Preprints.org

Article

Not peer-reviewed version

Investigation of Phishing Susceptibility with Explainable Artificial Intelligence

Zhengyang Fan *, Wanru Li, Kathryn Blackmond Laskey, Kuo-Chu Chang

Posted Date: 24 November 2023

doi: 10.20944/preprints202311.1540.v1

Keywords: Phishing Susceptibility; Cyber Security; Interpretable Artificial Intelligence; Machine Learning

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Investigation of Phishing Susceptibility with Explainable Artificial Intelligence

Zhengyang Fan *, Wanru Li, Kathryn Blackmond Laskey and Kuo-Chu Chang

Department of Systems Engineering and Operations Research, George Mason University; wli15@gmu.edu (W.L.); klaskey@gmu.edu (K.B.L.); kchang@gmu.edu (K.C.)

* Correspondence: zfan3@gmu.edu

Abstract: As artificial intelligence continues to advance, researchers are increasingly using machine learning algorithms to study the factors that make people more susceptible to phishing scams. Most studies in this area have taken one of two approaches: either they explore statistical associations between various factors and susceptibility, or they use complex models such as deep neural networks to predict phishing behavior. However, these approaches have limitations in terms of providing practical insights for individuals to avoid future phishing attacks and delivering personalized explanations regarding their susceptibility to phishing. In this paper, we propose a machine learning approach that leverages explainable artificial intelligence techniques to examine the influence of human and demographic factors on susceptibility to phishing attacks. Our analysis reveals that psychological factors such as impulsivity and conscientiousness, as well as appropriate online security habits, significantly affect an individual's susceptibility to phishing attacks. Furthermore, our individualized case-by-case approach offers personalized recommendations on mitigating the risk of falling prey to phishing exploits, considering the specific circumstances of each individual.

Keywords: phishing susceptibility; cyber security; Interpretable Artificial Intelligence; machine learning

1. Introduction

A phishing attack is a form of identity theft wherein a malicious website mimics a genuine one, with the intention of illicitly obtaining sensitive information like passwords, account details, or credit card numbers [1]. These exploits have caused significant losses to both corporations and government organizations. For example, the infamous cyber/phishing attack against the US Government Office of Personnel Management resulted in attackers gaining access to sensitive data on millions of government employees and contractors. In 2018, there was a 40% increase in phishing attacks targeted at US organizations [2]. According to the FBI Internet Crime Complaint Center, there were more than 2018 complaints resulting in losses of over 1.2 billion due to business email compromises. Based on the most recent FBI Internet Crime Annual Report, the incidence of phishing attacks has surged to its highest level since 2019, resulting in a significantly larger number of victims compared to personal data breaches, which ranked second in terms of victim count in 2022. Additionally, the financial losses associated with internet crimes, including phishing, reached a staggering 10.3 billion in 2022, nearly doubling the financial impact observed in 2021.

Conducting research to understand the factors contributing to an individual's susceptibility to phishing attacks is crucial in enhancing cybersecurity awareness and developing effective protective measures. However, some existing literature in this area has mainly focused on building models to achieve accurate predictions for phishing behavior. Although these models may improve performance, their interpretability can be challenging, making it difficult to guide researchers in developing targeted educational and awareness campaigns to prevent and mitigate the impact of phishing attacks [3,4]. Other works have relied on statistical tests to identify factors related to phishing susceptibility, but these approaches may not provide clear guidance on how the identified factors influence susceptibility [5–8]. To address these issues, this paper proposes a deep neural

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

network (DNN) approach powered by a local explainable artificial intelligence (XAI) technique called SHAP [9]. The proposed framework aims to provide not only accurate predictions on phishing susceptibility but also explanations of why an individual fell victim to a phishing attempt. In addition, it offers personalized recommendations on how to effectively minimize the risk of potential future phishing attacks.

The contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first attempt at employing XAI techniques to analyze susceptibility to phishing attacks. Our study aims to investigate various human factors associated with susceptibility to phishing attacks and to support decision-making through local interpretations.
- To the best of our knowledge, our study is the first of its kind in offering personalized recommendations aimed at mitigating the risk of potential future phishing attacks. These recommendations are based on local explanations tailored to each individual's unique circumstances.

The rest of the paper is organized as follows: Section 2 reviews the relevant literature; Section 3 describes the dataset used, which includes features, samples, and labels in our experiments; Section 4 introduces the deep learning and SHAP framework utilized in this study together with experimental results and analysis. Section 5 concludes the paper.

2. Related Literature

This section is structured as follows: In Section 2.1, we review the findings of previous phishing studies that have examined variables relevant to the ones used in our current study; Section 2.2 reviews the relevant applications of machine learning and XAI methodologies in phishing-related research.

2.1. Factors Related to Phishing Susceptibility

Parrish et al. [10] proposed a framework that aims to comprehend the potential effects of personality, experience, and demographic factors on phishing susceptibility. Based on this categorization, our study will examine literature pertaining to demographic, psychosocial, and experiential factors that may be associated with phishing susceptibility. Previous studies investigating the relationship between gender and phishing susceptibility have yielded inconsistent results. Some studies found that women were more likely to fall for phishing emails [6,11–13]. However, other studies found no significant gender differences or even suggested that males may be more susceptible in certain situations [5,14]. A study by Li et al. [2] similarly found no significant gender differences in falling for multiple phishing emails among university faculty and staff.

Studies examining the association between age and phishing susceptibility have produced inconsistent results as well. Most studies found that younger people (ages 18 to 25) were more likely to click on phishing emails than older age groups [12,15,16], while others found that the highest age group (over 59) was most susceptible [2]. In contrast, studies found no significant age differences in phishing susceptibility among university students and faculty [14,17]. Methodological differences between studies, such as the age ranges used and the type of phishing attack used, may account for the discrepant findings. Additionally, Li et al. [2] explored several additional factors that could potentially contribute to the age-related effects observed in previous studies.

Recent studies suggest that individuals who have fallen victim to phishing attacks in the past are more susceptible to future phishing attempts [18]. This raises the question of whether certain psychosocial factors contribute to phishing susceptibility. Psychosocial factors pertain to psychological aspects, such as personality traits or interpersonal behaviors, that can potentially influence a person's vulnerability to phishing attacks. Personality traits are stable and intrinsic characteristics of an individual's personality, such as Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism in the Big Five personality traits framework [19]. These traits can impact individuals' decision-making processes, interaction with others, and responses to various job-related pressures and uncertainties, all of which can be exploited by social engineering tactics.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

In a study of 200 university students, Alseadoon et al. [20] found that those with higher levels of Openness were more susceptible to social engineering. Similar results were reported in Halevi et al. [6]. Although these studies had small sample sizes and the latter study used a more stringent criterion. Workman [21] found a positive association between phishing susceptibility and Agreeableness, with higher levels of normative commitment, trust, and obedience to authority being linked to a greater likelihood of falling for social engineering attacks. Finally, Halevi et al. [6] study on Facebook privacy settings and phishing susceptibility found a positive correlation between Neuroticism and susceptibility to phishing emails. We refer interested readers to two recent excellent surveys by Desolda et al. [22] and Zhuo et al. [23], and the references therein for more detailed review on human factors that related to phishing susceptibility.

