Issues in Information Systems
Volume 26, Issue 4, pp. 382-397, 2025

DOI: https://doi.org/10.48009/4_iis_2025_131

Detecting phishing emails targeting healthcare practitioners: a domain-specific ensemble approach using diverse datasets

Gaston Elongha, *Marymount University, gle85144@marymount.edu*
Michelle Liu, *Marymount University, xliu@marymount..edu*

Abstract

The healthcare sector is increasingly targeted by sophisticated phishing emails generated by Large Language Models (LLMs) and advanced online tools. These emails often bypass traditional security measures, posing serious threats to patient safety and privacy. This study enhances domain-specific feature extraction by integrating advanced Natural Language Processing (NLP) techniques, specifically the transformer-based BERT model, to extract domain-specific phishing embeddings from health-related email content, improving overall detection accuracy. The proposed detection framework employs an ensemble stacking classifier that integrates Random Forest (RF) and Convolutional Neural Networks (CNN) as base models, with a Neural Network (NN) meta-learner for final classification. The models achieved high accuracy: RF at 96.4%, CNN at 97.3%, and NN at 97%, with ROC-AUC scores of 99% for CNN and RF, and 97% for NN. Evaluated on a dataset of 18,354 instances, including synthetic emails from ChatGPT-4o and Llama-3.1, and real-world samples from the Kaggle repository, the model demonstrated robust performance in identifying healthcare-targeted phishing attempts. While the results are promising, they are constrained by the dataset's characteristics. Future research will explore end-to-end transformer-based models to further enhance the detection of LLM-generated phishing threats in the healthcare sector.

Keywords: AI, domain-specific, healthcare phishing, cybersecurity, transformers, LLM

Introduction

Phishing, as a form of social engineering, exploits individuals' cognitive processes (Cranford et al., 2021), tricking them into disclosing confidential information or performing harmful actions (Altwaijry et al., 2024). It remains the primary vector in most cyber-attacks (Altwaijry et al., 2024), with incidents like spear phishing continuing to rise (Chua, 2021). The healthcare sector, which stores sensitive information such as Personally Identifiable Information (PII) and Protected Health Information (PHI/ePHI), has become a prime target.

Studies show that phishing is one of the top initial attack vectors leading to healthcare data breaches (Verizon, 2023). These breaches pose significant threats to patient privacy, potentially resulting in identity theft, ransomware, monetary loss, and exposure of confidential medical records, all of which fall under HIPAA regulation (Health Sector Cybersecurity Coordination Center (HC3), 2019). In addition to regulatory and financial consequences, such breaches damage institutional reputation and disrupt care delivery. In 2023, the average cost of a healthcare data breach reached $10.93 million per incident (Alder,

2023). These risks highlight the urgent need for advanced phishing detection systems tailored to healthcare environments.

There has been extensive research applying Machine Learning (ML) and Deep Learning (DL) algorithms to detect phishing emails. Commonly used ML techniques include Support Vector Machines (SVM) (Saleem, 2021), Random Forest (RF) (Espinoza et al., 2019), C4.5, CART, Decision Trees (DT), and K-Nearest Neighbors (K-NN) (Gholampour & Verma, 2023; Xiao & Jiang, 2020). DL methods such as Deep Neural Networks (DNN), Recurrent Neural Networks with Long Short-Term Memory (RNN-LSTM) (Li et al., 2022; Xiao & Jiang, 2020; Sachan et al., 2023), Convolutional Neural Networks (CNN) (Hussain et al., 2023; McGinley & Monroy, 2021; Alhogail & Alsabih, 2021; Atawneh & Aljehani, 2023), Feedforward Neural Networks (FNN), Restricted Boltzmann Machines (RBM), Deep Belief Networks (DBN), deep autoencoders (Brindha et al., 2023), and generic Neural Networks (NN) (Muralidharan & Nissim, 2023; Redondo-Gutierrez et al., 2022) have also been explored.

Despite these advancements, the focus has primarily been on analyzing email components like headers, bodies, URLs, or attachments. Some recent studies have expanded to whole-email analysis (Alshingiti et al., 2023; Muralidharan & Nissim, 2023; Rabbi et al., 2023; Saka et al., 2022), but few studies have examined AI-generated phishing emails in the healthcare domain, with one exception being Sameen et al. (2020), which focused solely on phishing URLs.

Attackers have increasingly turned to AI to automate and personalize phishing attacks at a scale. Meanwhile, defenders also leverage AI to identify and mitigate these threats (Kaur et al., 2023). Publicly accessible Large Language Models (LLMs) such as ChatGPT and Llama 3.1 further complicate this landscape (Malatji & Tolah, 2024; Sameen et al., 2020). These models, known for their ability to produce realistic phishing content (Roy et al., 2023; Langford & Payne, 2023), can generate highly fluent, context-aware phishing emails that mimic legitimate communications and bypass traditional rule-based security mechanisms, such as heuristic filters, static keyword detection, and warning banners (Newman, 2021).

Current defense mechanisms, though varied, often lack the dynamic adaptability required to counter these AI-generated phishing attacks. As a result, healthcare practitioners, who may lack cybersecurity expertise, are particularly vulnerable to these advanced phishing tactics. Phishing emails targeting healthcare practitioners often combine general phishing indicators with health-specific language, such as reference to patient care, accidents, and prescription access. Terms such as "*patients*," "*drugs*," and "*caregivers*" are commonly exploited. The advanced NLP capabilities of LLMs allow attackers to craft messages that closely resemble genuine healthcare communications (Newman, 2021), increasing the likelihood of deception (Malatji & Tolah, 2024).

Given the limitations of legacy email security systems in detecting domain-specific phishing threats, we developed a detection framework based on Ensemble Learning (EL) techniques to capitalize on the complementary strengths of multiple models. Building on earlier work (Author, SSRN preprint, blinded for review) which forms part of an ongoing doctoral dissertation focused on domain-specific phishing detection in healthcare, we expand the dataset by generating synthetic phishing emails using ChatGPT-4o and Llama 3.1 and supplement it with real-world healthcare-related phishing samples from Kaggle. The detection framework integrates Random Forest (RF) and Convolutional Neural Networks (CNN) as base models, and a Neural Network (NN) as the meta-learner.

To improve detection accuracy and generalization, we replace manual feature engineering with domain-specific embeddings automatically extracted using BERT (Bidirectional Encoder Representations from Transformers). This approach enables the model to better capture contextual indicators of phishing, particularly health-domain-specific terms such as "*prescription*," "*emergency care*," and "*patient records*."

