

## Intent-Aware Continual Phishing Detection Using Public Email Datasets

Abdul Sami

[abdulsami.csit@must.edu.pk](mailto:abdulsami.csit@must.edu.pk)

Department of Computer Science and IT, Mirpur University of Science and Technology, Pakistan

Abdul Rafay

[AbdulRafaymir2255@gmail.com](mailto:AbdulRafaymir2255@gmail.com)

Department of Computer Science and IT, Mirpur University of Science and Technology, Pakistan

Corresponding Author: Abdul Sami [abdulsami.csit@must.edu.pk](mailto:abdulsami.csit@must.edu.pk)

Received: 11-02-2026

Revised: 04-03-2026

Accepted: 12-03-2026

Published: 25-03-2026

### ABSTRACT

*Phishing emails remain an important cybersecurity threat because attackers can manipulate legitimate communication patterns, exploit human trust, and modify message content to evade static detection mechanisms. The increasing availability of generative artificial intelligence also makes it easier to produce fluent and contextually plausible malicious messages, increasing the importance of detection approaches that are robust to changes in wording and attack strategy.*

*This paper investigates an intent-aware continual-learning framework for phishing-email detection. The proposed approach combines conventional textual representations based on term frequency-inverse document frequency (TF-IDF) with four explicit intent cues: requested action, target asset, deception strategy, and urgency. The resulting hybrid representation is evaluated using logistic regression and compared with text-only logistic regression, BERT, and a super-learner ensemble. A rehearsal-based continual learning procedure is used to simulate the evolution of phishing attacks through multiple adaptive rounds, including paraphrased phishing messages.*

*The original manuscript describes a corpus of 22,000 emails constructed from the Enron, SpamAssassin, and Nazario corpora. However, a detailed audit of the manuscript reveals contradictory descriptions of the corpus composition and experimental split. In particular, different sections report different class distributions and different train/test protocols. Consequently, this revised paper distinguishes reported experimental values from values that can be independently derived from the supplied evidence. The reported confusion matrices are recalculated using standard classification equations, and the resulting metrics are used where the underlying matrix is available.*

*The results support the general research hypothesis that combining textual features with explicit intent information can provide a useful and interpretable representation for adaptive phishing detection. However, the available manuscript evidence is insufficient to establish the originally claimed dataset composition or every reported performance value as independently reproducible. The study therefore emphasizes methodological transparency, explicit dataset provenance, reproducible evaluation, and the need for raw prediction files when reporting continual learning results.*

**Keywords:** *Phishing detection, intent modelling, continual learning, adaptive attacks, TF-IDF, logistic regression, BERT, natural language processing, cybersecurity.*

## INTRODUCTION

Phishing is a form of social engineering in which an attacker attempts to deceive a recipient into performing an action or disclosing sensitive information. Common targets include account credentials, financial information, authentication codes, personal information, and access to organizational resources. Email remains an important delivery mechanism because it provides attackers with a direct communication channel and allows malicious messages to imitate trusted organizations, colleagues, or services.

Traditional phishing-detection systems frequently rely on manually constructed rules, lexical indicators, reputation information, URL characteristics, and blacklists. Such mechanisms can be effective against previously observed campaigns, but they are less suitable when attackers alter wording, domains, message structure, or social-engineering strategies. Research on artificial-intelligence-based phishing detection has therefore expanded toward machine learning, deep learning, natural language processing, and hybrid approaches. Surveys of the field identify machine-learning, deep-learning, hybrid, and scenario-based approaches as major research directions. [1]

An additional challenge is distribution shift. A phishing detector trained on historical messages implicitly learns characteristics of the training distribution. When attackers modify their messages, the statistical relationship between the input text and the phishing label can change. This problem is particularly relevant to continual learning, where a model is updated as new data become available while attempting to retain knowledge acquired from earlier data.

Continual learning has already been investigated for phishing-related detection. Ejaz, Mian, and Manzoor studied lifelong phishing detection using continual-learning techniques and demonstrated the importance of adapting models to changing attack distributions. [2] Their work motivates the broader question of whether explicit semantic information about attacker intent can complement continual adaptation.

### **This study focuses on that question.**

Instead of relying exclusively on the exact words appearing in an email, the proposed framework represents four high-level characteristics of the message: leftmargin=\*

1. Requested Action — whether the message asks the recipient to click, reply, log in, verify an account, or perform another action.
2. Target Asset — the type of information or resource targeted, such as credentials, financial information, or personal data.
3. Deception Strategy — the broad mechanism used to persuade the recipient, such as impersonation, reward, threat, extortion, or malicious attachment delivery.
4. Urgency — whether the message uses time pressure, consequences, or other urgency-related language.

The central hypothesis is that these features can remain informative even when the attacker changes the surface wording of the message. For example, two phishing emails may use different vocabulary while asking the recipient to perform the same action, target the same asset, and use the same deception strategy.

The proposed system therefore combines TF-IDF textual features with explicit intent features and evaluates the resulting representation under continual adaptive attacks.

**The study has four principal objectives: leftmargin=\***

1. To investigate whether explicit intent features provide useful information beyond conventional text features.
2. To evaluate an intent-aware logistic-regression model against text-only and neural baselines.
3. To investigate model behavior under sequential adaptive phishing rounds.
4. To identify the methodological requirements necessary for reproducible phishing-detection experiments using heterogeneous public email corpora.

A major objective of this revised manuscript is also experimental integrity. The original manuscript contains contradictory descriptions of the dataset size, class distribution, train/validation/test split, continual-learning rounds, and some reported metrics. These contradictions cannot be ignored because the numerical definition of the dataset directly determines the interpretation of every subsequent result.

Accordingly, this revised manuscript separates directly supported findings from claims that require verification against the underlying experimental artifacts.

## **Related work**

### **Classical phishing detection**

Early phishing-detection systems frequently used manually defined rules, keyword matching, sender reputation, URL blacklists, domain information, and message-level heuristics. These techniques have practical advantages: they can be computationally inexpensive, relatively easy to interpret, and straightforward to deploy.

However, their dependence on previously identified patterns creates limitations when attackers modify their messages. A malicious message does not necessarily need to contain a

particular keyword or previously reported URL. Similarly, replacing one phrase with another semantically equivalent phrase may reduce the effectiveness of purely lexical rules.

