage text only.

Table 3. Phishing Webpage Dataset.

Feature Category	Number of Features	Description
URL structure and syntax features	56	Features extracted from the URL, such as URL length, hostname length, number of dots, hyphens, slashes, special characters, IP address use, and digit ratio
Webpage content features	24	Features extracted from webpage content, including hyperlinks, iframe use, login forms, favicon use, anchor safety, and pop-up windows
External-service features	7	Features obtained from external services, such as WHOIS, domain age, DNS records, Google index status, web traffic, and page rank

 doi:10.20944/preprints202605.0920.v1

Table 4. Class Distribution of the Phishing Webpage Dataset.

Class	Number of URLs	Percentage
Phishing	5,715	50.00%
Legitimate	5,715	50.00%
Total	**11,430**	**100.00%**

3.3. Data Preprocessing

3.3.1. Email Data Preprocessing

The email datasets were first standardised into a common structure because the original sources contained different column formats. Column names were normalised by converting them to lowercase and removing leading or trailing spaces. The subject and body fields were extracted where available. If either field was missing, it was replaced with an empty string to prevent processing errors. The subject and body were then combined into a single text field. This combined field was used as the main input for the email classification models. The original class labels were also standardised. Labels such as spam, phishing, malicious, fraud, and 1 were mapped to 1, representing phishing or spam. Labels such as legitimate, ham, benign, and 0 were mapped to 0, representing legitimate emails. The final standardised email dataframe contained three core columns: source, text, and label_num. Table 5 illustrates the preprocessing pipeline of Email dataset.

Table 5. Preprocessing Pipeline of Email Dataset.

Preprocessing Step	Description	Purpose
Column standardisation	Converted column names to lowercase and stripped extra spaces	Ensured consistent column handling across datasets
Missing text handling	Replaced missing subject/body fields with empty strings	Prevented text-processing errors
Text construction	Combined subject and body into one text field	Created a unified text input for modelling
Label standardisation	Mapped phishing/spam labels to 1 and legitimate labels to 0	Created a consistent binary target variable
Source tracking	Added a source column	Preserved dataset origin for subgroup and cross-source analysis

3.3.2. Webpage Data Preprocessing

The webpage dataset required less preprocessing because it was already provided as a structured feature dataset. Column names were standardised, and the target column status was converted into a binary label, where phishing was mapped to 1 and legitimate was mapped to 0. The raw url column was retained for FT labelling but excluded from the baseline structured-feature modelling stage. This was done to ensure that the webpage models used the extracted numerical and categorical webpage features rather than raw URL text. Table 6 illustrates the preprocessing pipeline of Webpage dataset.

 doi:10.20944/preprints202605.0920.v1

Table 6. Preprocessing Pipeline of Webpage Dataset.

Preprocessing Step	Description	Purpose
Column standardisation	Converted column names to lowercase and removed extra spaces	Ensured consistent processing
Target encoding	Mapped phishing to 1 and legitimate to 0	Created a binary classification target
Raw URL handling	Used URL for FT labelling but excluded it from baseline structured modelling	Prevented mixing raw text with structured feature modelling
Feature selection	Removed target, label, FT, and grouping columns from model input	Prevented data leakage

3.4. Finance-Theme Labelling Strategy

As displayed in Table 7 below, a rule-based FT labelling approach was developed to identify emails and URLs containing financial, banking, payment, or transaction-related content. This step was important because the study aimed to evaluate phishing detection not only in general, but also within finance-related attack contexts. The finance-theme dictionary consisted of four groups: core finance terms, finance brands, finance-phishing context phrases, and generic phishing action terms.

Table 7. Finance-Theme Labelling Dictionary.

Dictionary Group	Example Terms or Phrases	Purpose
Core finance terms	bank, account, payment, invoice, refund, credit card, transaction, transfer, tax, loan, insurance	Captures general financial language
Finance brands and institutions	PayPal, Stripe, Wise, Western Union, Bank of America, HSBC, Commonwealth Bank, NAB, ANZ, Westpac, Visa, Mastercard	Identifies references to or impersonation of financial institutions and payment platforms
Finance-phishing context phrases	verify your account, account suspended, payment failed, update your billing, unauthorized transaction, pending refund, wire transfer	Captures strong finance-related phishing scenarios
Generic phishing action terms	verify, confirm, update, login, reset, urgent, action required, click here, security alert	Captures urgency and credential-request language commonly used in phishing

Before keyword matching, text was cleaned by converting it to lowercase, removing raw URLs from the text copy, removing punctuation and special characters, and reducing repeated whitespace. This made the keyword matching process more consistent and transparent.

A weighted finance score (See Table 8) was then calculated for each record:

$$Finance\ score = (finance\ keyword\ hits \times 1) + (finance\ brand\ hits \times 2) + (finance\ context\ phrase\ hits \times 3) + (phishing\ action\ hits \times 1)$$

Table 8. Finance-Score Methodology.

Component	Weight	Rationale
Finance keyword hits	1	Indicates general financial relevance
Finance brand hits	2	Captures potential impersonation of financial institutions or payment services
Finance context phrase hits	3	Represents stronger evidence of finance-related phishing behaviour
Phishing action hits	1	Captures urgency or action-based phishing language

 doi:10.20944/preprints202605.0920.v1

A record was labelled as FT if it met two conditions: Finance Score ≥ 2 and at least one finance-related signal is present. The finance-related signal requirement was used to reduce false positives from generic phishing language. For example, an email containing only urgent or click here would not be labelled as FT unless it also contained financial content such as bank, payment, invoice, or PayPal. The FT labelling process produced the following outputs as displayed in Table 9 below:

Table 9. Finance-Theme Outputs.

Output Column	Description
Finance keyword hits	Number of general finance keywords detected
Finance brand hits	Number of finance brand or institution names detected
Finance context hits	Number of finance-phishing context phrases detected
Phishing action hits	Number of generic phishing action terms detected
Finance score	Weighted finance-theme score
Finance theme	Binary finance-theme label: 1 = finance-themed, 0 = non-finance

3.5. Construction of Finance-Based Analytical Groups

After FT labelling, each email and webpage record was assigned to one of four analytical groups by combining the phishing label with the FT label. These four groups were used for subgroup evaluation. This allowed the study to test whether FA models performed better on FT phishing records while also checking whether adaptation reduced performance on non-finance phishing records.

Table 10. Finance-based Analytical Group.

Phishing Label	FT Label	Group	Interpretation
1	1	Finance phishing	Phishing record with finance-related content
1	0	Non finance phishing	Phishing record without finance-related content
0	1	Finance legitimate	Legitimate record with finance-related content
0	0	Non finance legitimate	Legitimate record without finance-related content

3.6. Generic and Finance-Adapted Model Design

Two model-training strategies were compared: **generic training** and **FA training**.

3.6.1. Generic Training Strategy

The generic training strategy used all available training records. For email data, this meant all available email texts. For webpage data, this meant all available structured webpage samples. This approach represented a conventional phishing detection model trained on broad phishing and legitimate examples.

3.6.2. Finance-Adapted Training Strategy

The FA strategy was designed to increase the model's exposure to finance-relevant phishing and legitimate samples. The FA training set included:

1. all FT phishing records,
2. all finance-related legitimate records,
3. a sampled subset of non-finance phishing and non-finance legitimate records.

 doi:10.20944/preprints202605.0920.v1

A non-finance ratio parameter was used to control the number of non-finance records included relative to the number of FT records. In this study, non-finance ratio = 1.0 was used, meaning that the number of sampled non-finance records was equal to the number of FT records where possible. This design allowed the study to examine whether finance-focused training improves detection in FT phishing contexts.

3.7. Email Model Development

The email models were developed using Term Frequency (TF)–Inverse Document Frequency (IDF) vectorisation and linear classifiers. The text input was the combined email subject and body. TF-IDF was used because it provides an interpretable numerical representation of text by weighting terms according to their importance within a document relative to the whole corpus. Specifically, TF measures how often a term appears in a document, while IDF reduces the weight of common terms and increases the weight of rarer, more informative terms. Thus, TF-IDF helps the model identify words that are more meaningful for distinguishing phishing/spam emails from legitimate emails. Both Logistic Regression (LR) and Linear Support Vector Classifier (LSVC) used class balancing to reduce bias toward the majority class. Since the email models are based on TF-IDF vectorisation, it produces a very high-dimensional and sparse feature space. In such settings, linear classifiers such as LR and LSVC are generally preferred over tree-based methods like Random Forest (RF). Using the TF-IDF vectoriser four email models were constructed as described in Table 11.