The contributions of this study are multifaceted: (1) It addresses the emerging challenge of detecting LLM-generated, healthcare-specific phishing emails; (2) It offers a strategy to mitigate risks to Protected Health Information (PHI) and ensure HIPAA compliance; (3) It supports uninterrupted patient care by enhancing email security for practitioners; and (4) It advances AI-based detection by leveraging transformer-based embeddings and ensemble learning for improved classification.

The remainder of the paper is structured as follows: Section Two reviews ML and DL algorithms for phishing detection. Section Three details the updated methodology. Section Four covers the experiments, including preprocessing feature engineering, performance measures, and presents the results. Section Five presents the discussion and limitations, and the last section concludes with recommendations and future research direction.

Literature Review

Detecting phishing emails is commonly approached as a classification task using either ML or DL) techniques. These approaches vary depending on which email components are analyzed—such as headers, body text, URLs, or attachments. To understand the current state of research, we reviewed systematic review articles across four main categories: header analysis, body-text detection, URL-based detection, and attachment-based detection. We also examined approaches that analyze entire email messages holistically to improve overall detection accuracy.

Machine Learning Approaches

ML-based phishing detection often uses supervised classification, where datasets are split into training and testing subsets following preprocessing. Fig.1 illustrates a typical ML classification architecture.

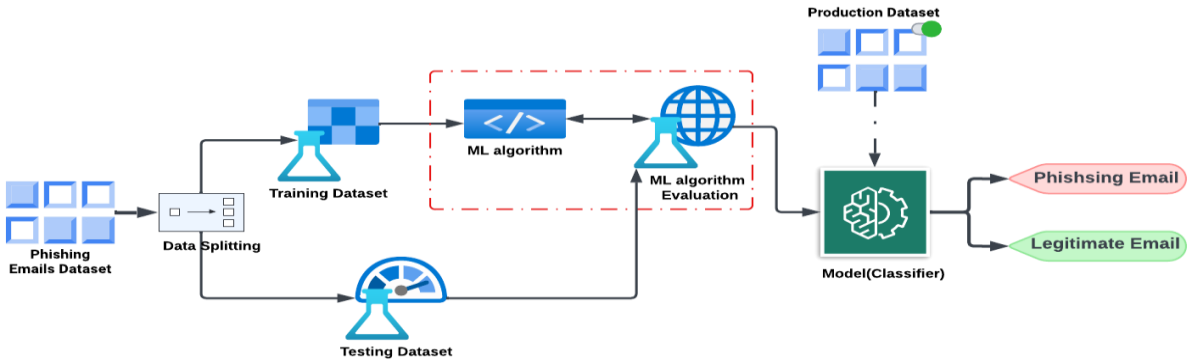


Figure 1. Machine Learning Phishing Classification Architecture

Studies focusing on header analysis have achieved strong results. Gholampour and Verma (2023) used K-Nearest Neighbors (KNN) algorithms, achieving 94% accuracy and validating their model with GPT-2-generated emails. Espinoza, et al. (2019) focused on email headers using RF and Logistic Regression (LR), achieving 96.77% accuracy. Saleem (2022) tested various ML models with psychological features, reporting accuracies of 95% for DT, 93% for Naive Bayes (NB), and 98% for RF. While most body-text detection studies rely on DL or ensemble models, few ML-specific experiments exist. For URL-based detection, Sameen et al. (2020) used an ensemble model with SVM, achieving 98% accuracy on AI-generated URLs and 97% on human-generated ones. Aung and Yamana (2022) reached 95.7–97.7% accuracy using a hybrid tokenizer that combined BERT and Word Segment techniques. Attachment-based detection studies also showed strong results. Scofield et al. (2020) analyzed attachments, such as PDF

document, using a heuristic rule-based classifier, achieving 98% accuracy. Tiruthani (2009) focused on detecting malicious links within attachments using a heuristic-based algorithm to achieve 99% accuracy. Lastly, research on analyzing entire emails holistically remains sparse. Rabbi et al. (2023) conducted a comprehensive study using six ML models to classify complete emails, concluding with RF and LR models both achieving 98% accuracy. Despite these results, they emphasized that DL models like CNN could further improve detection accuracy.

Deep learning Approaches

DL models often outperform ML counterparts due to their ability to learn representations directly from raw data (Basit et al., 2020). Fig. 2 illustrates the DL architecture used for phishing email detection, showing how data flows through various layers to classify emails.

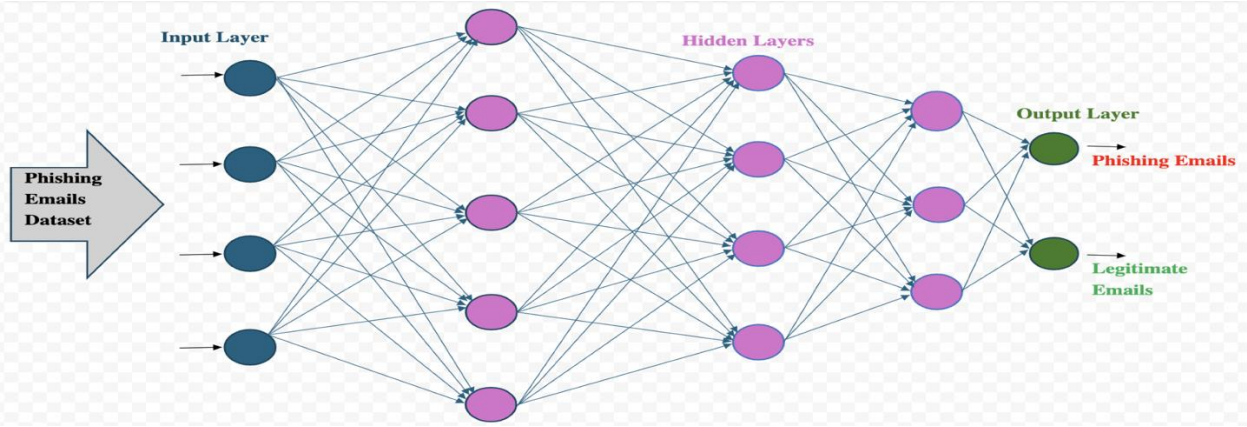


Figure 2. Deep Learning Phishing Classification Architecture

DL has demonstrated significant success in classifying email content, including body text, headers, and attachments. Xiao and Jiang (2020) used KNN and Bi-LSTM for multi-component detection, achieving 90% accuracy. Li et al. (2022) improved on LSTM models with 95% for body-text detection. Alhogail and Alsabih (2021) applied Graph Convolutional Networks (GCN) with NLP, reaching 98.2%, while McGinley and Monroy (2021) reported 99.14%. URL-based DL detection has also advanced. Hussain et al. (2023) achieved 99% with CNN. AlEroud and Karabatis (2020) used Generative Adversarial Networks (GANs) to create evasive phishing links. Siddiq et al. (2022) achieved 93–95% with CNN and NN. Aslam et al. (2023) used LSTM for body-text detection, reaching 92.46%.