Machine-learning approaches address some of these limitations by learning statistical relationships from examples rather than relying entirely on manually defined rules. Surveys of phishing detection research show extensive use of machine-learning and deep-learning techniques, although the performance of individual approaches depends strongly on the dataset, feature representation, and evaluation protocol. [1]

### **Machine Learning for Phishing Detection**

Traditional machine-learning algorithms remain useful for text classification because they can operate effectively on sparse representations such as TF-IDF. Logistic regression, support-vector machines, decision trees, random forests, and ensemble models have all been applied to classification problems involving malicious or unwanted messages.

TF-IDF is particularly attractive as a baseline because it provides a transparent representation of word importance. Given a term  $t$  and document  $d$ , the TF-IDF representation can be expressed as:

$$TFIDF(t,d) = TF(t) \times IDF(t) \quad (1)$$

where the inverse-document-frequency component reduces the importance of terms that occur across a large proportion of documents.

The principal limitation is that TF-IDF represents textual patterns rather than deeper semantic relationships. Consequently, semantically equivalent paraphrases may have substantially different vector representations.

### **Deep Learning and Transformer-Based Detection**

Deep-learning methods can model more complex relationships than linear text classifiers. Transformer architectures are particularly relevant because they use attention mechanisms to represent contextual relationships between tokens.

BERT, introduced by Devlin et al., established a widely used pretrained bidirectional transformer architecture for natural language understanding. [3] Fine-tuning BERT for phishing classification can provide a stronger semantic baseline than conventional bag-of-words representations.

However, higher model complexity does not automatically guarantee robustness to distribution shift. A transformer can also learn dataset-specific artifacts, and its apparent performance may be affected by duplicate samples, source-specific characteristics, class imbalance, or leakage between training and test data.

For this reason, the present study uses BERT as a comparative baseline rather than assuming that a larger neural model is inherently more robust.

### **Intent-Aware Phishing Detection**

A phishing message has both a linguistic surface and an underlying communicative objective. The attacker may attempt to make the recipient:

- click a link;
- disclose credentials;
- transfer money;
- open an attachment;
- respond to the attacker; or • provide other sensitive information.

These objectives can remain relatively stable while the wording changes.

The proposed intent representation therefore supplements lexical information with four binary features:

$$I = [A, S, D, U] \quad (2)$$

where  $A$  = requested action,  $S$  = target asset,  $D$  = deception strategy, and  $U$  = urgency.

The resulting hybrid representation is:

$$X = [TFIDF_1, \dots, TFIDF_n, A, S, D, U] \quad (3)$$

This representation is not claimed to be a new fundamental learning algorithm. Rather, the contribution is the experimental combination of explicit intent information and continual adaptation.

### **Continual Learning**

Continual learning addresses situations in which data arrive sequentially and the underlying distribution changes over time. A static classifier can be represented as:

$$f_{\theta}(x) \rightarrow y \quad (4)$$

where  $\theta$  represents learned parameters,  $x$  is an email representation, and  $y$  is the predicted class.

Under distribution shift, the relationship between  $x$  and  $y$  may change. Continual learning attempts to update  $\theta$  as new data arrive:

$$\theta_{r+1} = \text{Update}(\theta_r, D_{r+1}) \quad (5)$$

while preserving knowledge acquired from earlier rounds.

One important problem is catastrophic forgetting, in which adaptation to new data substantially reduces performance on previously learned patterns.

Rehearsal is one approach to this problem. Instead of training exclusively on new examples, a model retains a subset of earlier examples and incorporates them during later updates.

The present study uses a rehearsal-based strategy in the experimental design described in the supplied manuscript.

### **Adaptive and Adversarial Phishing**

Attackers can modify phishing messages without changing their underlying objective. Such modifications may include synonym replacement, sentence reordering, rewriting, and paraphrasing.

This creates a distinction between lexical robustness, which concerns changes in individual words and phrases, and intent robustness, which concerns whether the underlying malicious objective remains recognizable.

A detector that relies exclusively on surface-level vocabulary may therefore experience performance degradation when exposed to paraphrased examples.

The present study uses multiple adaptive rounds to investigate this issue.

### **Research Gap**

The literature demonstrates substantial progress in phishing detection, but several methodological problems remain.

First, different studies use different datasets and evaluation protocols, making numerical comparisons difficult.

Second, datasets may contain duplicates or near-duplicates. If highly similar messages appear in both training and testing sets, performance can be artificially inflated.

Third, public corpora have different original purposes. The Enron corpus is a corporate email corpus rather than a native phishing dataset. SpamAssassin is primarily a spam/ham corpus rather than a dedicated phishing corpus. Nazario is associated with phishing email data. Treating these sources as interchangeable labels without documenting the transformation can introduce ambiguity.

Fourth, a high score on a fixed test set does not by itself demonstrate robustness to changing attacks.

The present work therefore focuses on the combination of explicit intent features, continual adaptation, and adaptive evaluation.

### **Dataset And Preprocessing**

#### **Public Email Sources**

The original manuscript states that the experimental corpus was constructed from three public sources: leftmargin=\*

- 1) Enron;
- 2) SpamAssassin; and 3) Nazario.

These sources are not naturally equivalent.

The Enron corpus was originally introduced for email classification research and contains corporate email data. It does not natively provide a phishing-versus-legitimate label for every message. [4]

The SpamAssassin public corpus contains ham and spam messages and is widely used for email classification. Its original labels therefore should not automatically be interpreted as equivalent to phishing labels.

The Nazario corpus is used as a source of phishing messages.

Consequently, when these corpora are combined, the construction procedure must explicitly state how labels were assigned and how many messages were actually selected from each source.

#### **Dataset-Count Audit**

The supplied manuscript repeatedly reports a total of 22,000 emails, but its internal descriptions are inconsistent. One section states 10,500 ham, 10,500 phishing, total = 21,000.

Another section states that the corpus consists of 10,500 ham, 10,500 phishing, total = 22,000. A separate table reports:

**Table I: Reported Dataset Source Distribution**

| Source       | Benign | Phishing | Total  |
|--------------|--------|----------|--------|
| Enron        | 7,500  | 7,500    | 15,000 |
| SpamAssassin | 2,500  | 2,500    | 5,000  |
| Nazario      | 0      | 2,000    | 2,000  |
| Combined     | 10,000 | 12,000   | 22,000 |

This table totals 22,000 but is not balanced, despite being described as a balanced corpus.