2.2. ML and XAI in Phishing Study

There are several research works that have applied machine learning techniques for phishing studies. Abbasi et al. [24] utilized cluster analysis and an elaborate controlled experiment involving a large number of participants to identify and analyze user segments with high susceptibility to phishing, based on their demographics, perceptions, and behavior on phishing websites. Yang et al. [25] developed a model for predicting phishing victims by proposing a multidimensional phishing susceptibility prediction model. They used seven supervised learning techniques to forecast the vulnerability of the enlisted volunteers, based on their demographic, personality, knowledge experience, and security behavior, all of which were obtained through a questionnaire. Yang et al. [26] presented a user phishing susceptibility prediction model that incorporates both dynamic and static features. The model examines the impact of static factors, such as demographics, knowledge, and experience, as well as dynamic factors, such as design changes and eye tracking, on user susceptibility. To predict susceptibility accurately, a hybrid prediction model that combines Long Short-Term Memory (LSTM) and LightGBM was developed, resulting in a prediction accuracy of 92.34 percent. Rahman et al. [27] proposed a conditional generative adversarial network (C-GAN) model for both classification and data generation to find the potential associations between personality traits and phishing attacks. Cranford et al. [28] proposed a new approach that integrates cognitive modeling and machine learning to enhance training effectiveness. To select appropriate targets for intervention during the training process, they utilized a restless multi-armed bandit framework and incorporated a cognitive model of phishing susceptibility to inform the bandit model's parameters.

Other studies use machine learning techniques to detect phishing webpages or emails. They apply different approaches to extract phishing classification information from diverse sources, such as visual information like logos [29–32], textual information like URLs [33–42] and webpage content [43,44]. As the present paper focuses on human factors associated with phishing susceptibility, readers interested in applying machine learning for phishing detection are directed to surveys by Divakaran and Oest [45] and Singh et al. [46] for more details.

To our knowledge, only a limited number of studies have employed XAI techniques to investigate the phenomenon of phishing. Hernandes et al. [47] proposed an XAI approach for phishing detection using URL-based features. The authors used machine learning models along with various XAI techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and explainable boosting machine (EBM) to identify the most important URL features contributing to the model's prediction. Their results show that the most important URL features identified by the XAI techniques are consistent with common phishing characteristics. Chai et al. [48] proposed a multi-modal hierarchical attention model for developing meaningful phishing detection systems. The model includes two levels of attention mechanisms to enable the extraction of relevant features and informative interpretability across multiple levels. Lin et al. [49] proposed a hybrid deep learning-based approach called Phishpedia to visually identify phishing webpages with explainable visual annotations on the phishing page screenshot. Recently, Kluge and Eckhardt [50] proposed a user-focused anti-phishing measure that leverages XAI to improve users' understanding of the cues that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

contribute to the suspicion of phishing, and uncover the words and phrases in an e-mail that most relevant for identifying phishing attempts.

3. Materials and Methods

In this section, we will initially present the dataset employed in the present study in subsection 3.1. Subsequently, in subsection 3.2, a brief introduction to methodologies associated with deep neural networks will be provided. In subsection 3.3, a concise overview of the SHAP technique, an XAI approach facilitating a thorough analysis of the internal mechanisms of our model, will be presented.

3.1. Data

The data utilized in this paper originates from an experimental study performed by a research team that includes several authors of this paper. To enhance the comprehensiveness of the present study, an overview of the design of the phishing campaigns and survey questionnaire is provided in the Appendix. For more in-depth details of the simulated phishing experiment, readers are referred to the works of Li et al. [2] and Greitzer et al. [51]. The primary objective of the current study is to identify the specific characteristics associated with susceptibility to phishing attacks.

The research team, affiliated with George Mason University (GMU), conducted an extensive experimental study involving 6,938 participants consisting of GMU faculty and staff members. Their data collection process encompassed three key components: a pre-campaign survey, the actual phishing campaign, and a post-campaign survey. In our analysis, we specifically focused on the human factors associated with phishing susceptibility. Therefore, our study only relied on demographic data and pre-campaign survey data for our analysis. A comprehensive overview of the collected data, along with their corresponding descriptions, is provided in Table 1.

Table 1. Data Collected in the Experimental Study.

Data Type	Description
Demographic Data	Age, gender, position, and department type. Collected from HR records
Behavioral and Psychological Data	Personality (impulsivity, conscientiousness, emotional stability, agreeableness, perceived stress) and technical/cybersecurity related experience. Collected from pre-campaign survey

Due to the limited number of participants who completed the pre-campaign survey (504), our results and subsequent analysis in the following sections are based solely on these 504 samples, rather than the total number of targeted individuals (6,938).

3.1.1. Demographic Data

The study utilized human resources records to determine demographic factors such as age, gender, position, and department. Age groups were categorized in such a way that no individual's identity could be discerned through their demographic details. Positions were categorized as full-time faculty, adjunct faculty, wages staff, and other staff. The department type was grouped into administration, technical college (science and engineering-related fields), and other college (inclusive of non-administrative employees). Table 2 below summarizes the name, value type and descriptions for the demographic variables.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

Table 2. Variable Name, Value Type and Descriptions for the Demographic Data.

Variable Name	Value Type	Description
Age	Categorical	5 values: [19,27), [27,41), [41,49), [49,59), [59+)
Gender	Categorical	2 values: Female, Male
Department	Categorical	3 values: Technical college, Administrative, Other College
Position	Categorical	4 values: Full-time faculty, adjunct faculty, wage staff, other staff

3.1.2. Pre-campaign Survey Data

The data for the pre-campaign survey was collected to analyze behavioral, psychological, and personality factors, along with technical and cybersecurity-related experience. The survey consisted of three parts, with the first part consisting of 32 questions to assess personality traits and related psychological aspects, and the other two sections assessing technical knowledge and previous experience with phishing exploits. The questions in the survey's first section used a 1.0 to 5.0 scale, while those in the second and third sections used a mixed 1.0 to 5.0 scale and binary scale. To keep the survey's length manageable, a validated psychological/personality inventory test was used to test five psychological state/trait items of impulsivity, conscientiousness, emotional stability, agreeableness, and perceived stress in a highly condensed form [2,51]. Table 3 summarizes the variable name, value type and descriptions for the pre-campaign survey data.

Table 3. Variable Name, Value Type and Descriptions for the Pre-campaign Survey Data.

Variable Name	Value Type	Description
Impulsivity (impul)	Numeric	Averaged over question 1–10 of section 1. Range from 1.0 to 5.0 to measure impulsivity score
Conscientiousness (consc)	Numeric	Averaged over question 11–14 of section 1. Range from 1.0 to 5.0 to measure conscientiousness score
Emotional Stability (emo)	Numeric	Averaged over question 15–18 of section 1. Range from 1.0 to 5.0 to measure the emotional stability score
Agreeableness (agree)	Numeric	Averaged over question 19–22 of section 1. Range from 1.0 to 5.0 to measure the agreeableness score
Perceived Stress (stress)	Numeric	Averaged over question 23–32 of section 2. Range from 1.0 to 5.0 to measure the perceived stress score
Check Link (checklink)	Numeric	Response to the corresponding survey question in section 3. Range from 1.0 (never) to 5.0 (very often)
Privacy Setting (privacysetting)	Numeric	Response to the corresponding survey question in section 3. Range from 1.0 (never) to 5.0 (very often)
Check HTTPS (checkhttps)	Numeric	Response to the corresponding survey question in section 3. Range from 1.0 (never) to 5.0 (very often)
Click w/o Check (clickwocheck)	Numeric	Response to the corresponding survey question in section 3. Range from 1.0 (never) to 5.0 (very often)
Phished Before (phishbefore)	Binary	Response to the corresponding survey question in section 3. Binary valued: Yes = 1, No = 0
Phished in Last 3 Months (phishlast3mon)	Binary	Response to the corresponding survey question in section 3. Binary valued: Yes = 1, No = 0
Lose Info Due to Phishing (loseinfo)	Binary	Response to the corresponding survey question in section 3. Binary valued: Yes = 1, No = 0
Download Malware (downmalware)	Binary	Response to the corresponding survey question in section 3. Binary valued: Yes = 1, No = 0

Overall, a total of 17 factors/features are used in the subsequent analysis (4 from demographic data and 13 from pre-campaign survey data). Among 504 individuals who participated in the pre-campaign survey, 121 of them clicked the simulated phishing emails.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

3.2. Deep Neural Networks

Deep neural networks, also known as deep learning models, are a class of neural networks that consist of multiple layers of interconnected artificial neurons. These networks are designed to learn hierarchical representations of data by progressively extracting more abstract and complex features as information flows through the layers.