Table 11. Email Models.

Model	Feature Representation	Classifier	Training Strategy
Generic LR	TF-IDF	Logistic Regression	Generic
FA LR	TF-IDF	Logistic Regression	FA
Generic LSVC	TF-IDF	Linear Support Vector Classifier	Generic
FA LSVC	TF-IDF	Linear Support Vector Classifier	FA

Note: LR-Logistic Regression; FA-Finance-Adapted; LSVC-Linear Support Vector Classifier.

3.8. Webpage Model Development

The webpage models were developed using the structured features provided in the webpage phishing dataset. The raw URL, target labels, FT labels, finance scores, hit counts, and finance-group labels were excluded from model input to prevent data leakage. Six webpage models were constructed as described in Table 12 below.

Table 12. Webpage Models.

Model	Input type	Classifier	Training strategy
Generic LR	Structured webpage features	Logistic Regression	Generic
FA LR	Structured webpage features	Logistic Regression	FA
Generic LSVC	Structured webpage features	Linear Support Vector Classifier	Generic
FA LSVC	Structured webpage features	Linear Support Vector Classifier	FA
Generic RF	Structured webpage features	Random Forest	Generic
FA RF	Structured webpage features	Random Forest	FA

Note: LR-Logistic Regression; FA-Finance-Adapted; LSVC-Linear Support Vector Classifier; RF-Random Forest.

 doi:10.20944/preprints202605.0920.v1

3.9. Train/Test Splitting and Evaluation Subsets

The email and webpage datasets were split into training and testing sets using an 80/20 holdout design. The split was stratified by finance group to preserve the distribution of the four analytical groups in both the training and testing sets. After splitting, three evaluation subsets were created from the held-out test set as shown in Table 13. This evaluation design enabled the study to assess both the benefits and possible trade-offs of FA phishing detection.

Table 13. Model Evaluation Subset.

Evaluation Subset	Included Records	Purpose
Overall test set	All held-out test records	Measures general phishing detection performance
FT test set	Records where FT = 1	Tests performance on finance-related phishing contexts
NFT test set	Records where FT = 0	Checks whether finance adaptation harms general non-finance phishing detection

Note: FT-Finance-Themed; NFT-Non-Finance-Themed.

3.10. Performance Evaluation Metrics

Model performance was assessed using multiple binary classification metrics (see Table 14 below). This was necessary because phishing detection is not only about overall accuracy; it also requires careful attention to false positives and false negatives. F1-score was used as the main ranking metric in the FT subset comparison tables because it balances precision and recall, which is important when detecting finance-related phishing attacks.

Table 14. Model Evaluation Metrics.

Metric	Description	Relevance to Phishing Detection
Accuracy	Proportion of total correct predictions	Provides general correctness
Precision	Proportion of predicted phishing cases that are truly phishing	Measures false-alarm control
Recall	Proportion of actual phishing cases detected	Measures phishing detection coverage
F1-score	Harmonic mean of precision and recall	Balances missed phishing and false alarms
Balanced accuracy	Average recall across both classes	Useful when class distribution varies
Matthews correlation coefficient	Correlation between predicted and true labels	Robust overall binary classification measure
ROC-AUC	Ability to separate phishing and legitimate records across thresholds	Measures ranking/separation ability
PR-AUC	Precision–recall area under the curve	Useful for phishing-focused evaluation

3.11. SHAP-Based Explainability Analysis

To interpret webpage model behaviour, SHAP-based explainability and permutation importance were applied to the Random Forest webpage models. Random Forest was selected because SHAP is particularly suitable for tree-based tabular models, where feature contributions can be estimated efficiently and interpreted clearly. Two Random Forest models were trained for explanation: a Generic Webpage Phishing Model and an FA Webpage Phishing Model. Mean absolute SHAP values were used to rank feature importance, with a higher mean absolute SHAP

 doi:10.20944/preprints202605.0920.v1

value indicating that a feature had a stronger influence on the model's predictions. In addition, permutation importance was computed for both the Generic Webpage Phishing Model and the FA Webpage Phishing Model to provide a complementary measure of feature relevance based on the decline in predictive performance when a feature's values were randomly shuffled. The SHAP comparison table, together with the permutation importance results, was used to identify whether FA training changed the model's feature reliance compared with generic training. This provided interpretability evidence for the discussion section by showing not only whether finance adaptation changed predictive performance, but also whether it changed the model's decision logic.

SHAP was not applied to the email models because the email classifiers were trained on TF-IDF text representations, which generate a very high-dimensional and sparse feature space. In this setting, SHAP explanations can become computationally expensive and less concise because thousands of individual word or token features may receive separate attribution values. Moreover, the email models used linear classifiers, where model coefficients already provide a more direct and interpretable indication of term importance. Therefore, SHAP was reserved for the webpage models, where the structured tabular features and Random Forest architecture were more suitable for SHAP-based interpretation. The interpretability workflow involved the following steps.

Table 15. The SHAP Workflow.

Step	Description
Train generic model	Trained on all webpage training samples
Train FA model	Trained on FT records plus sampled NFT records
Extract transformed feature matrix	Retrieved pre-processed model input from the pipeline
Compute SHAP values	Calculated feature contributions for predictions on the FT webpage subset
Generate SHAP summary plots	Visualised the most influential features for each model
Create top-feature tables	Ranked features by mean absolute SHAP value
Compare explanations	Compared feature importance between generic and FA models

3.12. Hypothesis Testing Strategy

The hypotheses were evaluated using a structured comparative framework based on model performance across the predefined evaluation subsets (overall, FT, and NFT). For H1, which predicts that finance-adapted (FA) models outperform generic models in finance-themed contexts, model performance was compared on the FT test subset, with particular emphasis on F1-score and Matthews Correlation Coefficient (MCC) as primary evaluation metrics. For H2, which posits that the benefits of finance adaptation differ across modalities, a cross-modality comparison was conducted by examining performance differences between email-based and webpage-based models on the FT subset. For H3, which predicts a specialisation–generalisation trade-off, performance differences between FA and generic models were analysed across both FT and NFT subsets, focusing on whether gains in FT classification were accompanied by declines in NFT performance. To assess statistical significance, bootstrap confidence intervals were computed for performance differences (ΔF1 and ΔMCC), and McNemar's test was applied to paired model predictions on the FT subset to evaluate whether observed differences were statistically meaningful. This multi-level evaluation approach ensured that hypothesis testing captured not only overall performance differences but also domain-specific and cross-modality effects.

 doi:10.20944/preprints202605.0920.v1

4. Results

4.1. Performance of Email Phishing Models Across Evaluation Subsets

Table 16 presents the comparative performance of generic and FA email phishing classifiers across the overall, NFT, and FT evaluation subsets. A consistent pattern is observed across both Logistic Regression and Linear SVC. In the overall and NFT subsets, generic models achieved higher F1 and MCC values than FA models. In contrast, within the FT subset, FA models produced the best results for both model families. This suggests that finance adaptation improves detection accuracy when financially relevant cues are present, but its benefits are not universal across all phishing instances. These findings provide direct support for H1, which predicts that FA models outperform generic models on FT phishing tasks, and for H3, which predicts that FA models perform better in-domain but do not consistently outperform generic models on non-finance instances.

Table 16. Comparative Performance of Generic and Finance-Adapted Email Phishing Models.