Attachment detection using DL has benefited from LSTM-KNN hybrids (Li et al., 2022), which improved accuracy and reduced false results. Muralidharan and Nissim (2023) achieved 99% accuracy in full-email analysis. Saka et al. (2022) used clustering (e.g., K-Means, DBSCAN) and found context-rich features like subject lines and body text to be most effective. Alshingiti et al. (2023) validated DL's dominance, achieving 99.2% (CNN), 97.6% (LSTM), and 96.8% (CNN-LSTM hybrid).

Notably, several recent models have achieved near-perfect accuracy. Brindha et al. (2023) reported 99.72% using a DL framework optimized with Cuckoo Search. Atawneh and Aljehani (2023) achieved 99.61% using a combination of CNN, LSTM, RNN, and BERT. Sachan et al. (2023) reached 97–98% with LSTM and Bi-LSTM.

Synthesis of Systematic Review Work

Systematic reviews affirm that strong feature extraction underpins high-performing phishing detection. Basit et al. (2020) conducted a comprehensive review of various detection algorithms used for identifying URLs in emails. Their findings highlighted the critical role of feature extraction in achieving high model performance, identifying CNN and DNN as particularly effective for URL classification. They also noted that the use of ensemble or stacking models could enhance precision up to 97.61%. Basit et al. (2020) also revealed that, among ML algorithms, DT, RF, SVM, NN, LR, and Naive Bayes (NB) are the most effective for URL classification, particularly with proper feature extraction. They emphasized the importance of well-balanced feature sets. In terms of significance and high accuracy, C4.5, KNN, and SVM were often used based on DT classifier. Overall, ML algorithms such as SVM, RF, ANN, C4.5, K-NN, DT, LR, Local Coupled Extreme Learning Machine (LC-ELM) and XGB with features extraction techniques like ANOVA and RRFST achieved up to 99.2% accuracy, with RF alone reaching 95% accuracy for URL detection.

Catal et al. (2022) reviewed 43 articles and found DL models, especially RNN-LSTM, excelled even without extensive feature engineering. Safi & Singh (2023) echoed this, highlighting methods like Feature Fusion Depth Neural Network (FFDNN), Lattice Boltzmann Method (LBM), and DaEncoder, while noting RF remains widely used. Moreover, Bountakas (2021) conducted a comparative review of NLP and ML algorithms focusing on text-based email body detection through sentiment analysis. They concluded that NLP and sentiment analysis using Term Frequency-Inverse Document Frequency (TF-IDF), Word2vec, BERT are the most used techniques with high accuracy. The ML algorithms with the highest accuracy included RF, DT, LR, Gradient Boosting Trees (GBT), and NB. Similarly, Adesoji & Yamazaki (2023) reviewed articles detecting phishing emails while focusing on human factor and sentiment analysis using NLP and ML for classification. Their findings mirrored those of Bountakas (2021)'s, with Word2vec and FT-IDF being the most used for features extraction, while KNN, SVM, LR and RF emerging as the best performing models.

Finally, Valecha et al. (2022) and Nishikawa et al. (2020) emphasized the importance of persuasion cues in phishing detection. Both studies emphasized that consideration for persuasion techniques and human factors when developing phishing email detection is critical. The more persuasive phishing emails are, the more likely healthcare practitioners are to be deceived, potentially leading to breaches that compromise patient privacy and overall healthcare settings. Persuasion cues based on Cialdini's persuasion tactics, such as authority, scarcity (lost framing), consistency (promises), likeability (establishing trust), and reciprocity (reward or gain framing), could significantly enhance model accuracy and performance in detecting phishing emails. These reviews highlighted that DL techniques like CNN and Region-Based Convolutional Neural Networks (RCNN) are effective in extracting targeting human vulnerabilities. This knowledge was particularly relevant to our research because malicious actors increasingly leverage Generative AI's NLP capabilities to evade the existing healthcare email security by exploiting psychological persuasion.

Methodology

Drawing from prior literature and our own observations in healthcare cybersecurity, this study identifies key features in phishing emails targeting healthcare practitioners. These emails often employ urgent issues such as medical bill payments, account lockouts preventing prescription access, emergency care needs, or impersonation of patients, in addition to common phishing tactics (i.e., *authority, scarcity*, and *urgency*). While prior studies have shown that combining multiple email components (headers, bodies, and subject lines) improves classification performance, few efforts have emphasized domain-specific feature extraction. To address this, we apply a transformer-based BERT model to derive contextual embeddings from health-related email content. These embeddings are transferred to an EL classification model.

This advanced feature engineering methodology was inspired partially by the work of Vaswani et al. (2017), who presented how attention-based models can be leverage for extracting embeddings that can be used for diverse tasks, including classification. For the study feature engineering consists of merging the email text header and body. To generate the embeddings, we then applied BERT's special tokens, including the [CLS] representing the whole input sequence and [SEP] being the separator between distinctive segments of text, then processed as a single sequence input that encapsulated the entire email contextual information.

The main rationale is rooted in several factors that make EL particularly effective for phishing email detection. Phishing email detection is inherently a classification problem, requiring a model to accurately distinguish between legitimate and phishing emails. The EL approach enhances classification, robustness, and accuracy. It integrates predictions from multiple models, which allows it to deliver more reliable results than single-model approaches (Murel & Kavlakoglu, 2024). EL remains robust even when working with limited or diverse datasets. In the healthcare sector, data privacy concerns often result in smaller datasets, which can constrain the performance of standalone models. However, EL techniques, particularly those involving stacking classifiers, are known to mitigate overfitting and improve model generalizability, even with smaller datasets (Kunapuli, 2023). EL enables integration of ML and DL models to capture diverse features, including CNN and RF as base models, and a Neural Network (NN) as meta-learner to enhance detection accuracy and robustness by optimizing the combined outputs (Sameen et al., 2020).