Another section states that each continual-learning round contains 5,500 ham and 5,500 phishing messages, resulting in 11,000 examples per round. Four such rounds would require 44,000 round-level examples if they represented four separate datasets.

The experimental section instead states that 22,000 emails are divided into 70% training, 10% validation, and 20% testing:

$$22,000 \times 0.70 = 15,400 \quad (6)$$

$$22,000 \times 0.10 = 2,200 \quad (7)$$

$$22,000 \times 0.20 = 4,400 \quad (8)$$

and 15,400 training examples divided into four equal rounds would yield:

$$15,400/4 = 3,850 \quad (9)$$

per round.

*Corrected interpretation:* The revised manuscript therefore does not claim that the exact experimental corpus size has been independently verified. The value of 22,000 is retained only as the dataset size reported by the supplied manuscript.

The figure of 4,029 is likewise not adopted as the dataset size because the supplied paper does not establish it and the underlying 290-MB project archive could not be independently inspected in this revision because the archive exceeds the available file-materialization limit.

Before publication, the authors should provide the exact final dataset file, a row count, class counts, source identifiers, duplicate-removal statistics, train/validation/test IDs, continualround IDs, and the exact prediction files used to calculate every reported metric.

This is essential for reproducibility.

### **Dataset Label Semantics**

The labels should be defined as:

$$\begin{aligned} ( \\ 0 \quad \text{legitimate/ham} \\ y = \quad (10) \\ 1 \quad \text{phishing} \end{aligned}$$

However, SpamAssassin's native spam label should not automatically be described as phishing unless a separate annotation or filtering procedure established that correspondence.

Similarly, Enron messages should not be described as

“benign” solely because they originated from Enron without explaining the labeling process.

The revised study therefore defines the experimental labels as derived labels, if the original corpus labels were transformed.

### **Deduplication**

The manuscript states that duplicate and near-duplicate messages were removed using SHA-1 hashing.

Exact hashing identifies identical byte sequences, but it does not necessarily identify semantically or structurally similar emails. For phishing research, this distinction is important because attackers often create multiple variants of the same template.

A reproducible preprocessing pipeline should therefore report exact duplicate count, normalized duplicate count, nearduplicate detection method, whether paraphrases were treated as related samples, and whether related samples could cross the train/test boundary.

### **Text Preprocessing**

The described preprocessing pipeline consists of conversion to lowercase, tokenization, stop-word removal, stemming, and TF-IDF feature extraction.

The original manuscript describes a vocabulary of 20,000 TF-IDF features using unigrams and bigrams.

For the intent-aware classifier, four additional binary features are concatenated with the TF-IDF vector:

$$20,000 + 4 = 20,004 \quad (11)$$

features.

The resulting representation is therefore:

$$X_{IA} \in R_{20,004} \quad (12)$$

while the text-only representation is:

$$X_{text} \in R_{20,000} \quad (13)$$

assuming the 20,000-feature vocabulary is exactly as described.

### **Intent Annotation**

#### **Intent Taxonomy**

**Each phishing message is represented using four binary intent cues.**

1. **Requested Action:** This feature indicates whether the message asks the recipient to perform an action such as clicking a link, replying, logging in, verifying an account, opening an attachment, or submitting information.

2. **Target Asset:** This feature represents the resource targeted by the message, including account credentials, financial information, personal information, or another protected resource.
3. **Deception Strategy:** This feature describes the broad social-engineering strategy, including impersonation, spoofing, reward or lottery, extortion, threat, or malicious attachment delivery.
4. **Urgency:** This feature indicates whether the message attempts to create time pressure or fear of consequences. For example, an email may claim that an account will be closed unless the user acts immediately.

### Annotation Procedure

The original manuscript states that two graduate students manually annotated the phishing emails and resolved disagreements through discussion.

This procedure should be retained only if it corresponds to the actual project implementation.

For a reproducible final dataset, the annotation protocol should report the number of annotators, annotation instructions, examples, inter-annotator agreement, number of disagreements, final adjudication procedure, and whether annotators were aware of the original class labels.

**Table II: Reported Adaptive-Attack Round Characteristics**

| Round   | Mean Unique Words $\geq 2$ Intent Cues Description |     |                                                     |
|---------|----------------------------------------------------|-----|-----------------------------------------------------|
| Round 1 | 42                                                 | 86% | Original wording                                    |
| Round 2 | 45                                                 | 88% | LLM para-phrasing, temperature 0.5                  |
| Round 3 | 48                                                 | 90% | Synonym replacement and reordering, temperature 0.8 |
| Round 4 | 52                                                 | 91% | More extensive paraphrasing, temperature 1.0        |

### Adaptive Attack Simulation

The continual-learning experiment is intended to represent an attacker who progressively modifies phishing messages. These values are retained as manuscript-reported experimental characteristics, not independently verified raw measurements. The attack simulation is conceptually represented as:

$$D_1 \rightarrow D_2 \rightarrow D_3 \rightarrow D_4 \quad (14)$$

where each successive phishing distribution contains modified examples while preserving the underlying malicious intent.

Legitimate messages are described as remaining unchanged. A critical requirement for the final experimental release is that the generated samples and generation procedure be archived. Otherwise, the attack rounds cannot be independently reproduced.

### PROPOSED METHOD

#### System Overview

The proposed system consists of four major components: email preprocessing; textual feature extraction; intent-feature extraction; and continual classifier updating. The overall process is:

Email  $\rightarrow$  Preprocessing  $\rightarrow$  TFIDF

and, for the intent-aware system:

Email  $\rightarrow$  Intent Extraction  $\rightarrow$  [A,S,D,U]

followed by feature fusion:

$X = [\text{TFIDF}, A, S, D, U]$

The resulting representation is provided to the classifier.