Eval. Subset	Generic Model	F1 (Generic)	MCC (Generic)	FA Model	F1 (FA)	MCC (FA)	ΔF1	Better Model
Overall test	Generic LR	0.988	0.976	FA LR	0.984	0.967	-0.004	Generic
Overall test	Generic LSVC	0.996	0.992	FA LSVC	0.989	0.977	-0.007	Generic
NFT test	Generic LR	0.988	0.976	FA LR	0.982	0.965	-0.006	Generic
NFT test	Generic LSVC	0.996	0.993	FA LSVC	0.987	0.974	-0.009	Generic
FT test	Generic LR	0.990	0.971	FA LR	0.993	0.980	0.003	FA
FT test	Generic LSVC	0.996	0.988	FA LSVC	0.997	0.990	0.001	FA

4.2. Finance-Themed Email Classification Performance and Model Ranking

Table 17 shows that finance adaptation yields measurable performance gains in the target domain. Among all evaluated models, FA Linear SVC ranked first, achieving the highest F1 score, MCC, and balanced accuracy. Its performance exceeded that of the best generic alternative by a small but consistent margin, indicating that adaptation to FT content improved detection quality where finance-related cues were most relevant. This finding supports H1 by demonstrating superior performance of FA models on FT phishing tasks. At the same time, when interpreted alongside the overall and non-finance subset results, the table also supports H3, showing that the benefit of finance adaptation is concentrated within the target domain rather than extending uniformly across all phishing instances.

Table 17. Direct Finance Themed Email Performance Table.

Rank	Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	Bal. Accuracy	MCC	N
1	FA LSVC	0.9955	0.9974	0.9958	0.9966	0.9999	1.0000	0.9954	0.9900	4691
2	Generic LSVC	0.9945	0.9961	0.9955	0.9958	0.9999	1.0000	0.9939	0.9876	4691
3	FA LR	0.9910	0.9942	0.9923	0.9933	0.9994	0.9997	0.9904	0.9800	4691
4	Generic LR	0.9870	0.9872	0.9933	0.9902	0.9991	0.9995	0.9839	0.9708	4691

Figure 2 compares the four email classifiers on the FT subset using F1 and MCC. The FA models rank above their generic counterparts, with FA Linear SVC achieving the highest F1 and MCC, followed by Generic Linear SVC, FA Logistic, and Generic Logistic. This provides direct visual support for H1, showing that FA email models achieved the strongest performance on FT phishing classification tasks. The ranked comparison shows a clear ordering in favour of the FA email models.

 doi:10.20944/preprints202605.0920.v1

In particular, the FA Linear SVC model achieved the best overall performance, while the FA Logistic model also outperformed its generic counterpart. This visual pattern strengthens the evidence for H1 by showing that finance-focused training improved classification performance in the FT email setting.

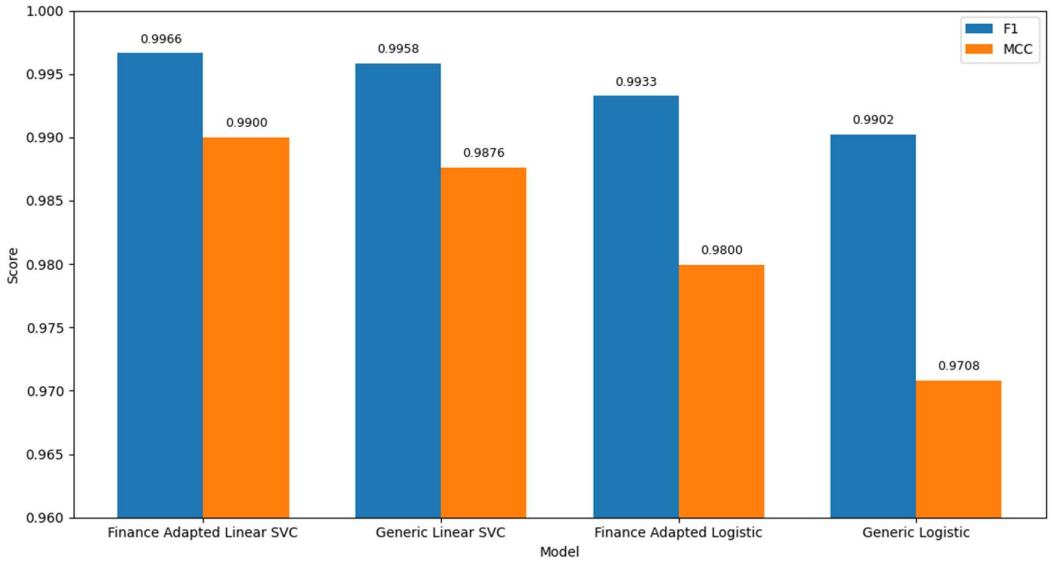

Figure 2. Ranked F1/MCC Comparison for Finance Themed Email Models.

4.3. Performance of Webpage Phishing Models Across Evaluation Subsets

Table 18 shows a mixed but informative pattern across the three evaluation conditions. On the overall test set, the FA models outperformed the generic models across all three model families, with the strongest gain observed for Random Forest, where the FA version achieved the highest F1 score (0.9915). This indicates that finance adaptation improved overall detection performance in the broader evaluation setting. However, on the NFT test subset, the pattern reversed. For Logistic Regression, Linear SVC, and Random Forest, the generic models achieved higher F1 scores than the FA models. This suggests that the FA models became less effective when the phishing instances lacked finance-related cues, indicating weaker generalization outside the target domain. On the FT test subset, the results were more nuanced. The generic models remained slightly stronger for Logistic Regression and Linear SVC, while the FA Random Forest achieved the best result overall with an F1 score of 0.9915, outperforming the generic Random Forest. This means that finance adaptation did not improve performance uniformly across all model families, but it did provide a measurable advantage for the strongest ensemble model in the target finance domain. Taken together, these findings provide partial support for H3. The results clearly show that finance adaptation does not consistently outperform generic modelling across all subsets, especially on the NFT subset, where generic models were superior across the board. At the same time, the FA Random Forest achieved the best performance on the FT subset, suggesting that domain-specific gains are possible when the model family is well suited to capture finance-related phishing patterns. Therefore, the evidence supports the idea that the contribution of finance adaptation is domain-specific rather than universal, although support for H1 is only partial, because FA models did not outperform generic models on the FT subset for every classifier.

 doi:10.20944/preprints202605.0920.v1

Table 18. Comparative Performance of Generic and Finance-Adapted Webpage Phishing Models.

Evaluation Subset	Model Family	Generic F1	FA F1	Better Model
Overall test	LR	0.9359	0.9792	FA
Overall test	LSVC	0.9368	0.9810	FA
Overall test	RF	0.9609	0.9915	FA
NFT test	LR	0.9289	0.9213	Generic
NFT test	LSVC	0.9292	0.9195	Generic
NFT test	RF	0.9594	0.9443	Generic
FT test	LR	0.9880	0.9792	Generic
FT test	LSVC	0.9828	0.9810	Generic
FT test	RF	0.9897	0.9915	FA

4.4. Finance-Themed Webpage Classification Performance and Model Ranking

Table 19 presents the ranking of webpage-based classifiers on the FT subset. The highest-performing model was FA Random Forest Web, which achieved the best values for F1, MCC, recall, and balanced accuracy, indicating that finance-focused adaptation improved detection performance for the Random Forest family in the target domain. However, this advantage was not consistent across all classifier types. The generic Logistic Regression and Linear SVC webpage models both outperformed their FA counterparts, showing that finance adaptation did not produce uniform gains within the FT webpage setting. These results provide partial support for H1, as the FA approach achieved the best overall FT webpage result but did not dominate across all model families. The findings also support H3, since the value of finance adaptation appears conditional and model-specific rather than universal.

Table 19. Direct finance-themed webpage performance table.

Rank	Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	Bal. Accuracy	MCC	N
1	FA RF	0.9844	0.9864	0.9966	0.9915	0.9914	0.9991	0.9316	0.9053	321
2	Generic RF	0.9813	0.9863	0.9931	0.9897	0.9897	0.9989	0.9299	0.8869	321
3	Generic LR	0.9782	0.9830	0.9931	0.9880	0.9606	0.9945	0.9132	0.8667	321
4	Generic LSVC	0.9688	0.9828	0.9828	0.9828	0.9591	0.9944	0.9081	0.8162	321
5	FA LSVC	0.9657	0.9861	0.9759	0.9810	0.9536	0.9935	0.9213	0.8076	321
6	FA LR	0.9626	0.9861	0.9725	0.9793	0.9633	0.9953	0.9196	0.7937	321

4.5. Cross-Modality Comparison of Finance Adaptation Effects

Table 20 compares the effect of finance adaptation across the email and webpage modalities on the FT subset. A clear cross-modality difference is observed. In the email setting, FA models outperformed generic models for both Logistic Regression and Linear SVC, with positive F1 gains in each case. In the webpage setting, however, the pattern was weaker and less consistent: the generic models remained superior for both Logistic Regression and Linear SVC. These findings support H2, which predicts that the performance improvement associated with finance adaptation is stronger in raw email phishing classification than in raw webpage phishing classification. The results suggest that finance-focused training is more beneficial when the model can learn directly from the richer semantic and linguistic cues present in email text, whereas webpage models continue to rely more heavily on broader structural phishing indicators. Overall, the table shows that the gains from finance

 doi:10.20944/preprints202605.0920.v1

adaptation are more consistent in email than in webpage classification, providing direct evidence for the modality-based argument developed in the theory section.