EL techniques can be broadly categorized into two types: parallel and sequential (Murel & Kavlakoglu, 2024). In this study, we applied the parallel method, where multiple base learning models are trained independently, and their predictions are then combined using a stacking classifier (Fig. 3). This approach was selected due to its effectiveness in handling classification tasks and optimizing prediction accuracy (Murel & Kavlakoglu, 2024).

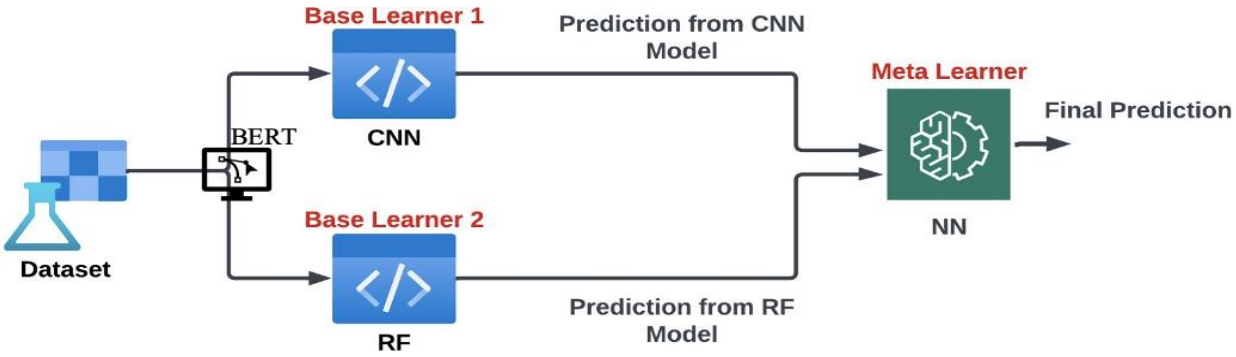


Figure 3. Proposed Stacking Classifier

The model architecture, illustrated in Fig. 3, shows how predictions from the base models (RF and CNN) are stacked and fed into the NN meta-learner for final classification. CNN and RF were selected due to their effectiveness in the literature review and robustness history. The NN meta-learner processes the outputs from CNN and RF by mimicking human-like pattern recognition (Pham & Sun, 2024). This approach leverages NN's capacity to self-optimize through multi-layer processing, allowing it to learn patterns and adapt weights effectively (Sameen et al., 2020). The integration of these models enables rapid and accurate classification of phishing emails, particularly those leveraged by LLMs' advanced NLP capabilities.

Datasets

We used two datasets for this experiment. The first is a real-world dataset containing 18,600 healthcare-related emails, both legitimate and phishing, sourced from a public Kaggle repository (Chakraborty, 2023). The second comprises two synthetic datasets generated by ChatGPT-4o and Llama 3.1 using healthcare-

specific prompts that incorporate both phishing tactics and domain-relevant terminology. These synthetic emails were developed based on insights from our healthcare cybersecurity experience and findings from the reviewed literature. Phishing emails, regardless of the target sector, often incorporate urgent keywords and persuasion tactics to exploit human vulnerabilities, as proven by Nishikawa et al. (2020). As an example, we identified a few healthcare-specific keywords commonly used in phishing emails, such as '*medical bill*,' '*account lockout*,' '*prescription*,' '*emergency care*,' and *'patient information*'.

Building on these validities, we developed a series of prompts that combined healthcare-specific elements with general phishing tactics, including keywords, urgency cues, and psychological manipulation strategies (see example in Fig. 4). This was critical in guiding LLMs (ChatGPT-4o mini and Llama3.1) used to learn from both general phishing email patterns and healthcare-associated scenarios, enabling them to generate phishing emails tailored to healthcare practitioners. The prompts ensured the two LLMs could effectively replicate the blend of sector-specific and general tactics observed in phishing emails, producing two datasets reflective of the threats faced by healthcare practitioners (see Table 1).

> """*Help me generate phishing emails targeting healthcare practitioners.*
>
> *Here are specific distributions or patterns I would like to see in the sender, subject lines, or body text:*
>
> *Mostly urgent problems such as medical bill payments, account lockout preventing access to prescription,*
>
> *drugs prescription, immediate care needs, emergency, and patients, executive, or manager impersonation.*
>
> *In addition to that, common phishing tactics such as authority, scarcity, and urgency like time-sensitive,*
>
> *act, click here, input needed asap, and impersonation. Learn from this distributions and built up mixed and advanced patterns* """

Figure 4. Sample of LLMs (ChatGPT-4o mini and Llama3.1) Prompt

As shown below, Table 1 provides a detailed breakdown of the health-related dataset collected from Kaggle and the synthetic dataset generated using the prompts for the experiment.

Table 1. Dataset Repartition

	Source	Size
Health-related email	Kaggle repository (Chakraborty, 2023)	**18,600** (legitimate: 11,322, and phishing: 7,328)
LLMs-generated	(ChatGPT-4o mini and Llama3.1)	**21,110** (legitimate: 10,888; phishing: 10,222)
	39,710 (legitimate 22, 210; phishing 17, 550**)**	

To eliminate redundancy in synthetic data, a cleaning function was applied to the merged datasets. This resulted in a total of 11,647 legitimate (63.5%) emails and 6707 phishing (36.5%) emails. The minority label in this case being the phishing emails with 36.5%, the validations from previous studies emphasized that the EL-stacking classifier's capacity to handle smaller datasets (Kunapuli, 2023; Kavlakoglu, 2024). However, dataset balance depends on the application and problem being solved. Ideally data is balanced

when it has equal number of class or label samples. Since phishing emails are typically underrepresented in real-world scenarios, this class imbalance was expected.

The acceptable class ratio in classification tasks typically ranges from 60:40 to 70:30, which remains manageable under the ensemble stacking approach employed in this study. To address class imbalance, we used the RandomOverSampler technique to increase the proportion of phishing emails from 36.5% to approximately 40%, achieving a moderate class distribution. Reducing the number of legitimate emails would risk data loss and potentially harm model generalizability, so oversampling the minority class was a more effective strategy. The final dataset, after oversampling, is summarized in Table 2 and reflects a class imbalance level that ensemble learning models are well-equipped to handle. The next section describes the experimental design, including the process for extracting domain-specific phishing embeddings and tuning model hyperparameters.