Unlike shallow neural networks with only one or a few hidden layers, deep neural networks have a greater depth, typically comprising multiple hidden layers. Each layer consists of a set of neurons that perform computations on the input data and pass the results to the next layer. The output of one layer serves as the input to the subsequent layer, enabling the network to learn increasingly sophisticated representations.

The depth of deep neural networks allows them to capture intricate patterns and relationships in data, making them particularly effective in handling large and complex datasets. Deep learning models have shown remarkable success in various domains, including computer vision, natural language processing, speech recognition, and reinforcement learning.

One of the key advantages of deep neural networks is their ability to automatically learn relevant features from raw data, eliminating the need for manual feature engineering. By leveraging the hierarchical nature of deep architectures, these models can extract high-level representations that capture meaningful information from the input data. This automated feature learning capability has contributed to the exceptional performance of deep learning models in many tasks, such as image classification, object detection, and language translation.

We trained a multi-layer perceptron (MLP) neural network to make predictions. The neural network architecture used in current study consists of four hidden layers. The first two layers have 256 neurons each and were followed by a rectified linear unit (ReLU) activation function. The third layer has 128 neurons and is followed by a hyperbolic tangent (Tanh) activation function. The fourth layer has 32 neurons with ReLU activation. The output layer is dense with a sigmoid activation function. Figure 1 below depicts the network architecture.

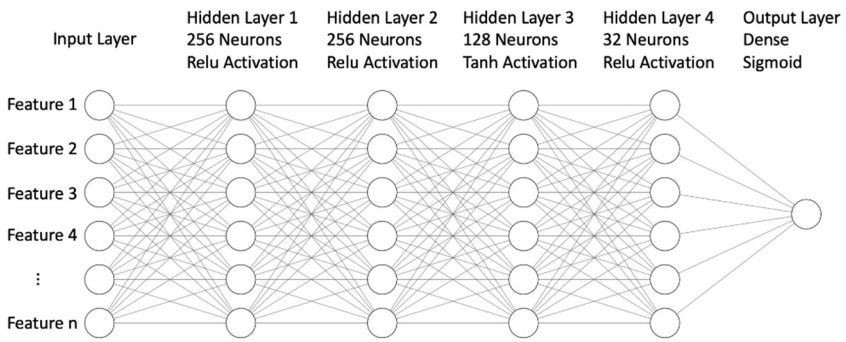

Figure 1. Network Architecture Used for the Neural Network Model for Prediction.

The purpose of using an MLP neural network is to learn a non-linear mapping between the input data and the target variable. The MLP neural network is a feedforward neural network where information flows from input to output in a unidirectional manner. The ReLU and Tanh activation functions were chosen due to their effectiveness in improving the performance of the neural network in handling non-linearity and avoiding the vanishing gradient problem. The sigmoid activation function in the output layer was used to output a probability score between 0 and 1, representing the likelihood of a user clicking on a phishing link. Overall, the network architecture was designed to be deep and intricate, enabling it to capture the complex relationships between the input features and the target variable.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

3.3. SHAP

SHAP (SHapley Additive exPlanations) is a model agnostic XAI method that provides a way to explain the predictions made by machine learning models. It is based on the concept of Shapley values, which is a method for assigning a contributing value to each player in a cooperative game. In the context of machine learning, each feature in a dataset can be considered as a 'player' in the game, and SHAP computes the contribution of each feature to the final prediction. Let $X = \{X_1, X_2, \dots, X_M\}$ be a set of M features and $f(\mathcal{X})$ be the model that needs to be explained. According to the Shapley value, given a sample point $x^* = \{x_1^*, x_2^*, \dots, x_M^*\}$, the amount that player/feature j contributes at sample x is

$$\phi_j(v) = \phi_j = \sum_{S \subseteq M - \{j\}} \frac{|S|!\,(|M| - |S| - 1)!}{|M|!} \big(v(S \cup \{j\}) - v(S)\big) \tag{1}$$

In the equation, $|\cdot|$ denotes the cardinality for a set. And $v(\cdot)$ is the value function for a subset of features that defined as conditional expectations of target function $f(\cdot)$:

$$v(S) = \mathbb{E}(f(X)|x_S = x_S^*) - \mathbb{E}\big(f(X)\big) \tag{2}$$

SHAP works by estimating the conditional expectation of the model output given the value of each feature, and then calculating the difference between the expected output and the actual output. This difference is referred to as the 'contribution' of the feature to the prediction. SHAP values are computed by averaging the contributions over all possible orderings of the features. In other words, SHAP provides a way to assign a value to each feature that reflects its contribution to the final prediction.

4. Results and Discussion

Within this section, we will initially outline the performance outcomes achieved by our deep learning model. Subsequently, we will employ the SHAP XAI method to conduct a comprehensive analysis of the data, enabling us to offer personalized training suggestions and provide valuable insights into the research domain.

4.1. Deep Learning Predictor

To prepare the data for the neural network model, the dataset was initially divided into training and testing sets in a 4:1 ratio, with 80% of the data allocated to training and 20% to testing. Due to the highly imbalanced nature of the data, where the number of non-click users was considerably higher than the number of click users in the training set, a technique called NearMiss was applied. NearMiss is an undersampling method that uses a K-nearest neighbors algorithm to reduce the number of majority class instances by selecting the samples closest to the minority class [52]. By using this technique, the imbalance in the training set was addressed, and a more balanced training set was obtained.

In addition, the dataset contains categorical and binary features such as gender and position, which cannot be used directly as input for the neural network model due to their nominal nature. To address this challenge, we employ a technique known as one-hot encoding to transform these non-ordinal features into numerical values that can be effectively utilized by the model. One-hot encoding represents each category as a binary vector with a length equal to the number of categories in the feature. Each category in the feature is then represented by a unique binary vector, with a value of 1 in the position corresponding to that category and 0s in all other positions. This allows the neural network to learn the relationship between each category and the target variable by treating each category as a separate feature with a numerical value.

The performance of the trained neural network predictor was evaluated using the testing set, and the receiver operating characteristic (ROC) curve is displayed in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

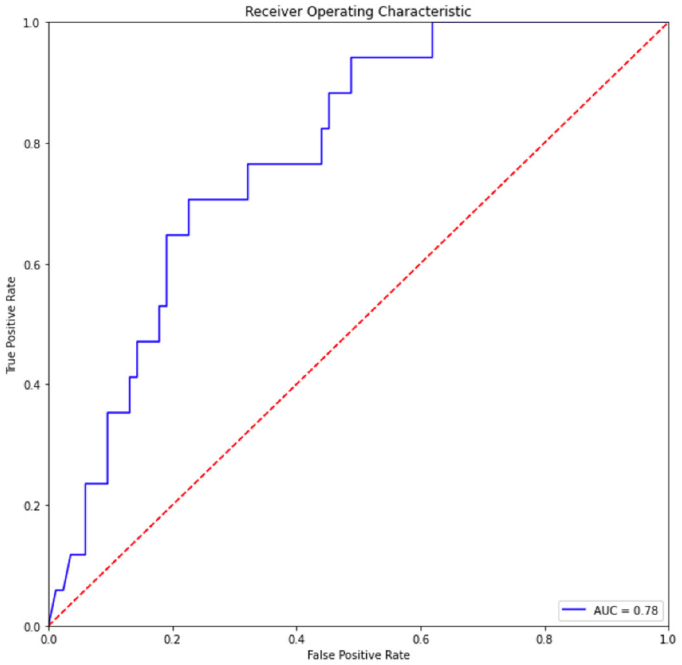

Figure 2. ROC Curve for the Learned Deep Learning Model. The aera under the ROC curve (AUC) is 0.78.