Table 20. Cross-Modality Comparison of Finance-Adaptation Gains in Email and Webpage Phishing Classification.

Modality	Model Family	Generic F1 (FT Subset)	FA F1 (FT Subset)	ΔF1	Better Model
Email	LR	0.9902	0.9933	+0.0030	FA
Email	LSVC	0.9958	0.9966	+0.0008	FA
Webpage	LR	0.9880	0.9792	-0.0088	Generic
Webpage	LSVC	0.9828	0.9810	-0.0018	Generic

Figure 3 represents comparison of finance-adaptation gain by modality. The figure compares the performance gain of FA models over generic models across the two modalities using the common model families, Logistic Regression and Linear SVC. In the email modality, finance adaptation produced positive gains for both classifiers on the FT subset, with ΔF1 = +0.0030 and ΔMCC = +0.0091 for Logistic Regression, and ΔF1 = +0.0008 and ΔMCC = +0.0024 for Linear SVC. In contrast, in the webpage modality, finance adaptation reduced performance for these same model families, with ΔF1 = −0.0088 and ΔMCC = −0.0730 for Logistic Regression, and ΔF1 = −0.0018 and ΔMCC = −0.0086 for Linear SVC. These results show that the benefit of finance adaptation is clearly stronger in email classification than in webpage classification. The formal cross-modality comparison shows that finance adaptation yields positive gains in the email modality but negative gains in the webpage modality for the common model families. For both Logistic Regression and Linear SVC, the FA email models outperformed their generic counterparts on the FT subset, whereas the FA webpage models underperformed relative to the generic webpage models. This provides direct empirical support for H2, indicating that finance adaptation is more effective when models learn from the richer finance-related semantic content of email text than from the more general structural indicators available in webpage data.

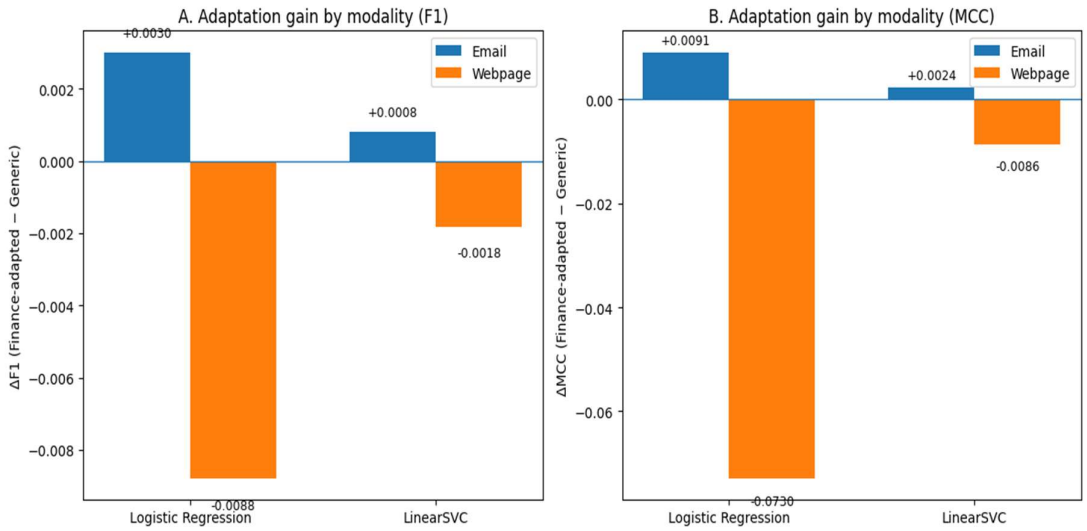

Figure 3. Formal Comparison of Finance Adaptation Gain by Modality.

4.6. Distribution of Finance-Themed Content Across Modalities

Table 21 shows a clear difference in finance-theme coverage between the two modalities. In the pooled email datasets, the FT labelling process identified 23,454 FT instances out of 164,972 total

 doi:10.20944/preprints202605.0920.v1

records, including both finance phishing (15,574) and finance legitimate (7,880) observations. This indicates that the email modality contains a meaningful amount of finance-related semantic content and provides a realistic basis for training and evaluating FA phishing models. In contrast, the webpage dataset showed no FT instances at the dataset-wide labelling stage. All 11,430 webpage records were classified as NFT, and the resulting finance-group distribution contained only non-finance legitimate and non-finance phishing categories. This suggests that the webpage modality in the available data is dominated by general structural phishing indicators rather than explicitly FT content. This is methodologically important for H2 because it helps explain why finance adaptation is expected to be more effective in email phishing classification than in webpage phishing classification. The email datasets contain richer and more directly observable FT cues, whereas the webpage dataset provides little dataset-wide evidence of explicit FT content. As a result, modality differences in finance-adaptation performance are theoretically and empirically plausible.

Table 21. Distribution of Finance-Themed Instances across Email and Webpage Datasets.

Modality	Total Instances	FT Instances	NFT instances	FT (%)	FT label groups present
Email	164,972	23,454	141,518	14.22%	Yes, finance phishing and finance legitimate
Webpage	11,430	0	11,430	0.00%	No, only non-finance phishing and non-finance legitimate

4.7. Generalisation Gap and Specialisation Trade-Off Analysis

Table 22 quantifies the generalisation gap between FT and NFT performance for generic and FA phishing models. In the email modality, FA models exhibit substantially larger positive gaps than generic models, indicating stronger concentration of performance within the finance domain. In the webpage modality, the same trade-off is most clearly visible for Random Forest, where finance adaptation improves the FT subset while reducing performance on the NFT subset. Taken together, these results support H3 by showing that finance adaptation increases domain-specific sensitivity but does not consistently improve performance outside the target finance context.

Table 22. Generalisation-Gap and Trade-Off Analysis for Finance Adaptation.

Modality	Model	Model	NFT F1	FT F1	Gen. Gap	Trade-off Interpretation
Email	LR	Generic	0.9880	0.9902	+0.0023	Small domain gap
Email	LR	FA	0.9823	0.9933	+0.0110	Stronger in-domain concentration
Email	LSVC	Generic	0.9962	0.9958	-0.0004	No meaningful domain concentration
Email	LSVC	FA	0.9870	0.9966	+0.0096	Stronger in-domain concentration
Webpage	LR	Generic	0.9289	0.9880	+0.0592	Strong domain gap
Webpage	LR	FA	0.9213	0.9792	+0.0579	Strong domain gap, but weaker overall
Webpage	LSVC	Generic	0.9292	0.9828	+0.0536	Strong domain gap
Webpage	LSVC	FA	0.9195	0.9810	+0.0615	Stronger in-domain concentration
Webpage	RF	Generic	0.9594	0.9897	+0.0303	Moderate domain gap
Webpage	RF	FA	0.9443	0.9915	+0.0472	Stronger in-domain concentration

Table 23 provides a concise trade-off view of finance adaptation across the FT and non-finance subsets. In the email modality, both Logistic Regression and Linear SVC show the pattern predicted by H3: finance adaptation improves performance on the FT subset but reduces performance on the non-finance subset. In the webpage modality, the same pattern is observed most clearly for Random

 doi:10.20944/preprints202605.0920.v1

Forest, which gains in the FT subset while declining in the non-finance subset. By contrast, webpage Logistic Regression and Linear SVC do not show a positive in-domain gain, indicating only partial support. Overall, the table supports H3 by showing that finance adaptation acts as a domain-specialisation strategy, yielding benefits mainly in the target finance context while weakening broader generalisation.

Table 23. Trade-Off View: Gain/Loss from Finance Adaptation.