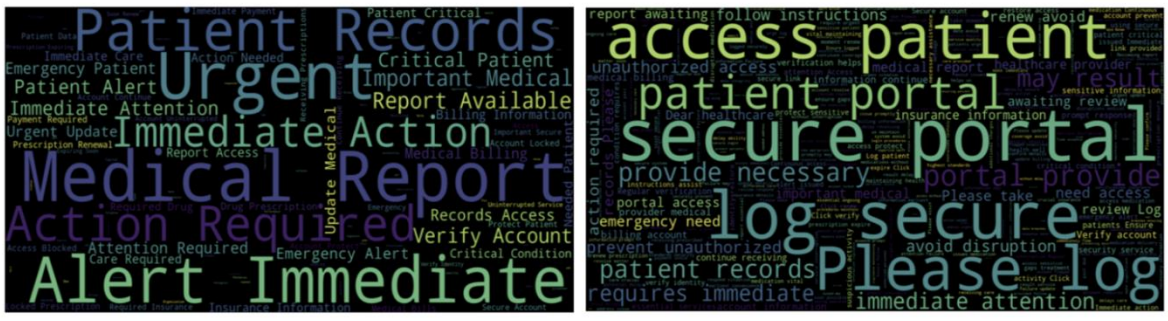
Table 2. Overall Dataset Distribution

	Initial Data Distribution	Resampled Data Distribution
Label 0 (legitimate emails)	11,647 (~63.5%)	11,647 (63.5%)
Label 1 (phishing emails)	6707 (~36.5%)	7,341 (40%)
		18, 988

Experiment Design

Experiments were conducted on a Dell OptiPlex 7050 Ubuntu 24.04.2 LTS OS with an Intel Core i7-7700 processor, 32 GB RAM, NVIDIA GeForce GTX 1650 graphics card, and a 2TB hard drive. Model development was implemented using Python within Visual Studio Code's Jupyter Notebook integration environment, supported all necessary for ML and DL libraries.

Textual processing functions including remove_stopwords(), strip_html(), and emove_between_square_brackets() were combined into a denoise_text() pipeline to eliminate irrelevant email text content. As shown in Fig. 5, phishing emails frequently contained urgent terms and persuasive tactics, while legitimate emails, as shown in Fig. 6, exhibited different tones, grammatical consistency, and domain-specific language to the healthcare sector. This analysis informed feature engineering and model training.



(a) Word Cloud for Phishing Emails-Header (b) Word Cloud for Phishing Emails-Body

Figure 5. Word clouds of Phishing Emails Content

(a) Word Cloud for Legitimate – Header (b) Word Cloud for Legitimate-Body

Figure 6. Word clouds of Legitimate Emails Content

Figure 7 illustrates the frequently trusted domains identified, reflecting typical communication patterns in healthcare organizations, including internal domains and trusted entities like government and medical partners. In real-world settings, a given dataset would reveal an additional and predefined domains list, often logged by email security systems such as Proofpoint.

Figure 7. Domain Cloud of Frequently Trusted Domains

Features Extraction using BERT

The feature engineering and extraction were updated from manually selected features to the use of transformer—pre-trained BERT and its tokenizer. This advanced approach directly extracts embeddings from text and semantic health related information of each email. Specifically, we employed *BertTokenizer* and *BertModel* from '*Bert-base-uncased*' variant, given its application and generalization, especially in name entity recognition, context, sentiment analysis, where subtle shifts in communication tone are critical. We employed a *def extract_bert_embeddings_batch* function to compute all embeddings, processing emails text in a 32-batch size.

We also ensured proper tensor formatting, with a maximum length of 512, handling overall model outputs, and truncating input sequences. Because attackers increasingly exploit the NLP capabilities of LLMs, a static list of keywords is insufficient to capture the nuanced intent within email text. Leveraging BERT, which is pre-trained on a large corpus, enabled our ensemble learning (EL) models to focus solely on the classification task without requiring additional large-scale training or feature engineering. As demonstrated by Vaswani et al. (2017) in "Attention is All You Need," BERT's transformer-based architecture provides powerful generalization for downstream tasks such as text classification and sentiment analysis, particularly when working with small or imbalanced datasets. This transfer learning approach mitigates overfitting risks and provides robust, pre-trained vector representations of input data, significantly reducing both the errors

associated with manual feature extraction and the computational cost of training EL models from scratch. The BERT embeddings were saved to a corresponding file and loaded as X_bert, with its corresponding encoded label (y), then split into training, testing and normalized using a StandardScaler after split before the models training.

Performance Metrics Evaluation

At their default settings, the base models, RF and CNN, produced the results shown in Table 3. These outcomes reflect the influence of default hyperparameters and residual class imbalance, despite oversampling to achieve a moderate 60:40 ratio. While the RF model's performance was modest, the EL approach demonstrated a more balanced and effective classification of domain-specific phishing emails in the healthcare context.

Table 3. Default Settings Metrics

Base Models	Accuracy	Precision	Recall	F1-score
RF	0.930 ~ 93%	0.932 ~ 93.2%	0.93 ~ 93%	0.93 ~ 93%
CNN	0.968 ~ 96.8%	0.968 ~ 96.8%	0.968 ~ 96.8%	0.968 ~ 96.8%
Meta Learner	**Accuracy**	**Precision**	**Recall**	**F1-score**
NN	0.968 ~ 96.8%	0.968 ~ 96.8%	0.968 ~ 96.8%	0.968 ~ 96.8%

The CNN model was tuned from the computed hyperparameters equal to adam optimizer, learning rate of 5e-5, dropout rate of 0.3, batch size of 64 and epochs 64, while using a callback early stopping on best validation recall and accuracy, with a patience of 10 and class weight {0: 1, 1: 2.5} for minority class. For the RF models, a RandomizedSeaarchCV tuning was applied with few interactions for faster tuning, using the Principal Component Analysis (PCA) that reduced dimensionality of scaled training and testing. The class weight was balanced, n_estimator of 200, min_samples_split and leaf of 7 and 2, and maximum depth of 25.

The NN-meta learner processing base models' predictions was tuned with optimal hyperparameters to avoid missing important predictions from the CNN and RF. Computed hyperparameters for the NN-meta learner were respectively 32 for batch size, 10 epochs, dropout rate of 0.4, half CNN callbacks early stopping patience. Any other hyperparameter tuning matched the previous CNN tuning parameters. As we tuned the base models, the metrics recorded (see Table 4) illustrate the domain-specific phishing emails classification capabilities of CNN and RF—base models and NN—meta learner's metrics after tuning. To demonstrate their capabilities of distinguishing legitimate email from LLMs-generated and advanced phishing emails targeting healthcare practitioners, this version added the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) scores to the computed tuning process. The ROC-AUC given the study objectives, and update, is crucial to demonstrate how accurate these tuned-model are to correctly classify and distinguish true phishing email from legitimate emails.