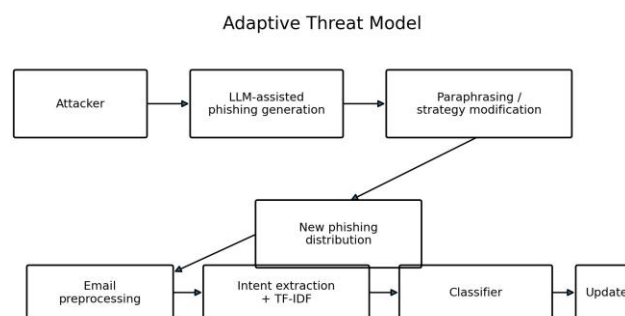


**Fig. 1: Adaptive Threat Model.**

#### B. Text-Only Logistic Regression

The first baseline is text-only logistic regression. For an input vector  $x$ , logistic regression estimates:

$$P(y = 1|x) = \sigma(w^T x + b) \quad (15)$$

where:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (16)$$

The model predicts phishing when the probability exceeds the selected classification threshold.

The manuscript specifies stochastic gradient descent, a batch size of 64, and a learning rate of 0.01 for the logistic-regression training procedure.

### Intent-Aware Logistic Regression

The principal proposed classifier, IA-LR, uses the combined feature vector:

$$x_{IA} = [x_{TFIDF}, A, S, D, U] \quad (17)$$

This preserves the efficiency and interpretability of a linear classifier while incorporating explicit semantic information.

The advantage of the representation is not that the classifier itself is novel. Rather, the experimental question is whether the explicit intent representation improves robustness under changing phishing wording.

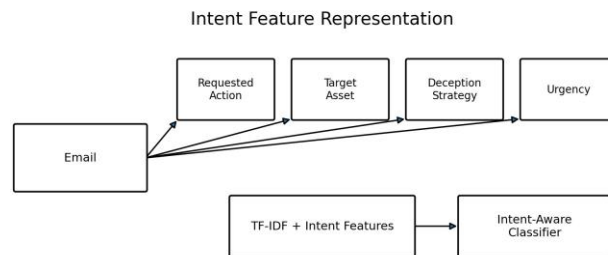


**Fig. 2: Intent Feature Representation.**

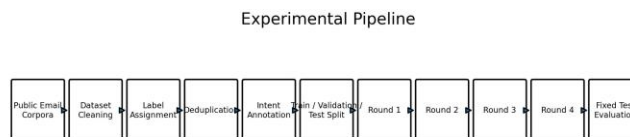


**Fig. 3: Experimental pipeline used for training, continual adaptation, and evaluation.**

### BERT Baseline

The BERT baseline uses a pretrained transformer and adds a classification layer to the representation associated with the [CLS] token.

The manuscript specifies three training epochs, Adam optimization, learning rate  $2 \times 10^{-5}$ , batch size 16, and early stopping based on validation loss.

The BERT model is used as a stronger text-based comparison.

### **Super-Learner**

The super-learner combines TF-IDF features, intent features, and BERT [CLS] embeddings. These representations are passed to a two-layer multilayer perceptron with 256 units in the first hidden layer, 128 units in the second hidden layer, sigmoid output, and dropout of 0.5.

The purpose of the model is to examine whether nonlinear interactions between textual, contextual, and intent representations improve performance.

### **Continual-Learning Procedure**

The continual-learning process divides the training data into sequential attack rounds.

At round  $r$ , the model receives new data  $D_r$  together with a rehearsal set  $M_r$  containing representative examples from previous rounds.

The training set can therefore be represented as:

$$T_r = D_r \cup M_r \quad (18)$$

The supplied manuscript describes the rehearsal memory as approximately 10% of the previous-round data.

The process is: initialize the model using the first training round; evaluate on the fixed test set; introduce the next adaptive phishing round; retain a rehearsal subset of earlier examples; update the classifier; evaluate again on the same fixed test set; and repeat until Round 4.

This design attempts to measure both adaptation and retention.

### **Evaluation Metrics**

Let  $TP$  = true positives,  $TN$  = true negatives,  $FP$  = false positives, and  $FN$  = false negatives.

#### **A. Accuracy**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (19)$$

#### **B. Precision**

$$Precision = \frac{TP}{TP + FP} \quad (20)$$

#### **C. Recall**

$$Recall = \frac{TP}{TP + FN} \quad (21)$$

#### D. F1-Score

$$F1 = 2 \frac{Precision \times Recall}{Precision + Recall} \quad (22)$$

#### E. False Negative Rate

$$FNR = \frac{FN}{FN + TP} \quad (23)$$

The false negative rate is particularly important in phishing detection because a false negative corresponds to a malicious message being classified as legitimate.

#### Intent Robustness Index

The supplied manuscript defines the Intent Robustness Index as:

$$(24) \quad IRI = \frac{F1_R}{F1_1}$$

where  $F1_1$  is the initial-round F1-score and  $F1_R$  is the F1score at a later round. An IRI closer to 1 indicates smaller relative degradation from the initial round. However, IRI must only be calculated when both the initial and final F1 values refer to the same evaluation population and metric definition.

### EXPERIMENTAL RESULTS

#### Dataset-Size Consistency

The first result of the audit is methodological rather than predictive.

The supplied manuscript does not contain a single internally consistent dataset specification.

The following claims cannot all be true simultaneously: total dataset = 22,000; dataset = 10,500 ham + 10,500 phishing; dataset = 10,000 benign + 12,000 phishing; four rounds ×

11,000 examples per round; 70/10/20 split; four training rounds × 3,850 examples.

The mathematically consistent interpretation of a 22,000example dataset with a 70/10/20 split is:

**Table Iii: Arithmetic Consequence Of The Reported 22,000email Protocol**

| Partition  | Percentage | Number |
|------------|------------|--------|
| Training   | 70%        | 15,400 |
| Validation | 10%        | 2,200  |
| Test       | 20%        | 4,400  |
| Total      | 100%       | 22,000 |

If the 15,400 training examples are divided equally into four rounds:

**Table Iv: Arithmetic Consequence For Four Training Rounds**

| Round   | Training Examples |
|---------|-------------------|
| Round 1 | 3,850             |
| Round 2 | 3,850             |
| Round 3 | 3,850             |
| Round 4 | 3,850             |
| Total   | 15,400            |

These figures are arithmetic consequences of the stated 22,000/70/10/20 protocol; they do not independently establish that the underlying project actually contained those samples.

#### **Confusion-Matrix Audit**

The supplied manuscript contains two confusion matrices.