Modality	Model Family	ΔF1 (FT Subset)	ΔF1 (NFT Subset)	H3 Pattern
Email	LR	+0.0030	-0.0057	Supports H3
Email	LSVC	+0.0008	-0.0092	Supports H3
Webpage	LR	-0.0088	-0.0076	Does not support in-domain gain
Webpage	LSVC	-0.0018	-0.0097	Weak support
Webpage	RF	+0.0018	-0.0151	Supports H3

Figure 4 shows a clear rank reversal across evaluation subsets, which directly supports H3. In the email panel, the FA models are weaker than the generic models on the overall and NFT subsets but move ahead on the FT subset. This reversal is especially clear for both Logistic Regression and Linear SVC, where finance adaptation improves performance only when finance-related cues are present. In the webpage panel, the same trade-off appears in a more mixed form. FA models do not consistently dominate across all subsets. On the NFT subset, generic models are stronger, while on the FT subset only the FA Random Forest clearly ranks highest; the generic Logistic Regression and Linear SVC remain slightly better than their FA versions. This shows that the benefit of finance adaptation is conditional and does not extend uniformly across all phishing instances or all model families. Overall, this figure provides strong visual evidence for H3, which predicts that FA phishing models should perform better on FT instances but not consistently outperform generic models on NFT instances. The observed reversal across subsets demonstrates the underlying specialisation-versus-generalisation trade-off: finance adaptation increases sensitivity in the target domain, but that gain is accompanied by weaker comparative performance outside that domain.

 doi:10.20944/preprints202605.0920.v1

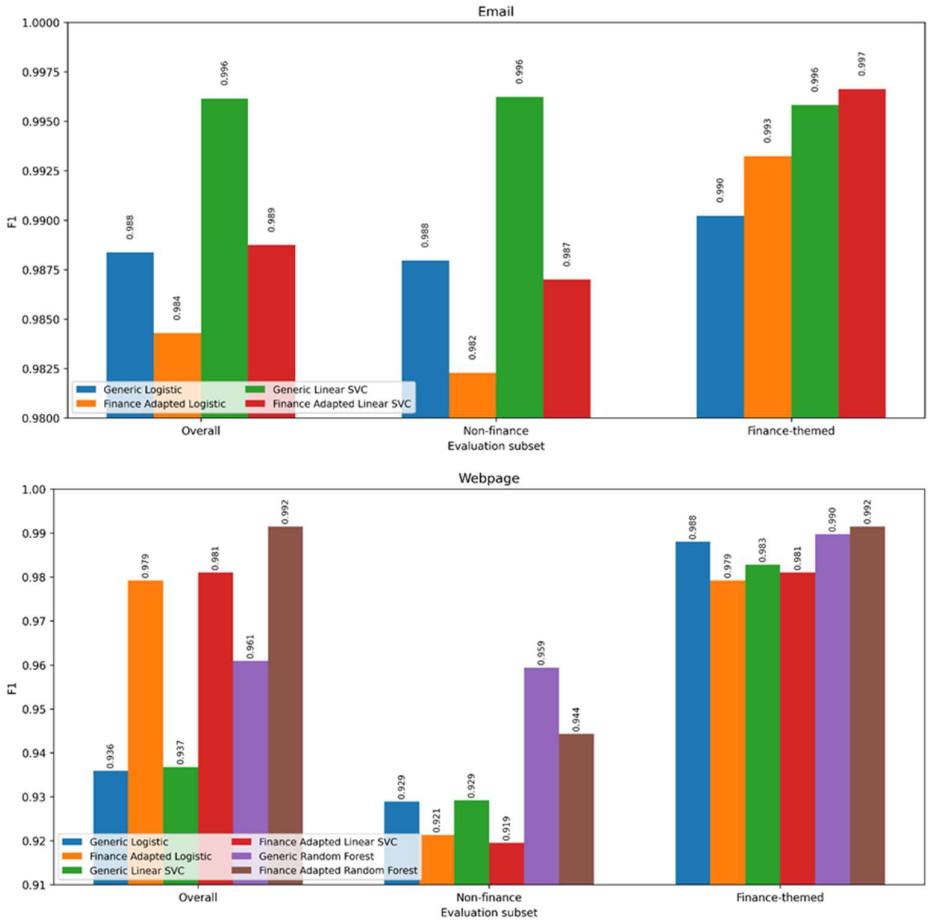

Figure 4. Rank Reversal of Generic and Finance Adapted Models Across Evaluation Subsets.

4.8. Statistical Validation of Finance-Adapted Model Gains

Figure 5 presents the performance gain of the FA email models over the generic models on the FT subset, shown as ΔF1 and ΔMCC with 95% bootstrap confidence intervals. For Logistic Regression, the FA model achieved a ΔF1 of +0.0141 with a 95% CI of [+0.0117, +0.0169], and a ΔMCC of +0.0327 with a 95% CI of [+0.0267, +0.0394]. For Linear SVC, the FA model achieved a ΔF1 of +0.0081 with a 95% CI of [+0.0060, +0.0102], and a ΔMCC of +0.0186 with a 95% CI of [+0.0145, +0.0238]. In all cases, the confidence intervals lie entirely above zero, indicating that the observed gains are statistically reliable rather than random fluctuations. This result is further supported by McNemar's test, which compares paired prediction disagreements between the generic and FA models on the same FT instances. For Logistic Regression, the FA model correctly classified 150 instances that the generic model missed, whereas the generic model correctly classified only 16 instances that the FA model missed ($p < 0.001$). For Linear SVC, the FA model was uniquely correct on 88 instances, compared with 12 instances uniquely correct for the generic model ($p < 0.001$). Together, these findings show that the FA models were not only numerically superior but also significantly more effective on the FT email subset. Accordingly, the figure provides strong additional support for H1, which predicts that FA phishing models outperform generic phishing models on FT phishing classification tasks. While the main results tables show the direction of improvement, this significance analysis demonstrates that the in-domain performance advantage is robust and statistically meaningful.

 doi:10.20944/preprints202605.0920.v1

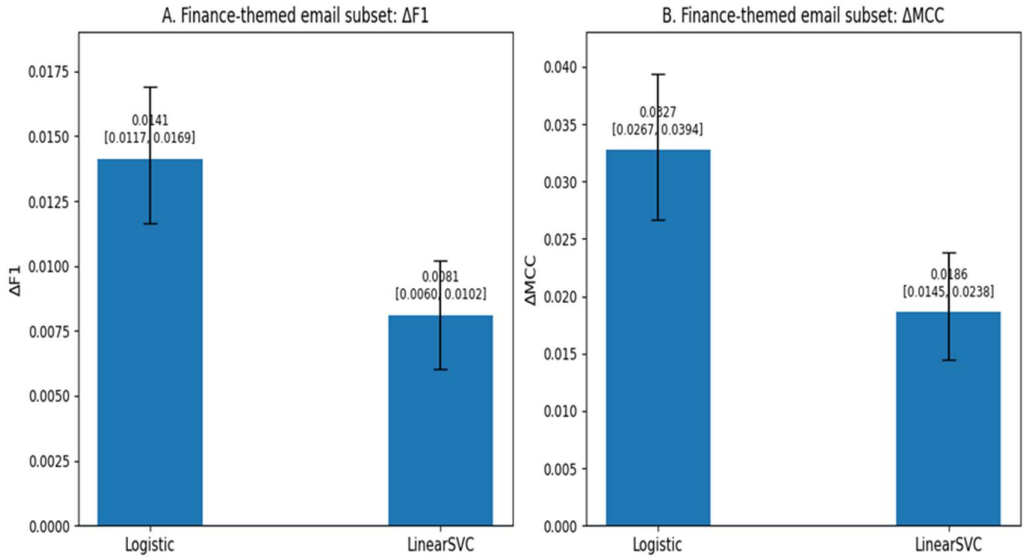

Figure 5. Statistical Validation of Finance Adapted Email Model Gains.

4.9. Explainability Analysis of Webpage Models Using SHAP and Permutation Importance

The Figure 6 SHAP comparison bar chart indicates that both the generic and FA webpage models rely mainly on structural indicators such as Google Index, Page Rank, Web Traffic, Phish Hints, Hyperlinks, Domain Age, and Hostname Length. Because these features reflect general webpage suspiciousness rather than explicit FT semantics, the figure provides strong support for H2. Specifically, it helps explain why finance adaptation produces weaker and less consistent gains in webpage phishing classification than in email phishing classification, where finance-related cues are more directly embedded in the text.