Table 4. Tuned-base EL Models Metrics

Base Models	Accuracy	Precision	Recall	F1-score	ROC-AUC
RF	0.964 ~ 96.4%	0.964 ~ 96.4%	0.964 ~ 96.4%	0.964 ~ 96.4%	0.994 ~ 99.4%
CNN	0.971 ~ 97.1%	0.971 ~ 97.1%	0.971 ~ 97.1%	0.971 ~ 97.1%	0.990 ~ 99%
Meta Learner	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC-AUC**
NN	0.971 ~ 97.1%	0.971 ~ 97.1%	0.971 ~ 97.1%	0.971 ~ 97.1%	0.970 ~ 97%

Phishing email detection is a challenging classification problem, particularly as attackers leverage LLMs and advanced AI tools to generate convincing phishing emails. Sensitivity (recall) and ROC-AUC were especially critical for this update version. Their higher scores (Table 4) tell on the models' ability to effectively distinguish true positives phishing emails from legitimate emails. They help ensure minority (phishing) and majority (legitimate) class balance in addition to the class weight tuning applied during the tuning process.

The displayed 97.1% sensitivity of the NN is particularly important given the imbalanced nature of phishing datasets ratio. This dataset nature is a general case for phishing email detection, since in real world settings, legitimate emails often outnumber phishing instances. Regardless, high sensitivity and ROC-AUC scores of ~0.97% being close to 1 indicate that the meta learner—NN and its base models—CNN and RF did not overlook the minority class—phishing emails—and demonstrated exceptional robustness in class distinction.

Figure 8demonstrates the NN's exceptional robustness through the ROC curve and Training Accuracy (TA) and Validation Accuracy (VA) plots. The ROC curve demonstrates a solid true positive rate (TPR) against false positive rate (FPR). Since the ROC curve is above and further from the random guess, it demonstrates that the NN has excellent discriminating class balance. The NN training and validation plot demonstrates more insights in addition to the ROC curve. The VA initially exhibited low accuracy but quickly improved, surpassing the TA accuracy after just one and half epochs. This demonstrates the NN's rapid generalization in understanding the existing class imbalance and its adaptive ability, leading to an excellent discriminatory power that balances class weight.

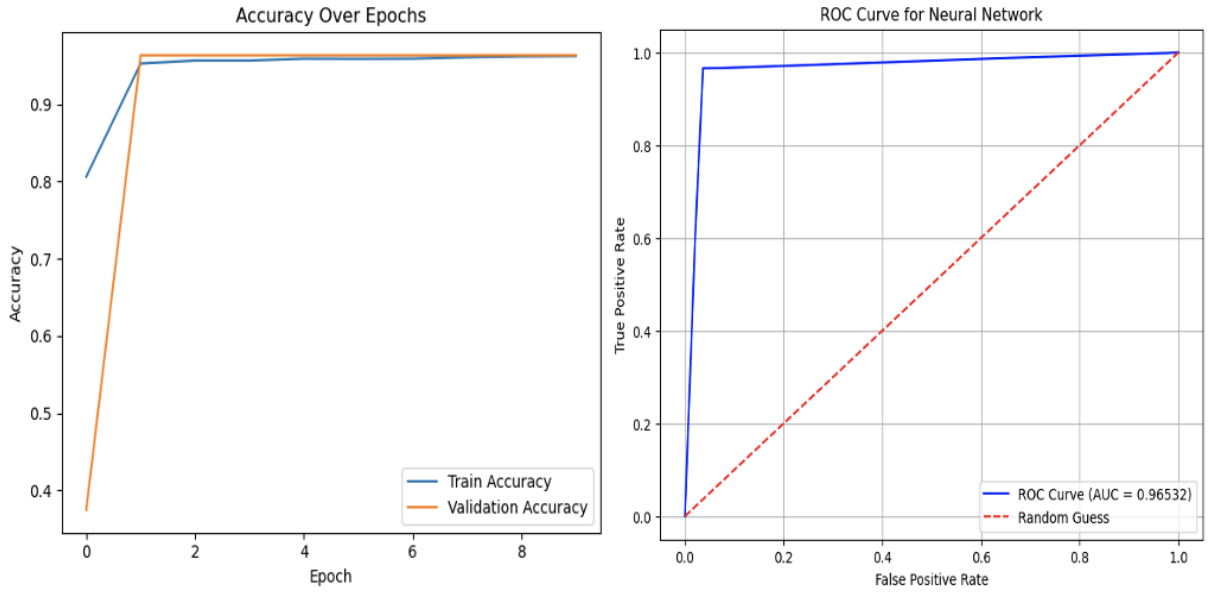


Figure 8. NN ROC Curve & TA/VA

Further testing revealed that using transformers-based approach for both feature extraction and classification as an alternative to the NN for final email classification. However, in this experiment, the NN and its base models required minimal computational resources, given the advanced BERT embeddings feature we extracted. The models relied on domain-specific embeddings and optimized predictions from the pre-tuned base models.

Discussion

The final performance metrics presented in Table 4 reflect the effectiveness of a simple model architecture when combined with advanced feature extraction techniques. While the architecture itself—comprising Random Forest (RF), Convolutional Neural Network (CNN), and a Neural Network (NN) meta-learner—may appear straightforward, the integration of transformer-based BERT embeddings enabled robust classification of domain-specific phishing emails targeting healthcare practitioners. The use of BERT, known for its advanced natural language processing (NLP) capabilities, allowed knowledge transfer to the underlying models, thereby eliminating the need for additional complex architecture or extensive manual feature engineering. These results are notable, particularly given the limited existing research on phishing email detection specific to the healthcare domain. As such, this study provides early evidence that BERT embeddings, when combined with ensemble learning, can support effective detection of sophisticated phishing emails tailored to the healthcare sector. The unique approach employed in this study highlights how domain-specific models can be developed even with moderate datasets, provided that NLP-based embeddings capture the contextual and semantic cues associated with phishing content.

While the findings are promising, there are several limitations. First, transformer-based models like BERT typically require large, diverse datasets and substantial computational resources to achieve optimal performance. Although BERT was used here for feature extraction rather than training from scratch, the broader adoption of transformer-based classification approaches would necessitate significantly more data and computing power. To address this, the ongoing doctoral dissertation associated with this study is working to expand the dataset and further explore the use of transformers for end-to-end phishing detection.