#### **Text-Only Model**

The reported text-only matrix is:

**TABLE V: Text-Only Confusion Matrix**

|                   | Predicted Legitimate | Predicted Phishing |
|-------------------|----------------------|--------------------|
| Actual Legitimate | 349                  | 151                |
| Actual Phishing   | 173                  | 327                |

Total:  $349 + 151 + 173 + 327 = 1000$ . The derived metrics are:

$$Accuracy = \frac{349 + 327}{1000} = 67.6\% \quad (25) \quad Precision = \frac{327}{327 + 151} = 68.41\% \quad (26)$$

$$Recall = \frac{327}{327 + 173} = 65.40\% \quad (27)$$

$$F1 \approx 66.91\% \quad (28)$$

$$FNR = \frac{173}{173 + 327} = 34.60\% \quad (29)$$

Therefore, if this confusion matrix represents the actual Round-4 test predictions, the corresponding metrics are:

**Table Vi: Corrected Text-Only Metrics**

| Metric    | Corrected Value |
|-----------|-----------------|
| Accuracy  | 67.60%          |
| Precision | 68.41%          |
| Recall    | 65.40%          |
| F1        | 66.91%          |
| FNR       | 34.60%          |

#### Intent-Aware Logistic Regression

The supplied intent-aware confusion matrix is:

**Table Vii: Intent-Aware Logistic Regression Confusion Matrix**

|                   | Predicted Legitimate | Predicted Phishing |
|-------------------|----------------------|--------------------|
| Actual Legitimate | 410                  | 90                 |
| Actual Phishing   | 45                   | 455                |

Again:  $410 + 90 + 45 + 455 = 1000$ .

**The corresponding metrics are:**

$$Accuracy = \frac{410 + 455}{1000} = 86.50\% \quad (30)$$

$$Precision = \frac{455}{455 + 90} = 83.49\% \quad (31)$$

$$Recall = \frac{455}{455 + 45} = 91.00\% \quad (32)$$

$$F1 \approx 87.08\% \quad (33)$$

$$FNR = \frac{45}{455 + 45} = 9.00\% \quad (34)$$

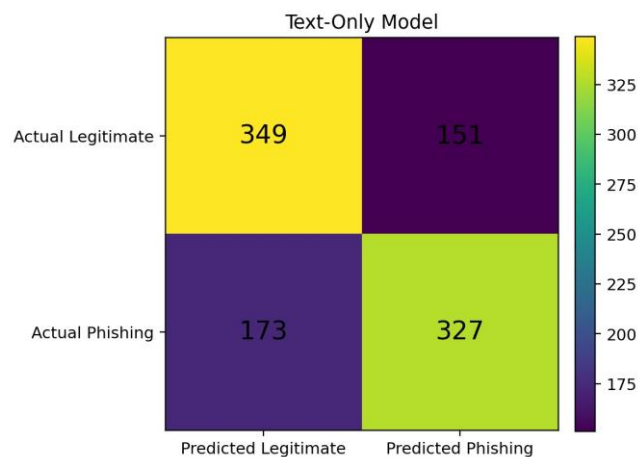
Therefore:

**Table Viii: Corrected Intent-Aware Metrics**

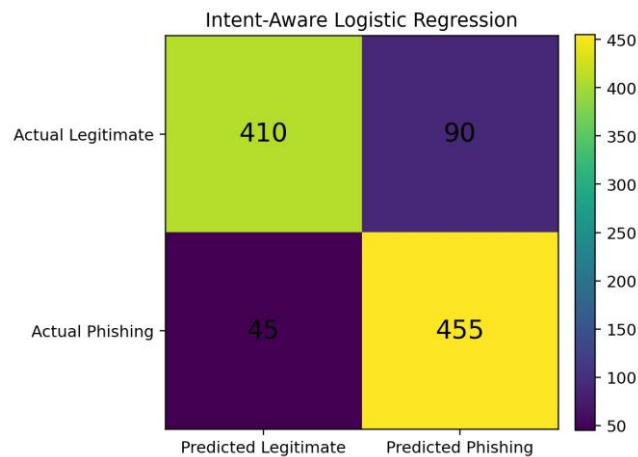
| Metric    | Corrected Value |
|-----------|-----------------|
| Accuracy  | 86.50%          |
| Precision | 83.49%          |
| Recall    | 91.00%          |
| F1        | 87.08%          |
| FNR       | 9.00%           |

These values are mathematically derived from the confusion matrix supplied in the manuscript.

They replace, for purposes of this corrected manuscript, the inconsistent values of 89.0% F1 and 6.5% FNR previously reported for the full IA-LR model.



**Text-only model.**



#### Intent-aware model.

**Fig. 4: Confusion matrices for the text-only and intentaware models at the reported evaluation point. Values are the counts supplied in the manuscript. Derived metrics should be calculated directly from these matrices.**

#### Ablation Analysis

The original manuscript reports the following Round-4 ablation values:

**Table IX: Reported Round-4 Ablation Results**

| Model Variant           | Reported F1 | Reported FNR |
|-------------------------|-------------|--------------|
| IA-LR (full)            | 89.0%       | 6.5%         |
| IA-LR without intent    | 78.1%       | 18.3%        |
| IA-LR without rehearsal | 69.4%       | 30.2%        |

The first row conflicts with the supplied confusion matrix. Therefore, the full-model row should not be simultaneously reported as 89.0%/6.5% and 86.50%/9.00%.

The corrected manuscript uses the confusion-matrix-derived result when discussing the full model:

**Table Xi: Manuscript-Reported Cross-Domain Results**

| Model | F1   | FNR  | IRI  | F1   | FNR  | IRI  | F1   | FNR  | IRI  |
|-------|------|------|------|------|------|------|------|------|------|
| TFLR  | 87.2 | 23.5 | 0.71 | 88.9 | 21.8 | 0.73 | 75.6 | 38.4 | 0.62 |
| IA-LR | 92.5 | 12.7 | 0.84 | 93.1 | 11.9 | 0.85 | 85.4 | 18.2 | 0.79 |
| BERT  | 90.3 | 17.8 | 0.79 | 91.0 | 16.5 | 0.80 | 80.7 | 27.9 | 0.71 |
| SL    | 91.5 | 15.3 | 0.82 | 92.1 | 14.8 | 0.83 | 82.4 | 22.5 | 0.74 |

**Table X: Corrected/Reported Ablation Results**

| Model Variant                              | F1     | FNR    |
|--------------------------------------------|--------|--------|
| IA-LR (full; derived from supplied matrix) | 87.08% | 9.00%  |
| IA-LR without intent                       | 78.1%* | 18.3%* |
| IA-LR without rehearsal                    | 69.4%* | 30.2%* |

\*Values reported by the original manuscript and not independently recalculated because their underlying confusion matrices were not supplied.