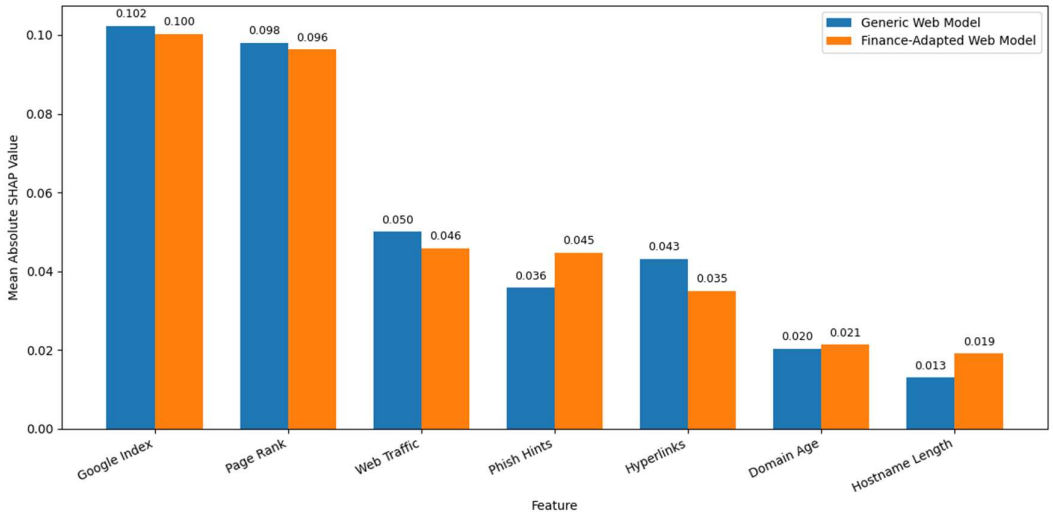

Figure 6. Shap Comparison Bar Chart for The Webpage Models.

Table 23 presents permutation importance results for the generic webpage phishing model. The table shows that the model relied almost exclusively on a small set of broad reputation and indexing signals, with page_rank dominating importance (mean importance = 0.008472), followed distantly by google_index, phish_hints, domain_age, and nb_hyperlinks. Most other features, including many structural phishing indicators, received zero importance. This narrow reliance on general webpage legitimacy cues suggests that the generic model does not learn finance-specific or even strongly

 doi:10.20944/preprints202605.0920.v1

phishing-distinctive patterns. This finding supports H2 by helping to explain why finance adaptation yields weaker gains in webpage classification than in email classification: the generic webpage model is already oriented toward broad suspiciousness signals rather than domain-specific content, leaving limited room for finance-focused improvement unless the adapted model shifts its feature reliance.

Table 23. Permutation Importance Results for the Generic Webpage Phishing Model.

Rank	Feature	Mean Importance	Standard Deviation
1	page_rank	0.008472	0.001702
2	google_index	0.001882	0.002731
3	phish_hints	0.001377	0.001676
4	domain_age	0.001051	0.001375
5	nb_hyperlinks	0.001024	0.002197
6	nb_www	0.000347	0.000693
7	web_traffic	0.000148	0.001415
8	nb_dollar	0.000000	0.000000
9	nb_space	0.000000	0.000000
10	nb_semicolumn	0.000000	0.000000
11	nb_comma	0.000000	0.000000
12	nb_star	0.000000	0.000000
13	nb_percent	0.000000	0.000000
14	nb_colon	0.000000	0.000000
15	nb_tilde	0.000000	0.000000
16	nb_com	0.000000	0.000000
17	nb_dslash	0.000000	0.000000
18	nb_at	0.000000	0.000000
19	ip	0.000000	0.000000
20	nb_or	0.000000	0.000000

Table 24 shows permutation importance results for the finance-adapted (FA) webpage phishing model. In contrast to the generic model, the FA model maintained page_rank as the top feature (mean importance = 0.013945) but also assigned substantial importance to a wider and more phishing-specific set of predictors, including web_traffic, phish_hints, nb_hyperlinks, links_in_tags, ratio_exterrors, ratio_extredirection, and external_favicon. This broader feature profile indicates that finance adaptation shifted the model beyond narrow reputation-based signals toward a richer combination of structural phishing indicators, particularly those involving hyperlink behaviour, external resource loading, and redirection patterns. These results provide partial support for H1 by showing that the FA model developed a more domain-relevant feature representation, even though overall FT subset performance did not improve for all classifiers. The findings also support H2, as the shift toward phishing-specific structural features is consistent with the argument that webpage modality contains more generalisable than finance-explicit cues, and finance adaptation reweights rather than transforms the feature space.

 doi:10.20944/preprints202605.0920.v1

Table 24. Permutation Importance Results for the FA Webpage Phishing Model.

Rank	Feature	Mean Importance	Standard Deviation
1	page_rank	0.013945	0.001996
2	web_traffic	0.005072	0.001518
3	phish_hints	0.004579	0.002008
4	nb_hyperlinks	0.003381	0.001310
5	links_in_tags	0.001692	0.000000
6	domain_age	0.001547	0.000922
7	ratio_exterrors	0.001523	0.000508
8	ratio_extredirection	0.001523	0.000508
9	google_index	0.001360	0.002408
10	ratio_extmedia	0.001354	0.000677
11	length_hostname	0.001337	0.001023
12	external_favicon	0.001184	0.000775
13	safe_anchor	0.001184	0.001082
14	char_repeat	0.001015	0.000829
15	ratio_intmedia	0.001015	0.000829
16	ratio_digits_url	0.001006	0.000840
17	prefix_suffix	0.000846	0.000846
18	avg_words_raw	0.000846	0.000846
19	longest_word_host	0.000846	0.000846
20	domain_registration_length	0.000846	0.000846

5. Implications, Limitations, and Conclusions

5.1. Theoretical Implications

This study makes several important contributions to the phishing detection and cybersecurity analytics literature by introducing a domain-aware perspective to phishing classification.

First, the findings challenge the conventional assumption that phishing detection should be treated as a fully generic binary classification problem. The results show that finance-themed phishing exhibits distinct predictive patterns, and that models trained with finance-specific emphasis outperform generic models within this domain. This supports the theoretical argument that phishing is not homogeneous but rather composed of domain-specific subtypes with unique semantic and behavioural characteristics.

Second, the study contributes to the literature on domain adaptation and specialised learning by providing empirical evidence of a specialisation–generalisation trade-off. While finance-adapted models improve performance on finance-themed instances, they do not consistently outperform generic models on non-finance-themed data. This finding extends existing machine learning theory by demonstrating that domain-specific optimisation improves in-domain sensitivity but may reduce broader generalisation, particularly in heterogeneous cybersecurity environments.

Third, the study advances multi-modal cybersecurity research by showing that the benefits of domain adaptation are modality-dependent. The stronger performance gains observed in email classification relative to webpage classification highlight that domain-specific signals are more explicitly encoded in textual data than in structured webpage features. This provides new insight

 doi:10.20944/preprints202605.0920.v1

into how feature representation interacts with domain adaptation, contributing to the broader literature on multi-modal machine learning and cybersecurity analytics.

Finally, the integration of SHAP-based explainability provides theoretical insight into how models learn domain-specific signals. The results demonstrate that FA models shift feature reliance toward more relevant predictors, supporting the argument that domain adaptation not only improves performance but also changes the underlying decision logic of models.

5.2. Practical Implications

The findings of this study have direct implications for financial institutions, fintech platforms, and cybersecurity practitioners.

First, the results suggest that organisations operating in financially sensitive environments, such as banks, payment platforms, and digital finance providers should consider deploying domain-adapted phishing detection systems rather than relying solely on generic models. Since finance-adapted models show superior performance in finance-themed contexts, their adoption can reduce false negatives, which are particularly costly in financial fraud scenarios.