Second, access to real-world healthcare phishing datasets remains a major barrier. Due to privacy, ethical, and proprietary restrictions, such datasets are rarely available in the public domain. Consequently, this study relied on synthetic data generated using LLMs and publicly available health-related datasets to simulate realistic phishing scenarios. This approach was informed by the research team's experience in healthcare cybersecurity and ongoing observations of phishing trends within the sector. While not a perfect substitute, this method enabled the development of a domain-relevant dataset that closely mirrors real-world threats. the cybersecurity healthcare settings. Lastly, the study focused on detecting phishing emails with static content (i.e., body text and headers) and did not incorporate dynamic or behavioral signals such as user interaction patterns, email reply to chains, or click-based engagement. This limits the applicability of the model in real-world settings where advanced phishing campaigns may evolve dynamically or involve multi-stage tactics. Future work could expand the detection scope to include behavioral features or time-series modeling of email sequences.

Conclusion

This study employed advanced feature extraction using BertTokenizer and BertModel from ‘Bert-base-uncased' variant, given its application and generalization, especially in name entity recognition, context, sentiment analysis, and healthcare where communication tone changes are critical. The BERT embeddings were leveraged through a transfer learning to train and validate the proposed Ensemble Learning (EL) detection architecture. The EL included base-models—CNN with 97.1% of sensitivity and 99% of ROC-AUC score; RF with 96.4% of sensitivity and 99,4% of ROC-AUC score. The final classifier was a stacking classifier, which was a meta-learner—NN with 97.1% of sensitivity and 97% of ROC-AUC score. This research is among the first to apply AI-driven domain-specific phishing detection in the healthcare sector, focusing on emails generated by LLMs and advanced tools targeting healthcare practitioners. The study drew on researchers' expertise in cybersecurity and showed that a domain-specific phishing email

classification is effective to mitigate these emerging phishing emails. The results demonstrate a robust classification, supported by the study's diverse datasets (comprising two LLMs-generated datasets and data collected from online repository), highlighting their validation in healthcare settings. Future research, including the ongoing doctoral dissertation, will fully leverage transformer-based approaches for classification of healthcare domain-specific phishing emails, enhancing both robustness and practical applicability. Combining datasets from multiple healthcare entities will improve model generalizability, particularly given the capability of transformers-based approaches to handle larger and more diverse datasets.

References

Adewale, A.E., & Yamazaki, T. (2023). Fundamental Sentiment Analysis by Natural Language Processing and Machine Learning for Email Classification. In Proceedings of the 2023 5th Asia Pacific Information Technology Conference (APIT '23). *Association for Computing Machinery*, New York, NY, USA, 103–105. https://doi.org/10.1145/3588155.3588171

Alder, S. (2023, December 15). The Average Cost of a Healthcare Data Breach is Now $9.42 Million. *The HIPAA Journal*. Retrieved August 15, 2024, from https://www.hipaajournal.com/average-cost-of-a-healthcare-data-breach-9-42-million-2021/

AlEroud, A. & Karabatis, G. (2020). Bypassing Detection of URL-based Phishing Attacks Using Generative Adversarial Deep Neural Networks. In Proceedings of the Sixth International Workshop on Security and Privacy Analytics (IWSPA '20). *Association for Computing Machinery*, New York, NY, USA, 53–60. https://doi.org/10.1145/3375708.3380315

Alhogail, A., & Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. *Computers & Security*, *110*, 102414. https://doi.org/10.1016/j.cose.2021.102414

Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Qazi Emad, U. H., Saleem, K., & Muhammad, H. F. (2023). A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN. *Electronics, 12*(1), 232. https://doi.org/10.3390/electronics12010232

Altwaijry, N., Al-Turaiki, I., Alotaibi, R., & Alakeel, F. (2024). Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models. *Sensors* (Basel, Switzerland), 24(7), 2077. https://doi.org/10.3390/s24072077

Aslam, S., Aslam, H., Manzoor, A., Chen, H., & Rasool, A. (2024). AntiPhishStack: LSTM-Based Stacked Generalization Model for Optimized Phishing URL Detection. *Symmetry*, 16(2), 248. https://doi.org/10.3390/sym16020248

Atawneh, S., & Aljehani, H. (2023). Phishing email detection model using Deep learning. *Electronics*, *12*(20), 4261. https://doi.org/10.3390/electronics12204261

Aung, E.S. & Yamana, H. (2022). Segmentation-based Phishing URL Detection. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '21). *Association for Computing Machinery*, New York, NY, USA, 550–556. https://doi.org/10.1145/3486622.3493983

Basit, A., Zafar, M. Q., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2020). 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

Bountakas, P., Koutroumpouchos, K., & Xenakis, C. (2021). A Comparison of Natural Language Processing and Machine Learning Methods for Phishing Email Detection. In Proceedings of the 16th International Conference on Availability, Reliability and Security (ARES '21). *Association for Computing Machinery*, New York, NY, USA, Article 127, 1–12. https://doi.org/10.1145/3465481.3469205

Brindha, R., Nandagopal, S., Azath, H., Sathana, V., Joshi, G. P., & Kim, S. W. (2023). Intelligent Deep learning-based Cybersecurity phishing email detection and classification. *Computers, Materials & Continua*, *74*(3), 5901–5914. https://doi.org/10.32604/cmc.2023.030784

Çatal, Ç., Giray, G., Tekinerdoğan, B., Kumar, S., & Shukla, S. (2022). *Applications of deep learning for phishing detection: a systematic literature review. Knowledge and Information Systems*, *64*(6), 1457–1500. https://doi.org/10.1007/s10115-022-01672-x

Chua, J. (2021). *Cybersecurity in the Healthcare Industry*. *The Journal of Medical Practice Management: MPM*, 36(4), 229-231. Retrieved February 20, 2024, from https://www.proquest.com/docview/2504559001?sourcetype=Scholarly%20Journals

Cranford, E. A., González, C., Aggarwal, P., Tambe, M., Cooney, S., & Lebière, C. (2021). Towards a Cognitive Theory of Cyber Deception. *Cognitive Science*, 45(7). https://doi.org/10.1111/cogs.13013

Espinoza, et al. (2019). Phishing Attack Detection: A Solution Based on the Typical Machine Learning Modeling Cycle, *in 2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, pp. 202-207. https://doi.org/10.1109/CSCI49370.2019.00041

Gholampour, P.M & Verma, R.M. (2023). Adversarial Robustness of Phishing Email Detection Models. In Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics (IWSPA '23). *Association for Computing Machinery*, New York, NY, USA, 67–76. https://doi.org/10.1145/3579987.3586567

Health Sector Cybersecurity Coordination Center (HC3). (2019). A cost analysis of healthcare sector data breaches. Retrieved June 8, 2024, from https://www.hhs.gov/sites/default/files/cost-analysis-of-healthcare-sector-data-breaches.pdf

Hussain, M., Cheng, C. K., Xu, R., & Afzal, M. (2023). CNN-Fusion: An effective and lightweight phishing detection method based on multi-variant ConvNet. *Information Sciences*, 631, 328–345. https://doi.org/10.1016/j.ins.2023.02.039

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 97, 101804. https://doi.org/10.1016/j.inffus.2023.101804

Kunapuli, G. (2023). Ensemble Methods for Machine learning. *Simon and Schuster*.