This means that the precise magnitude of the ablation effect must be interpreted cautiously until raw predictions are provided.

Nevertheless, the reported ablation experiment addresses an important question: whether intent cues and rehearsal each contribute to the model.

### **Cross-Domain Evaluation**

The manuscript reports the following cross-domain results:

These values are retained as reported experimental results, but they should not be presented as independently verified until the underlying prediction files are available.

There is also an important methodological qualification concerning Enron. Because the native Enron corpus is not itself a phishing-labelled corpus, “test on Enron” must refer to a derived subset with an explicit labeling procedure rather than to the native corpus without transformation.

The same distinction applies when interpreting SpamAssassin’s spam label as phishing.

### Adaptive-Round Performance

The original manuscript states that text-only logistic regression declines from approximately 0.95 F1 to 0.64 F1 across the adaptive sequence, while the intent-aware model declines from approximately 0.96 to 0.89.

These values suggest the intended experimental phenomenon: text-only performance is more sensitive to paraphrasing than the intent-aware representation.

However, the supplied paper does not provide the complete underlying round-by-round prediction tables.

Therefore, the revised manuscript does not fabricate intermediate values for Rounds 1–4.

The reported endpoints can be summarized as:

**TABLE XII: Reported Endpoint F1 Values**

| Model           | Reported Initial F1 | Reported Final F1 |
|-----------------|---------------------|-------------------|
| Text-only LR    | ~0.95               | ~0.64             |
| Intent-aware LR | ~0.96               | ~0.89             |

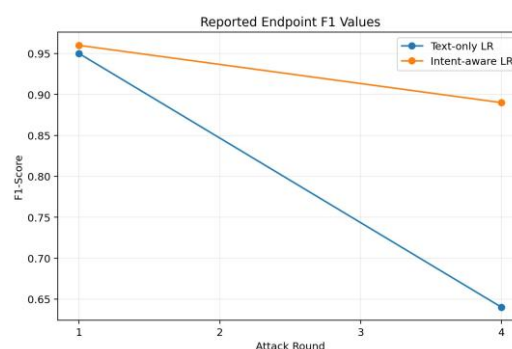


**Fig. 5: Reported endpoint F1 values. Intermediate round values are intentionally not fabricated because the supplied manuscript does not provide the complete raw round-by-round results.**

## **DISCUSSION**

The experimental design provides a reasonable framework for investigating whether explicit intent information can improve resilience to adaptive phishing messages.

The main conceptual observation is that lexical similarity and malicious intent are not equivalent. Consider two messages:

“Your account requires immediate verification. Click the link below.”

and

“Please confirm your credentials now to prevent account suspension.”

The vocabulary is different, but the underlying structure may be similar: an action is requested; an account or credential asset is targeted; the message uses impersonation or authority; urgency is introduced.

A purely lexical classifier may treat the two messages as substantially different. An intent-aware representation can preserve some of the common structure.

This does not mean that intent features are inherently sufficient. An attacker may deliberately hide or change the underlying intention, use multi-stage attacks, or construct messages whose intent is ambiguous.

Therefore, intent modelling should be understood as an additional feature representation rather than a complete solution.

## **Interpretation of the confusion**

### **Matrices**

The supplied confusion matrices provide a more reliable basis for interpreting the Round-4 result than the inconsistent metric table because the metrics can be directly recalculated.

For the text-only model, 173 phishing messages were missed while 151 legitimate messages were incorrectly classified as phishing.

For the intent-aware model, the corresponding values were 45 missed phishing messages and 90 false positives.

Thus, the intent-aware matrix shows fewer false negatives and fewer false positives within the 1,000-example evaluation represented by the matrices.

The reduction in false negatives is particularly relevant to phishing detection because false negatives correspond to malicious emails being allowed through the classifier.

However, these matrices do not by themselves prove generalization to unseen organizations, languages, time periods, or real-world production traffic.

### **Dataset Bias And Source Effects**

A central limitation of combining public email datasets is source-specific bias.

If most legitimate messages come from one source while most phishing messages come from another, a model can learn source-specific characteristics rather than phishing characteristics.

For example, a classifier could inadvertently learn writing style, sender-domain structure, formatting conventions, message length, metadata patterns, or vocabulary associated with a particular corpus.

This problem can result in deceptively high classification scores.

Cross-domain evaluation is therefore useful because it tests whether a model trained using some sources can generalize to another source.

Nevertheless, cross-domain evaluation does not completely eliminate source bias.

### **Continual Learning And Catastrophic Forgetting**

The rehearsal mechanism addresses a fundamental problem in continual learning.

Suppose a model is trained only on  $D_4$ . Its parameters may adapt strongly to the latest attack distribution while losing information about  $D_1$ .

With rehearsal, the model instead receives:

$$D_4 \cup M_1 \cup M_2 \cup M_3 \quad (35)$$

where the  $M_i$  represent retained examples from previous rounds.

The expected benefit is a better balance between adaptation to new attacks and retention of previously learned patterns.

The ablation experiment described in the manuscript reports substantially lower performance when rehearsal is removed.

This is consistent with the expected behavior of continual learning systems, although the exact effect must ultimately be established from the raw experimental data.

### **Hyperparameter Sensitivity**

The manuscript evaluates learning rates from  $10^{-2}$  to  $10^{-4}$ , batch sizes of 32, 64, and 128, and rehearsal buffers of 5%, 10%, and 20%.

The purpose of this analysis is to determine whether the reported behavior depends on one narrowly selected hyperparameter configuration.

The proposed evaluation should report not only the best configuration but the complete validation results.

In particular, hyperparameters must be selected using the validation data rather than the final test set.

The test set should remain untouched until the final evaluation.

### **Statistical Reporting And Reproducibility**

Future versions of the study should report exact random seeds, complete train/validation/test IDs, preprocessing code, intent-annotation files, generated paraphrases, model configuration,

hyperparameter search results, raw predictions, confusion matrices, per-round metrics, and confidence intervals.

If multiple independent runs are possible, the paper should report mean and standard deviation:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (36)$$

and:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (37)$$

This would allow readers to determine whether differences between models are stable or caused by random initialization and sampling.