Second, the study highlights the importance of modality-specific model design. Email-based phishing detection systems benefit more from finance adaptation due to the presence of rich semantic cues, whereas webpage-based systems rely more on general structural indicators. This implies that organisations should adopt a hybrid or layered defence strategy, where:

- Email filters use domain-adapted NLP models
- Webpage detection systems rely on robust structural and behavioural features

Third, the identified specialisation–generalisation trade-off has important operational implications. While finance-adapted models improve detection in finance-related attacks, they may underperform in broader phishing scenarios. Therefore, practitioners should avoid a "one-size-fits-all" approach and instead consider:

- Ensemble systems combining generic and domain-adapted models
- Context-aware model selection, depending on the threat environment

Fourth, the explainability analysis provides actionable insights for cybersecurity analysts. By identifying the most influential features (e.g., page rank, hyperlinks, financial keywords), SHAP-based interpretation helps:

- Improve threat intelligence
- Enhance manual investigation processes
- Support model validation and trust

5.3. Policy and Regulatory Implications

This study also has implications for regulators and policymakers in financial cybersecurity.

First, the results emphasise that finance-themed phishing represents a high-risk category of cyber threats, requiring targeted detection strategies. Regulators may consider encouraging or mandating domain-specific cybersecurity frameworks for financial institutions, rather than relying on general-purpose phishing detection systems.

Second, the findings support the need for industry-specific cybersecurity standards, particularly in fintech ecosystems where phishing attacks often target:

- Online banking platforms
- Payment systems
- Digital wallets

In this context, regulatory bodies could promote:

- Domain-aware AI models
- XAI requirements for fraud detection systems

 doi:10.20944/preprints202605.0920.v1

Third, the use of SHAP-based explainability aligns with growing regulatory emphasis on AI transparency and accountability. The ability to interpret model decisions is critical for:

- Auditing automated fraud detection systems
- Ensuring compliance with AI governance frameworks
- Building trust in automated decision-making

5.4. Methodological Implications

From a methodological perspective, this study demonstrates the importance of:

- Explicit subdomain labelling (FT vs NFT) in cybersecurity datasets
- Evaluating models across multiple subsets rather than aggregate performance only
- Incorporating explainability techniques to understand model behaviour

These practices provide a more nuanced and realistic evaluation framework, which can be applied to other domains of cybersecurity and fraud detection.

5.5. Limitations

Despite providing important insights into finance-themed phishing detection, this study has several limitations that should be acknowledged.

First, the study relies on secondary datasets compiled from publicly available sources, particularly for the email data. While the dataset is large and diverse, it may not fully capture the latest real-world phishing strategies, especially those evolving in response to modern fintech systems, AI-generated phishing (e.g., LLM-based attacks), and region-specific financial fraud patterns. As a result, the external validity of the findings may be limited in rapidly changing threat environments.

Second, the finance-theme labelling approach is rule-based, relying on predefined keyword dictionaries and weighted scoring. Although this method is transparent and replicable, it may introduce:

- False positives (generic messages containing financial terms)
- False negatives (finance-related phishing without explicit keywords)

This means that the FT classification may not perfectly capture all finance-related phishing instances, potentially affecting model training and evaluation.

Third, there is a modality imbalance in finance representation. While the email dataset contains a meaningful proportion of FT instances, the webpage dataset shows no dataset-wide finance-themed observations. This limits the ability to fully test finance adaptation in the webpage context and may partially explain why finance-adapted gains are weaker in webpage models.

Fourth, the study focuses primarily on traditional machine learning models (Logistic Regression, Linear SVC, Random Forest). While these models are appropriate for interpretability and benchmarking, they do not capture more advanced architectures such as:

- Deep learning models (e.g., LSTM, CNN)
- Transformer-based models (e.g., BERT)

These models may capture richer semantic relationships, especially in phishing emails.

Fifth, the study uses a static train–test split (80/20 holdout design). While this is standard practice, it does not account for temporal dynamics, where phishing tactics evolve over time. As such, the results may not fully reflect model robustness under real-world deployment conditions.

Sixth, the explainability analysis is limited to webpage models using SHAP and permutation importance. While this is appropriate for structured data, interpretability for email models is based only on linear coefficients, which may not fully capture complex interactions in text data.

5.6. Future Research Directions

Building on these limitations, several avenues for future research are proposed.

 doi:10.20944/preprints202605.0920.v1

First, future studies can explore dynamic and time-aware phishing detection models by incorporating:

- Temporal splits
- Rolling-window evaluation
- Concept drift analysis

This would provide a more realistic understanding of how models perform as phishing tactics evolve.

Second, the finance-theme labelling process can be enhanced using data-driven or hybrid approaches, such as:

- Supervised classification for theme detection
- Large language models (LLMs) for semantic labelling

This would improve the accuracy and flexibility of domain classification beyond keyword-based methods.

Third, future research can incorporate advanced deep learning and transformer-based models, particularly for email phishing detection. Models such as BERT or domain-specific language models may:

- Capture deeper contextual meaning
- Improve detection of subtle phishing cues
- Enhance performance in finance-themed classification tasks

Fourth, there is a need to develop richer finance-themed webpage datasets. Since the current webpage dataset lacks explicit finance-themed instances, future work should:

- Collect real-world finance-specific phishing websites
- Label webpage datasets using both content and structural signals
- Enable a more balanced cross-modality comparison

Fifth, future studies can explore multi-modal phishing detection frameworks, integrating:

- Email text
- Webpage structure
- User behaviour signals

This would reflect the real-world phishing process more accurately and may improve detection performance across attack stages.

Sixth, further research can investigate ensemble and hybrid model strategies, combining:

- Generic models (broad detection)
- Domain-adapted models (targeted detection)

Such approaches may help mitigate the specialisation–generalisation trade-off identified in this study.

Seventh, future work can extend explainability analysis by applying:

- SHAP to deep learning models
- Attention-based interpretation methods
- Cross-model explainability comparisons

This would provide deeper insights into how different model types learn phishing patterns.

Overall, while this study provides important evidence on the role of domain adaptation in phishing detection, future research should move toward dynamic, multi-modal, and AI-driven approaches to better capture the evolving and context-specific nature of cyber fraud.

5.7. Conclusion

This study examined whether finance-adapted phishing detection models provide measurable advantages over generic models in identifying finance-themed phishing attacks, and whether such advantages vary across email and webpage modalities. By explicitly distinguishing finance-themed

 doi:10.20944/preprints202605.0920.v1

from non-finance-themed instances and comparing generic and finance-adapted models across multiple classifiers and datasets, the study introduced a domain-aware framework for phishing detection.

The findings provide clear and nuanced evidence. First, consistent with H1, finance-adapted models demonstrated superior performance in the finance-themed email setting, where domain-specific semantic cues are strongly present. In particular, finance-adapted Linear SVC achieved the highest performance across key metrics, and statistical validation confirmed that these gains were both robust and significant. However, in the webpage setting, support for H1 was partial, with finance-adapted models outperforming generic models only for the Random Forest classifier, indicating that the benefits of domain adaptation are model- and modality-dependent.

Second, the results strongly support H2, showing that the effectiveness of finance adaptation is substantially stronger in email classification than in webpage classification. This reflects a fundamental difference in data representation: email data contains rich linguistic and contextual finance-related signals, whereas webpage data relies more heavily on general structural and technical indicators that are not uniquely tied to financial contexts. The SHAP and permutation importance analyses further reinforce this interpretation by showing that webpage models depend primarily on broad reputation and structural features, limiting the scope for finance-specific learning.

Third, the study provides strong support for H3, highlighting a clear specialisation–generalisation trade-off. While finance-adapted models improve detection performance within the finance-themed domain, they do not consistently outperform generic models on non-finance-themed instances and may even underperform in broader settings. This demonstrates that domain adaptation enhances in-domain sensitivity but may reduce out-of-domain generalisation, a finding that is both theoretically and practically significant.

Overall, the study makes three key contributions. It demonstrates that phishing detection should not be treated as a purely generic task, but rather as a context-dependent classification problem where domain-specific cues matter. It shows that the benefits of domain adaptation depend critically on data modality and feature representation. Finally, it highlights the importance of balancing specialised detection performance with broader generalisation capability, particularly in heterogeneous cybersecurity environments.

From a practical perspective, the findings suggest that financial institutions and fintech platforms can improve phishing detection by incorporating domain-adapted models, especially for email-based threats. However, such models should be deployed alongside generic detection systems to maintain coverage across diverse phishing scenarios. From a methodological perspective, the study underscores the value of combining domain-aware labelling, multi-modal analysis, and XAI techniques to achieve a more comprehensive understanding of model behaviour.