Langford, T., & Payne, B. (2023). Phishing Faster: Implementing ChatGPT into Phishing Campaigns. In *Lecture notes in networks and systems* (pp. 174–187). https://doi.org/10.1007/978-3-031-47454-5_13

Li, Q., Cheng, M., Wang, J. and Sun, B. (2022). LSTM Based Phishing Detection for Big Email Data, *IEEE Transactions on Big Data*, vol. 8, no. 01, pp. 278-288. https://doi/org/10.1109/TBDATA.2020.2978915

Malatji, M., & Tolah, A. (2024). Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. AI And Ethics. https://doi.org/10.1007/s43681-024-00427-4

McGinley, C. and Monroy, S. (2021). "Convolutional Neural Network Optimization for Phishing Email Classification," *in 2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021 pp. 5609-5613. https://doi.org/10.1109/BigData52589.2021.9671531

Muralidharan, T., & Nissim, N. (2023). Improving malicious email detection through novel designated deep-learning architectures utilizing entire email. *Neural Networks*, 157, 257–279. https://doi.org/10.1016/j.neunet.2022.09.002

Murel, J., & Kavlakoglu, E. (2024, March). What is ensemble learning? IBM. Retrieved July 20, 2024, from https://www.ibm.com/topics/ensemble-learning

Newman, L. H. (2021, August 7). AI wrote better phishing emails than humans in a recent test. *WIRED*. Retrieved June 8, 2024, from https://www.wired.com/story/ai-phishing-emails/

Nishikawa, H., et al. (2020). *Analysis of Malicious Email Detection using Cialdini's Principles*, in 2020 15th Asia Joint Conference on Information Security (AsiaJCIS), Taipei, Taiwan, pp. 137-142. https://doi.org/10.1109/AsiaJCIS50894.2020.00032

.

Pham, H. L., & Sun, J. (2024). Certified continual learning for neural network regression. *In Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2024)* (pp. 806–818). Association for Computing Machinery. https://doi.org/10.1145/3650212.3680322

Rabbi, M., A. Champa, and M. Zibran (2023). Phishy? Detecting Phishing Emails Using ML and NLP, in 2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA), Orlando, FL, USA, pp. 77-83. https://www2.cose.isu.edu/~minhazzibran/resources/MyPapers/Phishing_SERA23_Published.pdf

Redondo-Gutierrez, L., Francisco Jáñez-Martino, Eduardo Fidalgo, Enrique Alegre, Víctor González-Castro, and Rocío Alaiz-Rodríguez. (2022). Detecting malware using text documents extracted from spam email through machine learning. In Proceedings of the 22nd ACM Symposium on Document Engineering (DocEng '22). *Association for Computing Machinery*, New York, NY, USA, Article 17, 1–4. https://doi.org/10.1145/3558100.3563854

Roy, S. S., Naragam, K. V., & Nilizadeh, S. (2023, May 9). *Generating Phishing Attacks using ChatGPT*. arXiv.org. https://arxiv.org/abs/2305.05133

Sachan, S., Doulani, K., Adhikari, M. (2023). Semantic Analysis and Classification of Emails through Informative Selection of Features and Ensemble AI Model. In Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing (IC3-2023). *Association for Computing Machinery*, New York, NY, USA, 181–187. https://doi.org/10.1145/3607947.3607979

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. https://doi.org/10.1016/j.jksuci.2023.01.004

Saka, T., Kami Vaniea, and Nadin Kökciyan (2022). Context-Based Clustering to Mitigate Phishing Attacks. In Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security (AISec'22). *Association for Computing Machinery*, New York, NY, USA, 115–126. https://doi.org/10.1145/3560830.3563728

Saleem, B. M. (2021). *T*he P-Fryer: Using Machine Learning and Classification to Effectively Detect Phishing Emails (Order No. 28646335). Available from ProQuest Dissertations & Theses Global. (2572551978).

Sameen, M., Han, K., & Hwang, S. O. (2020). PhishHaven—An efficient Real-Time AI Phishing URLs detection System. *IEEE Access*, 8, 83425–83443. https://doi.org/10.1109/access.2020.2991403

Scofield, D., Miles, C., Kuhn, S. (2020). Automated Model Learning for Accurate Detection of Malicious Digital Documents. Digital Threats 1, 3, Article 15 (September 2020), 21 pages. https://doi.org/10.1145/3379505

Siddiq, A., Arifuzzaman, M., Islam, M. S. (2022). Phishing Website Detection using Deep Learning. In Proceedings of the 2nd International Conference on Computing Advancements (ICCA '22). *Association for Computing Machinery*, New York, NY, USA, 83–88. https://doi.org/10.1145/3542954.3542967

Chakraborty, S. (2023). Phishing Email Detection [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/6090437

Tiruthani, N., S. Nargundkar and W. Yu. (2009). PhishCatch - A Phishing Detection Tool, in 33rd Annual IEEE International Computer Software and Applications Conference (COMPSAC 2009), Seattle, WA, 2009 pp. 451-456.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention is all you need*. arXiv.org. https://arxiv.org/abs/1706.03762

Valecha, R., Mandaokar, P. and Rao, H. (2022). Phishing Email Detection Using Persuasion Cues, in *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 02, pp. 747-756. https://doi.ieeecomputersociety.org/10.1109/TDSC.2021.3118931

Verizon. (2023). *2023 Data Breach Investigations Report*. https://www.verizon.com/business/en-nl/resources/reports/dbir/2023/

Xiao, D., & Jiang, M. (2020, August). Malicious mail filtering and tracing system based on KNN and improved LSTM algorithm. In *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)* (pp. 222-229). IEEE.