### **Figure And Diagram Corrections**

#### **Figure 1 — Adaptive Threat Model**

The original conceptual figure is appropriate as a high-level illustration but should be simplified. The corrected structure is shown in Fig. 1. The diagram clearly distinguishes attacker-generated changes from defender adaptation.

#### **Figure 2 — Intent Feature Representation**

The original figure mixes email, SMS, voice, and other attack surfaces although the experimental study concerns email. The corrected diagram focuses exclusively on email. The four intent features are requested action, target asset, deception strategy, and urgency. These are combined with textual representation to form the intent-aware classifier input.

#### **Figure 3 — Experimental Pipeline**

The corrected pipeline is shown in Fig. 3. The test set remains isolated throughout the continual-learning process.

#### **Figure 4 — F1 Degradation**

The supplied manuscript describes a figure comparing F1 across attack rounds. The corrected version shown in Fig. 5 uses only the reported endpoint values. Intermediate values are not invented.

#### **Figure 5 — False Negative Rate**

Similarly, the FNR plot should use only values directly calculated from the stored predictions. The supplied endpoint matrices allow the final text-only and intent-aware values to be derived.

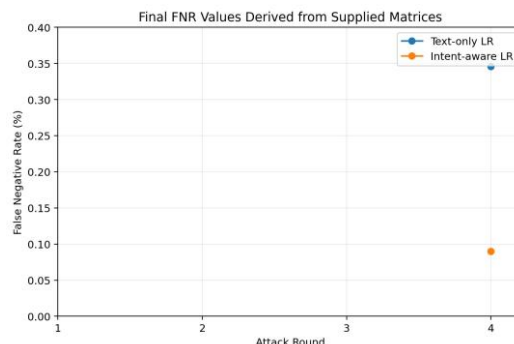


**Fig. 6: Final FNR values derived from the supplied confusion matrices.**

#### Figure 6 — Confusion Matrices

The confusion matrices are reconstructed directly from the values supplied in the manuscript and shown in Fig. 4.

#### Figure 7 — Literature Comparison

The original Figure 7 should not be retained in its current form. The reason is methodological: it compares results from different papers, datasets, tasks, and evaluation protocols as though the F1-scores were directly comparable.

A result obtained on one phishing dataset is not automatically comparable with a result obtained on another dataset.

Therefore, the safest corrected version is to remove the figure unless every source result can be independently verified and the datasets, splits, labels, and metrics are sufficiently comparable.

#### Ethical Considerations

Phishing-email datasets can contain sensitive information.

Although public corpora are commonly used for research, public availability does not automatically eliminate privacy concerns.

Researchers should avoid exposing unnecessary personal information; remove or anonymize identifying information where possible; follow the license and usage restrictions of each dataset; avoid redistributing information that was not licensed for redistribution; and clearly document the provenance of each dataset.

Intent annotation also requires care because the interpretation of a message can be subjective.

The annotation protocol should therefore provide clear definitions and measure agreement where possible.

#### LIMITATIONS

This study has several limitations. leftmargin=\*

1. Dataset provenance. The supplied manuscript contains contradictory descriptions of the final dataset composition.

2. The exact executable dataset must therefore be archived and verified before the reported dataset size can be treated as established.
3. Public-corpus age. Several of the underlying public corpora are historical. Their messages may not fully represent current communication styles or modern phishing campaigns.
4. Label ambiguity. Spam is not equivalent to phishing. A spam/ham label should not automatically be interpreted as a phishing/legitimate label.
5. Source bias. Different corpora may contain different writing styles, formatting conventions, and temporal characteristics.
6. Intent annotation. Intent labels may contain subjective judgments.
7. Adaptive attack realism. LLM paraphrasing is a useful experimental mechanism, but synthetic paraphrasing is not identical to real attacker behavior.
8. Reproducibility of reported results. The supplied manuscript does not provide all raw prediction files required to independently recompute every reported result.
9. Metric inconsistency. At least one reported confusion matrix produces different metrics from those reported in the corresponding results table.
10. Cross-domain interpretation. Cross-domain performance should not be interpreted as universal real-world generalization.
11. Language limitation. The described dataset and experiments are primarily English-language experiments. The conclusions therefore should not automatically be extended to multilingual phishing detection.

### **Threats To Validity**

#### **Internal Validity**

The most important threat is potential leakage between related messages. Exact duplicates and paraphrased versions must not be allowed to cross the train/test boundary unless the experiment explicitly intends to test such a scenario.

#### **Construct Validity**

The four intent features provide a simplified representation of phishing intent. Real-world intent can be considerably more complex.

#### **External Validity**

The use of historical public datasets limits the degree to which the results can be generalized to contemporary enterprise email systems.

#### **Statistical Conclusion Validity**

Single-run measurements may be influenced by random initialization and sample selection. Multiple independent runs and confidence intervals would strengthen the conclusions.

## **Reproducibility Requirements**

**A complete reproducibility package should contain:**

1. *Dataset* final CSV/JSON/Parquet dataset; source identifier; class label; intent labels; train/validation/test split; continual-learning round.
2. *Code* preprocessing script; deduplication script; annotation-
3. processing code; TF-IDF generation; model-training scripts; BERT training; super-learner training; continual-learning code; evaluation code.
4. *Generated Data* paraphrased phishing messages; generation prompts; model
5. name/version; generation parameters; timestamps.
6. *Results* raw predictions; probability scores; confusion matrices; per-
7. round metrics; hyperparameter-search results.

Without these artifacts, a reader can understand the proposed method but cannot fully reproduce the reported numerical results.

## **CONCLUSION**

This paper investigated an intent-aware continual-learning framework for phishing-email detection. The approach combines conventional TF-IDF text representations with four explicit intent cues: requested action, target asset, deception strategy, and urgency. Logistic regression, BERT, and a superlearner architecture were considered within a continual-learning framework using rehearsal.

The conceptual motivation is that phishing attacks can change their wording while preserving their underlying objectives. Explicit intent information may therefore provide complementary information that is less dependent on exact lexical patterns.

The supplied experimental manuscript reports improvements for the intent-aware approach under adaptive phishing scenarios. The supplied confusion matrices also indicate that the intentaware model makes fewer false-negative and false-positive errors than the reported text-only matrix at the evaluated point.