In conclusion, this study provides evidence that finance-themed phishing detection benefits from domain-aware modelling, but that such benefits are context-specific rather than universal. As phishing attacks continue to evolve in complexity and scope, future research and practice should move toward adaptive, multi-modal, and explainable detection frameworks that can effectively balance precision within targeted domains with robustness across broader threat environments.

Author Contributions: Conceptualization, Istiaque Bhuiyan and Tanvir Bhuiyan; Methodology, Istiaque Bhuiyan and Tanvir Bhuiyan; Software, Istiaque Bhuiyan; Validation, Istiaque Bhuiyan and Tanvir Bhuiyan; Formal analysis, Istiaque Bhuiyan and Tanvir Bhuiyan; Investigation, Istiaque Bhuiyan; Writing – original draft, Istiaque Bhuiyan and Tanvir Bhuiyan; Writing – review & editing, Tanvir Bhuiyan; Supervision, Tanvir Bhuiyan. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

 doi:10.20944/preprints202605.0920.v1

References

Alawida, M., Omolara, A. E., Abiodun, O. I., & Al-Rajab, M. (2022). A deeper look into cybersecurity issues in the wake of Covid-19: A survey. *Journal of King Saud University -Computer and Information Sciences, 34*(10), 8176-8206.

AlBenJasim, S., Takruri, H., Al-Zaidi, R., & Dargahi, T. (2024). Development of cybersecurity framework for FinTech innovations: Bahrain as a case study. *International Cybersecurity Law Review, 5*, 501-532.

Almutairi, A., Kang, B., & Alhashimy, N. (2025). Business email compromise: A systematic review of understanding, detection, and challenges. *Computers & Security*, *158*, 104630.

Al-Subaiey, A., Al-Thani, M., Alam, N. A., Antora, K. F., Khandakar, A., & Zaman, S. A. U. (2024). Novel interpretable and robust web-based AI platform for phishing email detection. *Computers and Electrical Engineering*, *120*, 109625.

Asiri, S., Xiao, Y., Alzahrani, S., & Li, T. (2024). Phishing RTDS: A real-time detection system for phishing attacks using a Deep Learning model. *Computers & Security*, *141*, 103843.

Baltuttis, D., & Teubner, T. (2024). Effects of visual risk indicators on phishing detection behavior: An eye-tracking experiment. *Computers & Security*, *144*, 103940.

Bustio-Martínez, L., Herrera-Semenets, V., García-Mendoza, J.L., Álvarez-Carmona, M.Á., González-Ordiano, J.Á., Zúñiga-Morales, L., Quiróz-Ibarra, J.E., Santander-Molina, P.A. & Van Den Berg, J. (2024). Uncovering phishing attacks using principles of persuasion analysis. *Journal of Network and Computer Applications*, *230*, 103964.

Bustio-Martínez, L., Herrera-Semenets, V., González-Ordiano, J. A., Pérez-Guadarrama, Y., Zúniga-Morales, L. N., Montoya-Godínez, D., & Van Den Berg, J. (2025). Enhanced phishing detection using multimodal data. *Knowledge-Based Systems*, *334*, 115105.

Butavicius, M., Taib, R., & Han, S. J. (2022). Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails. *Computers & Security*, *123*, 102937.

Căciulescu, A. R., Rughiniș, R., Țurcanu, D., & Radovici, A. (2024). Mapping cyber-financial risk profiles: Implications for european cybersecurity and financial literacy. *Risks*, *12*(12), 200.

Choudhary, T., Mhapankar, S., Bhddha, R., Kharuk, A., & Patil, R. (2023). A machine learning approach for phishing attack detection. *Journal of Artificial Intelligence and Technology, 3*(3), 108-113.

Doshi, J., Parmar, K., Sanghavi, R., & Shekokar, N. (2023). A comprehensive dual-layer architecture for phishing and spam email detection. *Computers & Security*, *133*, 103378.

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

Haider Rizvi, S. M., Imran, R., & Mahmood, A. (2025). Text classification using graph convolutional networks: A comprehensive survey. *ACM Computing Surveys*, *57*(8), 1-38.

Hannousse, A., & Yahiouche, S. (2021). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. *Engineering Applications of Artificial Intelligence*, *104*, 104347.

Jabir, R., Le, J., & Nguyen, C. (2025). Phishing attacks in the age of generative artificial intelligence: A systematic review of human factors. *AI*, *6*(8), 174.

Jafri, J. A., Amin, S. I. M., Rahman, A. A., & Nor, S. M. (2024). A systematic literature review of the role of trust and security on Fintech adoption in banking. *Heliyon*, *10*(1).

Javaheri, D., Fahmideh, M., Chizari, H., Lalbakhsh, P., & Hur, J. (2024). Cybersecurity threats in FinTech: A systematic review. *Expert Systems with Applications*, *241*, 122697.

Kavya, S., & Sumathi, D. (2024). Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection. *Artificial Intelligence Review*, *58*(2), 50.

Khadka, K., Ullah, A. B., & Martinez Marroquin, E. (2026). Unmasking persuasion in phishing: a content analysis of principles of persuasion in emails and subject lines. *Information & Computer Security*, *34*(1), 104-121.

Korkmaz, Y. (2026). Fraud E-mail detection using intelligent algorithms: Comparison of traditional approaches with deep learning techniques. *Information Processing & Management*, *63*(2), 104416.

Laxman, V., Ramesh, N., Prakash, S. K. J., & Aluvala, R. (2024). Emerging threats in digital payment and financial crime: A bibliometric review. *Journal of Digital Economy*, *3*, 205-222.

 doi:10.20944/preprints202605.0920.v1

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *13*(2), 1-41.

Loggen, J., Moneva, A., & Leukfeldt, R. (2024). A systematic narrative review of pathways into, desistance from, and risk factors of financial-economic cyber-enabled crime. *Computer Law & Security Review*, *52*, 105858.

Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S., & Porras, J. (2023). Mitigation strategies against phishing attacks: A systematic literature review. *Computers & Security*, *132*, 103387.

Ofusori, L., Bokaba, T., & Mhlongo, S. (2024). Artificial intelligence in cybersecurity: A comprehensive review and future direction. *Applied Artificial Intelligence*, *38*(1), 2439609.

Opara, C., Chen, Y., & Wei, B. (2024). Look before you leap: Detecting phishing web pages by exploiting raw URL and HTML characteristics. *Expert Systems with Applications*, *236*, 121183.

Rashid, F., Doyle, B., Han, S. C., & Seneviratne, S. (2024). Phishing URL detection generalisation using unsupervised domain adaptation. *Computer Networks*, *245*, 110398.

Ribeiro, L., Guedes, I. S., & Cardoso, C. S. (2024). Which factors predict susceptibility to phishing? An empirical study. *Computers & Security*, *136*, 103558.

Ribeiro, L., Guedes, I. S., & Cardoso, C. S. (2026). Eyes on phishing emails: an eye-tracking study: Liliana Ribeiro. *Journal of Experimental Criminology*, *22*(1), 189-211.

Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University-Computer and Information Sciences*, *35*(2), 590-611.

Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, *57*(12), 324.

Shafin, S. S. (2025). An explainable feature selection framework for web phishing detection with machine learning. *Data Science and Management*, *8*(2), 127-136.

Sturman, D., Bell, E. A., Auton, J. C., Breakey, G. R., & Wiggins, M. W. (2024). The roles of phishing knowledge, cue utilization, and decision styles in phishing email detection. *Applied Ergonomics*, *119*, 104309.

Tanveer, M. H., Fatima, Z., Zardari, S., & Guerra-Zubiaga, D. (2023). An in-depth analysis of domain adaptation in computer and robotic vision. *Applied Sciences*, *13*(23), 12823.

Tjingaete, M., & Juremi, J. (2023). Anti-phishing email detection framework for new-age phishing attacks. *Journal of Applied Technology and Innovation*, *7*(1).

Zhang, Y., Qian, L., Zhang, Q., Li, P., & Liu, G. (2022, November). A structure-aware method for cross-domain text classification. In *Pacific Rim International Conference on Artificial Intelligence* (pp. 283-296). Cham: Springer Nature Switzerland.