The model yielded an accuracy of 0.78, with a recall of 0.71, a precision of 0.57 and a F-1 score of 0.64. In contrast, Greitzer et al.'s study [51] on the same population employed a linear logistic regression model achieving 0.71 accuracy, a recall of 0.17 and a precision of 0.11 on the IT dataset. These models are not directly comparable because Greitzer, et al. [51] reported predictive accuracy only for a model that did not make use of any behavioral factors. These results indicate that our model exhibits reasonable predictive capabilities and can be valuable in identifying potential click risks in real-world scenarios. Table 4 summarizes the model performance results.

Table 4. Model Performance.

Evaluation Metric	Value
Accuracy	0.78
Precision	0.57
Recall	0.71
F-1 Score	0.64

4.2. SHAP Explanation

Figure 3 below illustrates the impact of different factors on phishing susceptibility as revealed by the SHAP XAI method.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

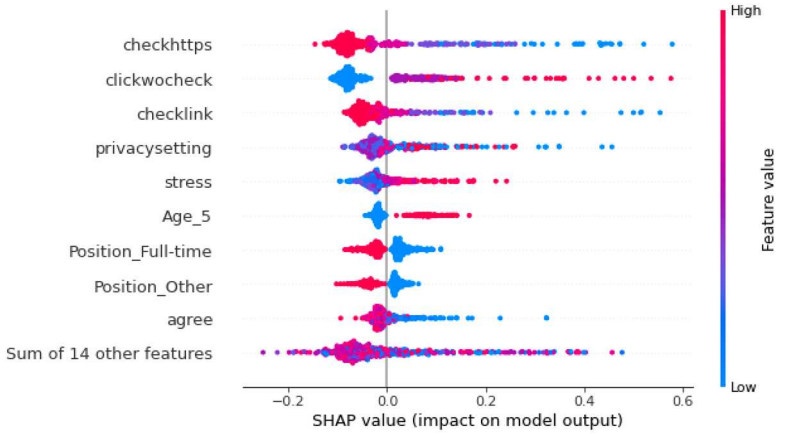

Figure 3. SHAP Values for Features: In this plot, each data point represents an observation in the dataset. Features are displayed along the *y*-axis, and the *x*-axis represents the corresponding SHAP values. The color bar on the right serves as a reference, with red indicating high feature values and blue indicating low values. For instance, focusing on the "checkhttps" feature, instances with higher "checkhttps" values (depicted in red) exhibit negative SHAP values, ranging between -0.2 and 0. This observation aligns seamlessly with our intuitive understanding.

The behavioral habits-related factors have the largest impact on phishing behavior. Interestingly, the checkhttps factor has the largest variability in terms of Shapley values. It is observed that individuals who seldom check for secure websites are assigned large positive Shapley values, which increases their probability of being phished. Conversely, high values of checkhttps may have a moderate impact on reducing the probability of being phished. On the other hand, the clickwocheck factor has exactly the opposite impact. Higher values of clickwocheck correspond to individuals who always click email links without checking their legitimacy and are assigned large positive Shapley values, which increases their probability of being phished. Lower values of chickwocheck lead to a decrease in the probability of being phished. Similar interpretations apply to other behavior-related factors such as checklink and privacysetting.

It is noteworthy that these results are inconsistent with a previous study based on the same data set [51], which did not identify any behavior-related factors having a significant impact on phishing susceptibility. This is possibly due to the fact that the previous study relied on stepwise logistic regression, which is a linear method and cannot capture nonlinear relationships between behavioral factors and phishing susceptibility. In contrast to the previous study, the current study used a deep neural network model to predict phishing susceptibility and then applied SHAP as a post-hoc method to identify important features. This approach is able to model the complicated nonlinear relationship between behavioral factors and phishing susceptibility. To provide evidence for our hypothesis regarding the nonlinear relationship between phishing susceptibility and behavioral factors, we then present some partial dependence plots for these features.

Partial dependence plots are a type of visualization tool used in the SHAP method of model interpretation. It shows the marginal effect one or two features have on the predicted outcome of a machine learning model by fixing the values of all other features and varying the values of one or two features of interest. Thus, partial dependence plots provide an estimate of how the predicted outcome of a model changes as a function of one or two features and enables us to explore and visualize the relationships between input features and model predictions and can help identify non-linear relationships between features and target variables that may not be apparent in simple scatterplots or correlation matrices.

Figure 4 depicts partial dependence plots for four variables: checkhttps (a), checklink (b), clickwocheck (c), and stress (d), respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

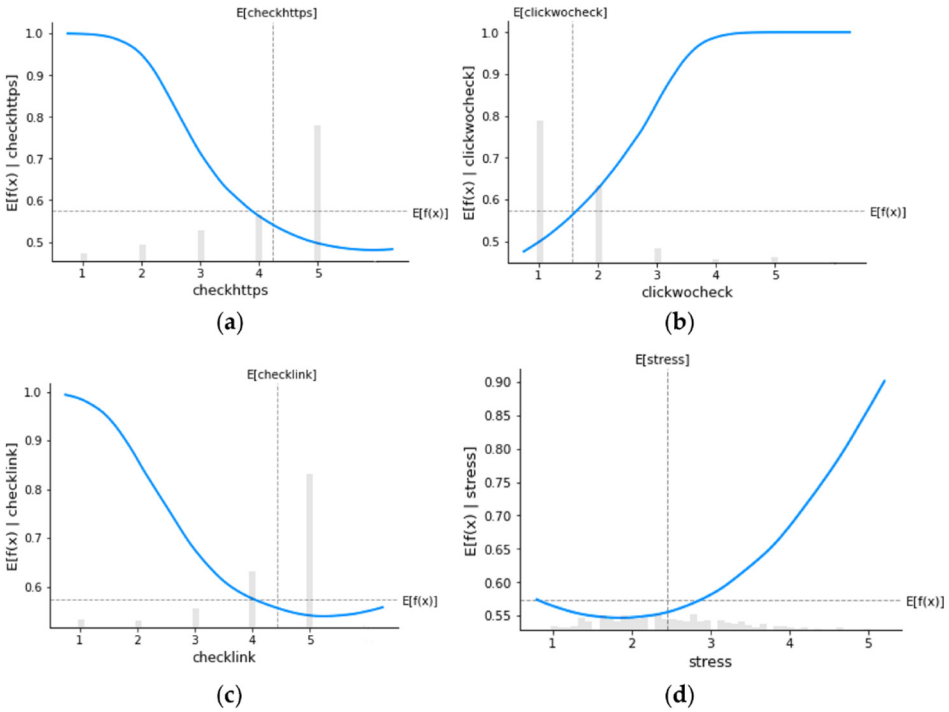

Figure 4. Partial dependence plots: (**a**) partial dependence plot for checkhttps; (**b**) partial dependence plot for clickwocheck; (**c**) partial dependence plot for checklink; (**d**) partial dependence plot for stress.

The above plots illustrate the relationship between each variable and the probability of clicking on a phishing email. The horizontal dashed line represents the expected click probability, while the vertical dashed line represents the expected value of each variable. Additionally, the shaded bar in the background shows the histogram of each feature. Figure 4a shows a non-linear and non-increasing relationship between checkhttps and the likelihood of clicking on a phishing email. As the value of checkhttps increases, the probability of clicking decreases. Similarly, Figure 4b indicates a non-decreasing relationship between clickwocheck and the probability of being phished. Interestingly, when clickwocheck is at 4, the probability of clicking is already near 1, suggesting that individuals who never check the legitimacy of emails (clickwocheck = 5) need to cultivate the habit of checking the legitimacy of emails to at least 3 (often check the legitimacy of the email) to decrease their risk of being phished. Figure 4c,d also demonstrate similar relationships between checklink and stress with the probability of clicking on a phishing email. These plots also provide insight into why our dataset is highly imbalanced, as shown in the histograms of Figure 4a to Figure 4c. The majority of individuals in our dataset exhibit good online habits and "security hygiene" making them less susceptible to phishing attacks.