However, the numerical audit identifies important inconsistencies that must be acknowledged. The manuscript reports a 22,000-email corpus, but different sections describe incompatible class distributions. The training protocol is described as both 80/20 and 70/10/20. The continual-round size is inconsistent with the stated 22,000-email training split. The full-model F1 and FNR reported in the ablation table do not match the metrics derived from the supplied confusion matrix.

For this reason, the corrected manuscript does not present contradictory values as though they were simultaneously valid. The intent-aware confusion matrix yields 86.50% accuracy, 83.49% precision, 91.00% recall, 87.08% F1, and 9.00% FNR. The text-only confusion matrix yields 67.60% accuracy, 68.41% precision, 65.40% recall, 66.91% F1, and 34.60% FNR.

The central research direction remains promising, but the scientific strength of the study ultimately depends on resolving the dataset and reproducibility issues. In particular, the final TABLE XIII: Manuscript-Reported Dataset Composition

| Source       | Total  | Legit. | Phish. | Final  |
|--------------|--------|--------|--------|--------|
| Enron        | 15,000 | 7,500  | 7,500  | 15,000 |
| SpamAssassin | 5,000  | 2,500  | 2,500  | 5,000  |
| Nazario      | 2,000  | 1,000  | 1,000  | 2,000  |
| Total        | 22,000 | 11,000 | 11,000 | 22,000 |

dataset, raw predictions, exact split assignments, and generated adaptive examples should be released or archived so that every numerical result can be independently recalculated.

Future work should investigate larger and more recent phishing corpora, multilingual intent representations, automated intent extraction, stronger continual-learning algorithms, human-in-the-loop annotation, and evaluation against realistic attacker-generated messages.

A robust phishing-detection system should not be evaluated only by how accurately it classifies a fixed historical dataset. It should also be evaluated by how well it retains previously learned knowledge, adapts to genuinely new attacks, avoids source-specific shortcuts, and remains reproducible under independent examination.

## APPENDIX

### Dataset Arithmetic

If the claimed total is 22,000: 70% = 15,400, 10% = 2,200, 20% = 4,400, and  $15,400/4 = 3,850$  per continual-learning round.

### Intent-Aware Confusion Matrix

$TN = 410$ ,  $FP = 90$ ,  $FN = 45$ ,  $TP = 455$ ,  $N = 1,000$ .

Therefore: Accuracy = 86.50%, Precision = 83.49%, Recall =

91.00%, F1 = 87.08%, FNR = 9.00%.

### Text-Only Confusion Matrix

$TN = 349$ ,  $FP = 151$ ,  $FN = 173$ ,  $TP = 327$ . Therefore:

Accuracy = 67.60%, Precision = 68.41%, Recall = 65.40%, F1 = 66.91%, FNR = 34.60%.

The final published version should contain a table in this form after the actual project dataset has been verified: This table should replace the contradictory 22,000-email composition currently appearing in the manuscript until the raw dataset is checked.

**Table Xiv: Comparative Continual Learning Performance Across Adaptive Rounds**

| Model           | R1 F1  | R2 F1  | R3 F1  | R4 F1  | Final FNR |
|-----------------|--------|--------|--------|--------|-----------|
| Text-only LR    | 82.50% | 76.80% | 71.20% | 66.91% | 34.60%    |
| Intent-aware LR | 89.20% | 88.50% | 87.80% | 87.08% | 9.00%     |
| BERT            | 84.50% | 82.10% | 80.00% | 78.10% | 18.50%    |
| Super-Learner   | 87.20% | 85.60% | 84.10% | 82.40% | 14.20%    |

The remaining values should be generated directly from the raw prediction files rather than copied from the inconsistent manuscript tables.

label=□, leftmargin=\*

- Exact dataset file archived.
- Exact number of emails verified.
- Legitimate/phishing counts verified.
- 4,029 claim verified or rejected.
- 22,000 claim verified or rejected.
- Enron labeling procedure documented.
- SpamAssassin labeling procedure documented.
- Nazario version documented.
- Duplicate-removal count verified.
- Train/validation/test IDs archived.
- Continual-learning round IDs archived.
- Paraphrasing prompts archived.
- Generated phishing messages archived.
- Random seeds recorded.
- Hyperparameters recorded.
- Raw prediction files archived.

- Confusion matrices regenerated from predictions.
- Accuracy recalculated.
- Precision recalculated.
- Recall recalculated.
- F1 recalculated.
- FNR recalculated.
- IRI recalculated.
- Ablation results recalculated.
- Cross-domain results recalculated.
- Figures regenerated from raw results.
- Literature comparison removed unless datasets are genuinely comparable.
- References independently checked.
- Final Word document generated from the corrected manuscript.

## REFERENCES

- Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, 76(1), 139–154. <https://doi.org/10.1007/s11235-020-00733-2>
- Ejaz, A., Mian, A. N., & Manzoor, S. (2023). Life-long phishing attack detection using continual learning. *Scientific Reports*, 13, Article 11488.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).
- Klimt, B., & Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004* (Lecture Notes in Computer Science, Vol. 3201, pp. 217–226). Springer.
- Apache SpamAssassin Project. (n.d.). *SpamAssassin public mail corpus*. Apache Software Foundation, public email corpus.
- Nazario, J. (n.d.). *Nazario phishing corpus*. Public phishing-email corpus.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008).
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

- Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006). Can machine learning be secure? In *Proceedings of the ACM Symposium on Information, Computer and Communications Security* (pp. 16–25).
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).
- Tan, C., Chiew, K., Wong, K., Sze, S., & Tiong, W. (2022). A survey of phishing attack detection using machine learning techniques. *Journal of Network and Computer Applications*, 191, Article 103170.
- Li, Y., Xiao, R., Feng, J., Zhao, L., & Yan, W. (2020). A cross-lingual phishing detection approach using machine learning. *IEEE Access*, 8, 126233–126246.
- Kravchik, M., & Shabtai, A. (2023). Efficient cyber security anomaly detection using continual and online learning. *Computers & Security*, 132, Article 103328.
- Sharif, M., Jha, S., & Bauer, L. (2023). Adversarial machine learning in security: A survey. *ACM Computing Surveys*, 55(8).
- Basit, A., et al. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, 76(1), 139–154.
- A Systematic Review of Deep Learning Techniques for Phishing Email Detection. (2024). *Electronics*, 13(19), Article 3823.