In the next step of our analysis, we use SHAP to provide local explanations for individual instances using waterfall plots. This approach allows us to provide personalized suggestions for people to reduce their risk of being phished based on their specific features and characteristics. The waterfall plot provides a detailed explanation for the prediction of a single instance by showing how each feature contributes to the final prediction. It displays a horizontal bar for each feature, with the length representing the impact of that feature on the prediction, and the color indicating the direction of the impact (positive or negative). The waterfall plot is useful for identifying the most important features that contribute to a particular prediction and understanding how changes in those features can affect the prediction. Figures 5–7 depict waterfall plots for three individuals who clicked on phishing emails during our experiments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

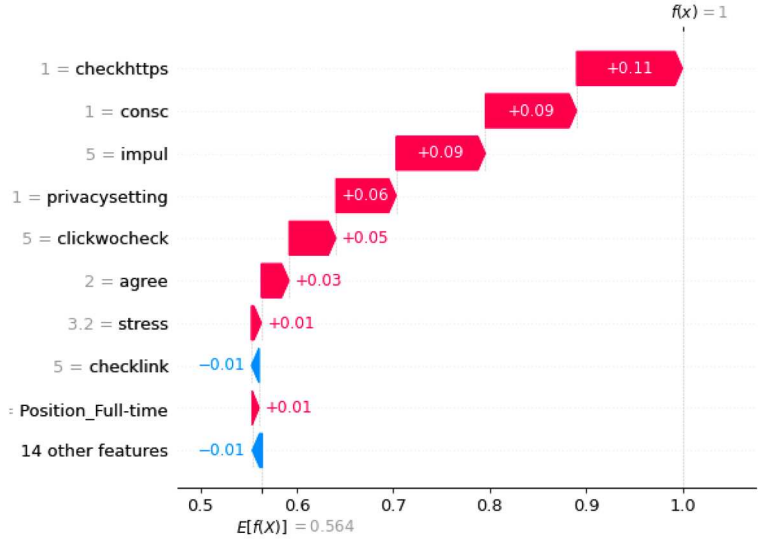

Figure 5. Waterfall plot for individual No.1.

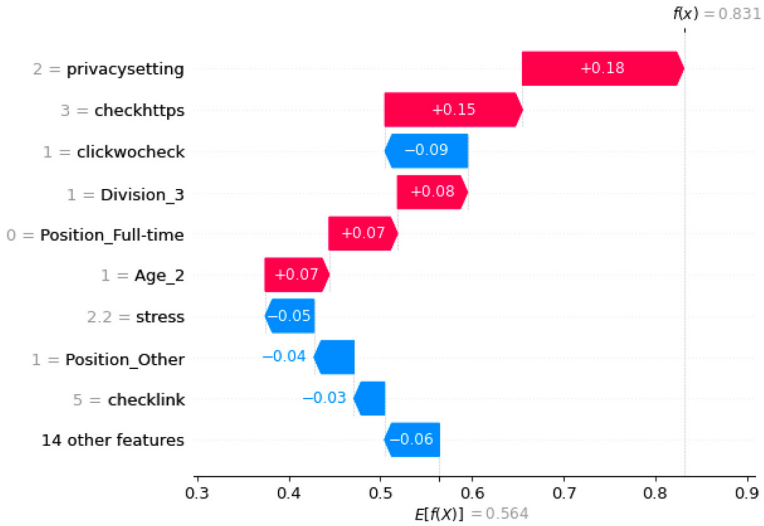

Figure 6. Waterfall plot for individual No.2.

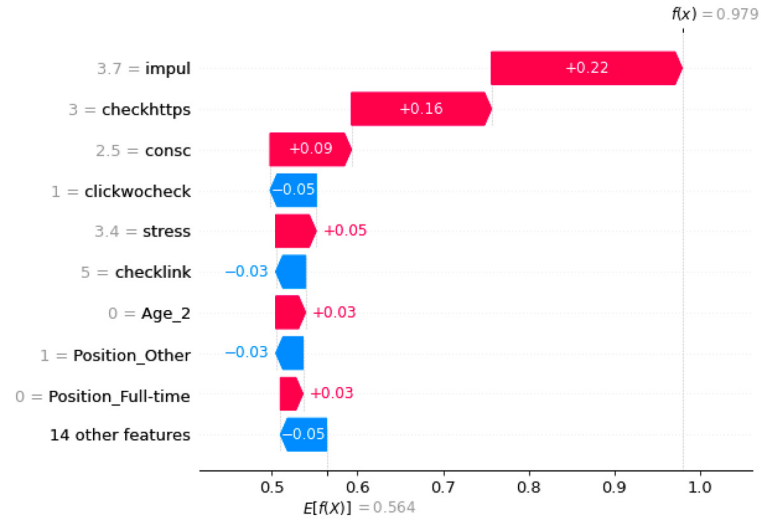

Figure 7. Waterfall plot for individual No.3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

In Figure 5, this individual was predicted to be a victim of phishing with a probability of 1.00, which is much higher than the average probability of 0.564 for people to be phished. Note that the calculated average click probability, which is 56.4%, is obtained by taking the expectation with respect to features X using a down-sampled training dataset. The down-sampling technique, specifically the NearMiss method, is employed to address the issue of imbalanced data by reducing the number of non-clickers. This average click probability is directly related to the trained prediction model $f(X)$. It is important to note that this value differs from the marginal click rate of 20% (121 clickers over 504 total data points), which is calculated based on the response variable Y (i.e., $\mathbb{E}(Y)$.

Two behavioral factors, checkhttps and privacysetting, as well as two psychological factors, consc and impul, contribute most to this positive prediction. These results can be used to create a personalized anti-phishing training program for this individual by focusing on improving their security behavior related to checking https and privacy settings. By combining these results with Figure 8a, which shows a scatter plot of checkhttps with its corresponding Shapley values, we can expect the probability of this individual being phished to drop to 0.79 if they develop a habit of frequently checking the https of email links (i.e., checkhttps = 5).

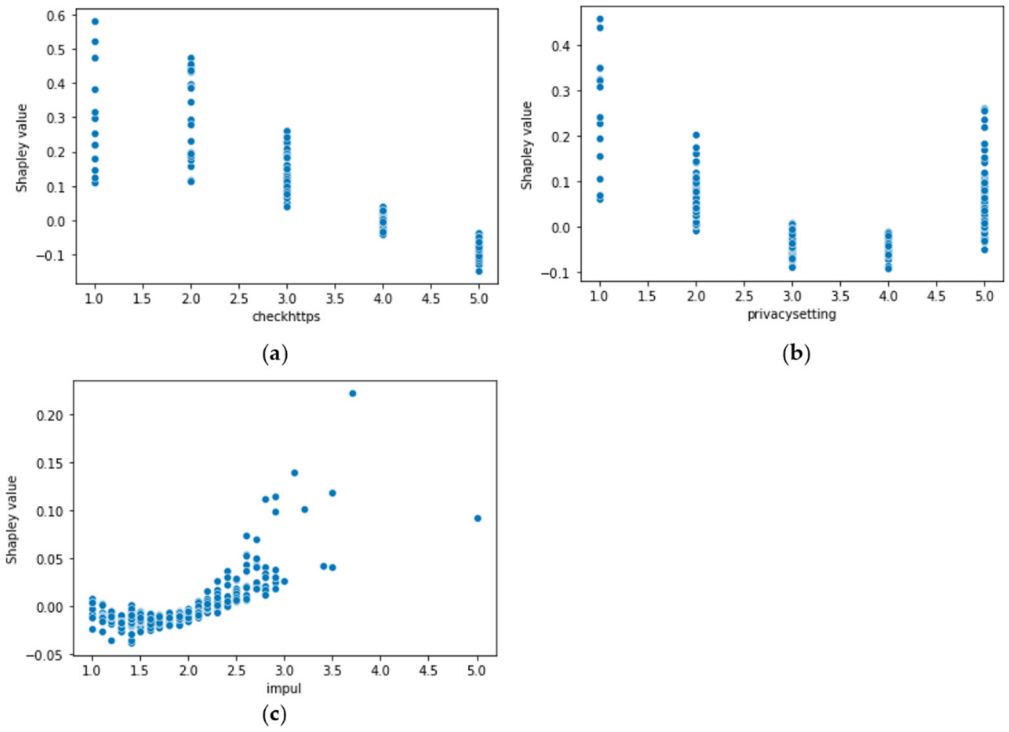

Figure 8. Scatter plots: (**a**) scatter plot for check https; (**b**) scatter plot for privacy setting; (**c**) scatter plot for impulsivity.

In Figure 6, we see a waterfall plot for another individual who has been classified as a victim of phishing with a probability of 0.831. Unlike the previous example, privacysetting is the feature that contributed most to this positive prediction, with a Shapley value of 0.18. Therefore, a personalized anti-phishing training program can be created for this individual to improve their privacy setting skills. By combining the scatter plot of privacysetting (Figure 8b), we can estimate that the probability of this individual being phished would decrease to around 0.58 if their privacysetting score is improved to 3 or 4.

The third individual's waterfall plot is displayed in Figure 7, and this individual has been classified as a clicker with a predicted phishing probability of 0.979. The highest contributing factor to this prediction is impul, a psychological factor, with a Shapley value of 0.22. In addition, checkhttps also contributes significantly to the prediction with a Shapley value of 0.16, which is higher compared to the previous two instances. Therefore, to reduce the risk of this individual being phished in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

future, the personalized anti-phishing training program should focus on reducing impulsive behavior and encouraging habitual checking of https. By examining the scatter plots for impul (Figure 8c) and checkhttps (Figure 8a), it can be observed that by reducing the individual's impul score to 1 and increasing the checkhttps score to 5 after training, the probability of being phished is expected to decrease to approximately 0.54. This estimation is based on the expected Shapley value for impul = 1 being around 0 and the expected Shapley value for checkhttps = 5 being around -0.05.

5. Conclusions

In this paper, we present a machine learning approach that utilizes SHAP, an XAI technique, to investigate the influence of human and demographic factors on susceptibility to phishing attacks. Our study reveals that, on a global level, security hygiene habits exhibit the most significant influence on individuals' susceptibility to phishing. Among these habits, checkhttps, clickwocheck, checklink, and privacysetting are identified as the top four factors that significantly impact phishing susceptibility. On the other hand, at the local/individual level, individuals often possess their own unique set of factors that contribute to their susceptibility to phishing attacks. Therefore, based on the local Shapley value analysis, our approach proposes personalized recommendations for each individual to mitigate their susceptibility to phishing scams based on their unique circumstances. For example, our study shows that impulsivity is the most influential factor contributing to the susceptibility of one particular individual to phishing attacks. A personalized training program for this individual would therefore focus on impulsivity. In general, personalized recommendations aim to address the specific factors that render each person vulnerable to phishing attacks.

The analysis conducted in this study relies exclusively on demographic information and pre-campaign survey responses. Despite achieving an accuracy of 78% with our deep learning model, the true positive rate, representing the probability of correctly identifying individuals susceptible to phishing, is approximately 75%. To improve this, we plan to incorporate post-survey data into our analysis in future work. This inclusion will allow us to capture any shifts in participants' behavior and attitudes towards phishing after the campaign, and potentially improve the accuracy of our model. Additionally, we aim to build a natural language processing model to analyze the open-ended responses in the post-survey data. This approach will yield more detailed and nuanced insights into participants' experiences and perceptions of the phishing campaign. By doing so, we can obtain valuable information to guide the development of more effective anti-phishing interventions.

Author Contributions: Conceptualization, Z.F., W.L. and K.C.; formal analysis, Z.F. and W.L.; investigation, Z.F. and W.L.; Methodology, Z.F., W.L., K.B.L. and K.C.; supervision: K.B.L. and K.C.; writing-original draft, Z.F. and W.L.; writing-reviewing and editing, K.B.L. and K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

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

Appendix A. Design of Phishing Campaign

In this appendix, we include supplementary details and data essential described in Li et al. [2] for a comprehensive understanding and reproducibility of the research presented in the main text. The design of three phishing emails within distinct contexts, each incorporating urgency cues to prompt user interaction, is outlined as follows:

- IT Helpdesk (IT): An IT helpdesk email informs the user of suspicious overnight activity leading to account deactivation, with instructions to click a link for activity review and account reactivation; Sender domain: "support@masonhelpdesk.com"
- Package Delivery (PD): A package delivery service email communicates a failed delivery due to an invalid postal code, urging the user to click on the link to download a shipping label necessary for picking up the package; Sender domain: "pkginfo@vapostal.com"

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

- Credit Charge Warning (CC): An email alerts users to a suspicious credit card charge, emphasizing large purchase notifications and prompting the user to click a link for charge review and notification settings adjustment. Sender domain: "service@acubank.co"

To mitigate potential suspicion and account for confounding factors such as day of the week and order of receipt, we established nine user groups (A1, A2, A3, B1, B2, B3, C1, C2, C3) using stratified sampling, ensuring similar age, gender, and departmental composition within each group. Each group received one email per week on a different day, resulting in the receipt of all three emails over a three-week period, see Table A1.

Table A1. Phishing Campaign Schedule [2].

Week	Email to send	Tuesday	Wednesday	Thursday
Week 1	Email #1	A1	B1	C1
	Email #2	A2	B2	C2
	Email #3	A3	B3	C3
Week 2	Email #1	C3	A3	B3
	Email #2	C1	A1	B1
	Email #3	C2	A2	B2
Week 3	Email #1	B2	C2	A2
	Email #2	B3	C3	A3
	Email #3	B1	C1	A1

Users were given a minimum of one week opportunity to click on each email; for those emails given a longer window, almost all clicks were observed during the first week. Click data, including operating system and time of clicks, were recorded to correlate click behavior with IT data, identifying technical indicators of susceptibility to phishing.

For users clicking on phishing links, the Landing Page (LP) to which they were redirected varied, aiming to assess the impact of feedback on subsequent behavior. Three LPs were randomly assigned:

- Standard 404 error message (least informative feedback).
- Webpage displaying a simple message about the simulated phishing link.
- Webpage providing detailed information about the phishing link, explaining the study, and featuring a training video on identifying suspicious emails (most educational feedback).

To address the impact of LP, two hypotheses were formulated:

- Users notified of clicking on a phishing link would be less likely to click on a future link.
- Users receiving stronger notification (i.e., training video LP) would be less likely to click on a future link than those receiving a simple message notification.

To ensure an adequate sample size for the first hypothesis, the probability distribution for LPs was set at 25% for a simple message, 25% for a training video, and 50% for a standard 404.

Appendix B. Pre-Campaign Survey

The pre-campaign survey included three sections addressing the following topics:

- Section 1: Personality measures
- Section 2: Perceived stress
- Section 3: Technical/cyber security related experience

Appendix B.1. Personality Measures

Directions: The following statements describe different sorts of attributes or behaviors. Please read the statement carefully and select the response that best describes you are you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself. Your responses to this survey will be kept in absolute confidence.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

Indicate for each statement whether it is 1: Very Inaccurate; 2: Moderately Inaccurate; 3: Neither Accurate nor Inaccurate; 4: Moderately Accurate; 5: Very Accurate as a description of you.

Table B1. Pre-Campaign Survey: Personality Measures [51]. Each question has six choices: 1: Very Inaccurate; 2: Moderately Inaccurate; 3: Neither Accurate nor Inaccurate; 4: Moderately Accurate; 5: Very Accurate; 6: Decline to Answer.

Survey Questions
1. Act without thinking
2. Jump into things without thinking
3. Stick to my chosen path
4. Often make last-minute decisions
5. Make rash decisions
6. Like do crazy things
7. Like to avoid mistakes
8. Choose my words with care
9. Like to act on a whim
10. Ruch into things
11. Make a mess of things
12. Often forget to put things back in their proper place
13. Get chores done right away
14. Like order others
15. Get upset easily
16. Seldom fell blue
17. Have frequent mood swings
18. Am relaxed most of the time
19. Am not interested in other people's problems
20. Am not really interested in others
21. Feel others' emotions
22. Sympathize with others' feelings

General instructions: This survey asks you to indicate how various attributes and behaviors may apply to you. Please respond honestly, with confidence that your answers will be confidential. If you so choose, there is an option to decline to answer.

Appendix B.2. Perceived Stress

Directions: Indicate for each statement whether it is 1: Very Infrequent; 2: Moderately Infrequent; 3: Neither Frequent nor Infrequent; 4: Moderately Frequent; 5: Very Frequent as a description of your experience.

Table B2. Pre-Campaign Survey: Perceived Stress [51]. Each question has six choices: 1: Very Infrequent; 2: Moderately Infrequent; 3: Neither Frequent nor Infrequent; 4: Moderately Frequent; 5: Very Frequent; 6: Decline to Answer.

Survey Questions
1. In the last month, how often have you been upset because of something that happened unexpectedly?
2. In the last month, how often have you felt that you were unable to control the important things in your life?
3. In the last month, how often have you felt nervous and stressed?
4. In the last month, how often have you felt confident about your ability to handle your personal problems?
5. In the last month, how often have you felt that things were going your way?
6. In the last month, how often have you found that you could not cope with all the things that you had to do?
7. In the last month, how often have you been able to control irritations in your life?
8. In the last month, how often have you felt that you were on top of things?
9. In the last month, how often have you been angered because of things that were outside of your control?
10. In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

General instructions: This survey asks you to indicate how various attributes and behaviors may apply to you. Please respond honestly, with confidence that your answers will be confidential. If you so choose, there is an option to decline to answer.

Appendix B.3. Technical/Cyber Security Related Experience

Directions: For each statement indicate whether you do this: 1: Never; 2: Seldom; 3: Sometimes; 4: Often; 5: Very Often; 6: Not Applicable; 7: Decline to Answer.

Table B3. Pre-Campaign Survey: Technical/Cyber Security Related Experience (Part 1) [51]. Each question has seven choices: 1: Never; 2: Seldom; 3: Sometimes; 4: Often; 5: Very Often; 6: Not Applicable; 7: Decline to Answer.

Survey Questions
1. I check every email link before clicking
2. I review my privacy settings on social media sites
3. I check for secure website (https) before entering my personal information
4. I click links in email messages without ensuring that they are legitimate

Directions: Social engineering attacks occur in phishing email messages or other types of social media in which the attacker attempts to fool the user into clicking on a link that exposes the victim to malicious software or viruses.

Table B4. Pre-Campaign Survey: Technical/Cyber Security Related Experience (Part 2) [51]. Each question has three choices: 1: Yes; 2: No; 3: Decline to Answer.

Survey Questions
1. Have you ever clicked on a link that turned out to be a phishing attempt?
2. In the past 3 months, have you clicked on a link that turned out to be a phishing attack?
3. If 'Yes' to either of the previous questions, did you lose information as a result?
4. If 'Yes' to either of the previous questions, did you unintentionally download malware as a result?

References

1. Greitzer, F.L.; Strozer, J.R.; Cohen, S.; Moore, A.P.; Mundie, D.; Cowley, J. Analysis of Unintentional Insider Threats Deriving from Social Engineering Exploits. In Proceedings of the 2014 IEEE Security and Privacy Workshops; May 2014; pp. 236–250.
2. Li, W.; Lee, J.; Purl, J.; Greitzer, F.; Yousefi, B.; Laskey, K. *Experimental Investigation of Demographic Factors Related to Phishing Susceptibility*; 2020; ISBN 978-0-9981331-3-3.
3. Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **2019**, *40*, 44–58, doi:10.1609/aimag.v40i2.2850.
4. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115, doi:10.1016/j.inffus.2019.12.012.
5. Diaz, A.; Sherman, A.T.; Joshi, A. Phishing in an Academic Community: A Study of User Susceptibility and Behavior. *Cryptologia* **2020**, *44*, 53–67, doi:10.1080/01611194.2019.1623343.
6. Halevi, T.; Lewis, J.; Memon, N. Phishing, Personality Traits and Facebook 2013.
7. Pethers, B.; Bello, A. Role of Attention and Design Cues for Influencing Cyber-Sextortion Using Social Engineering and Phishing Attacks. *Future Internet* **2023**, *15*, 29, doi:10.3390/fi15010029.
8. Qi, Q.; Wang, Z.; Xu, Y.; Fang, Y.; Wang, C. Enhancing Phishing Email Detection through Ensemble Learning and Undersampling. *Appl. Sci.* **2023**, *13*, 8756, doi:10.3390/app13158756.
9. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., 2017; Vol. 30.
10. James, P.J.; Bailey, J.; Courtney, J. A Personality Based Model for Determining Susceptibility to Phishing Attacks. *Little Rock Univ. Ark.* **2009**, 285–296.
11. Jagatic, T.N.; Johnson, N.A.; Jakobsson, M.; Menczer, F. Social Phishing. *Commun. ACM* **2007**, *50*, 94–100, doi:10.1145/1290958.1290968.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

12. Sheng, S.; Holbrook, M.; Kumaraguru, P.; Cranor, L.F.; Downs, J. Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Association for Computing Machinery: New York, NY, USA, April 10 2010; pp. 373–382.
13. Blythe, M.; Petrie, H.; Clark, J.A. F for Fake: Four Studies on How We Fall for Phish. In Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Association for Computing Machinery: New York, NY, USA, May 7 2011; pp. 3469–3478.
14. Mohebzada, J.G.; Zarka, A.E.; Bhojani, A.H.; Darwish, A. Phishing in a University Community: Two Large Scale Phishing Experiments. In Proceedings of the 2012 International Conference on Innovations in Information Technology (IIT); March 2012; pp. 249–254.
15. Lin, T.; Capecci, D.E.; Ellis, D.M.; Rocha, H.A.; Dommaraju, S.; Oliveira, D.S.; Ebner, N.C. Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content. *ACM Trans. Comput.-Hum. Interact.* **2019**, *26*, 32:1-32:28, doi:10.1145/3336141.
16. Parsons, K.; Butavicius, M.; Delfabbro, P.; Lillie, M. Predicting Susceptibility to Social Influence in Phishing Emails. *Int. J. Hum.-Comput. Stud.* **2019**, *128*, 17–26, doi:10.1016/j.ijhcs.2019.02.007.
17. Downs, J.S.; Holbrook, M.B.; Cranor, L.F. Decision Strategies and Susceptibility to Phishing. In Proceedings of the Proceedings of the second symposium on Usable privacy and security; Association for Computing Machinery: New York, NY, USA, July 12 2006; pp. 79–90.
18. Canham, M.; Posey, C.; Strickland, D.; Constantino, M. Phishing for Long Tails: Examining Organizational Repeat Clickers and Protective Stewards. *SAGE Open* **2021**, *11*, 2158244021990656, doi:10.1177/2158244021990656.
19. Digman, J.M. Personality Structure: Emergence of the Five-Factor Model. *Annu. Rev. Psychol.* **1990**, *41*, 417–440, doi:10.1146/annurev.ps.41.020190.002221.
20. Alseadoon, I.; Chan, T.; Foo, E.; Nieto, J.G. Who Is More Susceptible to Phishing Emails?: A Saudi Arabian Study. *ACIS 2012 Proc.* **2012**.
21. Workman, M. Wisecrackers: A Theory-Grounded Investigation of Phishing and Pretext Social Engineering Threats to Information Security. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 662–674, doi:10.1002/asi.20779.
22. Desolda, G.; Ferro, L.S.; Marrella, A.; Catarci, T.; Costabile, M.F. Human Factors in Phishing Attacks: A Systematic Literature Review. *ACM Comput. Surv.* **2021**, *54*, 173:1-173:35, doi:10.1145/3469886.
23. Zhuo, S.; Biddle, R.; Koh, Y.S.; Lottridge, D.; Russello, G. SoK: Human-Centered Phishing Susceptibility. *ACM Trans. Priv. Secur.* **2023**, *26*, 24:1-24:27, doi:10.1145/3575797.
24. Abbasi, A.; Zahedi, F.M.; Chen, Y. Phishing Susceptibility: The Good, the Bad, and the Ugly. In Proceedings of the 2016 IEEE Conference on Intelligence and Security Informatics (ISI); September 2016; pp. 169–174.
25. Yang, R.; Zheng, K.; Wu, B.; Li, D.; Wang, Z.; Wang, X. Predicting User Susceptibility to Phishing Based on Multidimensional Features. *Comput. Intell. Neurosci.* **2022**, *2022*, e7058972, doi:10.1155/2022/7058972.
26. Yang, R.; Zheng, K.; Wu, B.; Wu, C.; Wang, X. Prediction of Phishing Susceptibility Based on a Combination of Static and Dynamic Features. *Math. Probl. Eng.* **2022**, *2022*, e2884769, doi:10.1155/2022/2884769.
27. Rahman, A.U.; Al-Obeidat, F.; Tubaishat, A.; Shah, B.; Anwar, S.; Halim, Z. Discovering the Correlation Between Phishing Susceptibility Causing Data Biases and Big Five Personality Traits Using C-GAN. *IEEE Trans. Comput. Soc. Syst.* **2022**, 1–9, doi:10.1109/TCSS.2022.3201153.
28. Cranford, E.; Jabbari, S.; Ou, H.-C.; Tambe, M.; Gonzalez, C.; Lebiere, C. Combining Machine Learning and Cognitive Models for Adaptive Phishing Training.
29. Bozkir, A.S.; Aydos, M. LogoSENSE: A Companion HOG Based Logo Detection Scheme for Phishing Web Page and E-Mail Brand Recognition. *Comput. Secur.* **2020**, *95*, 101855, doi:10.1016/j.cose.2020.101855.
30. Chiew, K.L.; Chang, E.H.; Sze, S.N.; Tiong, W.K. Utilisation of Website Logo for Phishing Detection. *Comput. Secur.* **2015**, *54*, 16–26, doi:10.1016/j.cose.2015.07.006.
31. Chiew, K.L.; Choo, J.S.-F.; Sze, S.N.; Yong, K.S.C. Leverage Website Favicon to Detect Phishing Websites. *Secur. Commun. Netw.* **2018**, *2018*, e7251750, doi:10.1155/2018/7251750.
32. Panda, P.; Mishra, A.K.; Puthal, D. A Novel Logo Identification Technique for Logo-Based Phishing Detection in Cyber-Physical Systems. *Future Internet* **2022**, *14*, 241, doi:10.3390/fi14080241.
33. Liu, D.-J.; Geng, G.-G.; Zhang, X.-C. Multi-Scale Semantic Deep Fusion Models for Phishing Website Detection. *Expert Syst. Appl.* **2022**, *209*, 118305, doi:10.1016/j.eswa.2022.118305.
34. Yang, L.; Zhang, J.; Wang, X.; Li, Z.; Li, Z.; He, Y. An Improved ELM-Based and Data Preprocessing Integrated Approach for Phishing Detection Considering Comprehensive Features. *Expert Syst. Appl.* **2021**, *165*, 113863, doi:10.1016/j.eswa.2020.113863.
35. Sahingoz, O.K.; Buber, E.; Demir, O.; Diri, B. Machine Learning Based Phishing Detection from URLs. *Expert Syst. Appl.* **2019**, *117*, 345–357, doi:10.1016/j.eswa.2018.09.029.
36. Akinyelu, A.A.; Adewumi, A.O. Classification of Phishing Email Using Random Forest Machine Learning Technique. *J. Appl. Math.* **2014**, *2014*, e425731, doi:10.1155/2014/425731.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2023 doi:10.20944/preprints202311.1540.v1

37. AlEroud, A.; Karabatis, G. Bypassing Detection of URL-Based Phishing Attacks Using Generative Adversarial Deep Neural Networks. In Proceedings of the Proceedings of the Sixth International Workshop on Security and Privacy Analytics; Association for Computing Machinery: New York, NY, USA, March 16 2020; pp. 53–60.
38. Yerima, S.Y.; Alzaylaee, M.K. High Accuracy Phishing Detection Based on Convolutional Neural Networks. In Proceedings of the 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS); March 2020; pp. 1–6.
39. Fang, Y.; Zhang, C.; Huang, C.; Liu, L.; Yang, Y. Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism. *IEEE Access* **2019**, *7*, 56329–56340, doi:10.1109/ACCESS.2019.2913705.
40. Wang, Y.; Ma, W.; Xu, H.; Liu, Y.; Yin, P. A Lightweight Multi-View Learning Approach for Phishing Attack Detection Using Transformer with Mixture of Experts. *Appl. Sci.* **2023**, *13*, 7429, doi:10.3390/app13137429.
41. Roy, S.S.; Awad, A.I.; Amare, L.A.; Erkihun, M.T.; Anas, M. Multimodel Phishing URL Detection Using LSTM, Bidirectional LSTM, and GRU Models. *Future Internet* **2022**, *14*, 340, doi:10.3390/fi14110340.
42. Butnaru, A.; Mylonas, A.; Pitropakis, N. Towards Lightweight URL-Based Phishing Detection. *Future Internet* **2021**, *13*, 154, doi:10.3390/fi13060154.
43. Wen, T.; Xiao, Y.; Wang, A.; Wang, H. A Novel Hybrid Feature Fusion Model for Detecting Phishing Scam on Ethereum Using Deep Neural Network. *Expert Syst. Appl.* **2023**, *211*, 118463, doi:10.1016/j.eswa.2022.118463.
44. Alhogail, A.; Alsabih, A. Applying Machine Learning and Natural Language Processing to Detect Phishing Email. *Comput. Secur.* **2021**, *110*, 102414, doi:10.1016/j.cose.2021.102414.
45. Divakaran, D.M.; Oest, A. Phishing Detection Leveraging Machine Learning and Deep Learning: A Review 2022.
46. Singh, C.; Meenu Phishing Website Detection Based on Machine Learning: A Survey. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS); March 2020; pp. 398–404.
47. Galego Hernandes, P.R.; Floret, C.P.; Cardozo De Almeida, K.F.; Da Silva, V.C.; Papa, J.P.; Pontara Da Costa, K.A. Phishing Detection Using URL-Based XAI Techniques. In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI); December 2021; pp. 01–06.
48. Chai, Y.; Zhou, Y.; Li, W.; Jiang, Y. An Explainable Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence. *IEEE Trans. Dependable Secure Comput.* **2022**, *19*, 790–803, doi:10.1109/TDSC.2021.3119323.
49. Lin, Y.; Liu, R.; Divakaran, D.M.; Ng, J.Y.; Chan, Q.Z.; Lu, Y.; Si, Y.; Zhang, F.; Dong, J.S. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages.; 2021; pp. 3793–3810.
50. Kluge, K.; Eckhardt, R. Explaining the Suspicion: Design of an XAI-Based User-Focused Anti-Phishing Measure. In Proceedings of the Innovation Through Information Systems; Ahlemann, F., Schütte, R., Stieglitz, S., Eds.; Springer International Publishing: Cham, 2021; pp. 247–261.
51. Greitzer, F.L.; Li, W.; Laskey, K.B.; Lee, J.; Purl, J. Experimental Investigation of Technical and Human Factors Related to Phishing Susceptibility. *ACM Trans. Soc. Comput.* **2021**, *4*, 8:1-8:48, doi:10.1145/3461672.
52. Inderjeet, M.; Zhang, J. kNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. *Proc. Workshop Learn. Imbalanced Datasets* **2003**, *126*, 1–7.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.