future internet

MDPI

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, Fairfax, VA 22030, USA; wli15@gmu.edu (W.L.); klaskey@gmu.edu (K.B.L.); kchang@gmu.edu (K.-C.C.)
* Correspondence: zfan3@gmu.edu

Abstract: Phishing attacks represent a significant and growing threat in the digital world, affecting individuals and organizations globally. Understanding the various factors that influence susceptibility to phishing is essential for developing more effective strategies to combat this pervasive cybersecurity challenge. Machine learning has become a prevalent method in the study of phishing susceptibility. 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. The machine learning model yielded an accuracy of 78%, with a recall of 71%, and a precision of 57%. 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

Citation: Fan, Z.; Li, W.; Laskey, K.B.; Chang, K.-C. Investigation of Phishing Susceptibility with Explainable Artificial Intelligence. *Future Internet* **2024**, *16*, 31. https://doi.org/10.3390/fi16010031

Academic Editors: Christoph Stach, Clémentine Gritti and Iouliana Litou

Received: 28 November 2023
Revised: 28 December 2023
Accepted: 16 January 2024
Published: 17 January 2024

1. Introduction

A phishing attack is a form of identity theft wherein a malicious website mimics a genuine one, to illicitly obtain 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 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 present study builds upon two prior research studies [2,10] that focused on identifying the most important factors associated with susceptibility to phishing attacks at a broader scale. While these studies yielded valuable insights, their results lack the specificity needed to offer personalized guidance to individuals in reducing their vulnerability to phishing attempts. Notably, these results are derived at a population level and may not account for the unique circumstances of individuals. In contrast, our current research aims to pinpoint the key factors associated with susceptibility to phishing attacks at the individual level, leveraging an XAI method. Our primary goal is to offer personalized guidance to individuals, empowering them to proactively reduce their risk of succumbing to phishing attacks. In addition, we aim to complement and extend the findings of prior phishing studies through our approach. Furthermore, we seek to enhance the understanding of the diverse factors influencing phishing susceptibility, enabling the development of more tailored and effective educational and preventive measures. Another objective is to contribute to the broader discourse on cybersecurity awareness, particularly in the context of evolving phishing tactics.

The contributions of this paper are summarized as follows:

1. 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;
2. To the best of our knowledge, our study is the first of its kind to offer 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

Demographic, psychosocial, and experiential factors stand out as the three most crucial factors linked to phishing susceptibility [11]. Accordingly, we will review the current literature focusing on these factors to better understand their associations with susceptibility to phishing attacks.

Prior studies yield inconsistent findings regarding the correlation between demographic factors, such as gender and age, and phishing susceptibility. Some studies assert that women exhibit higher susceptibility to phishing emails [6,12–14], while others report

no significant gender differences or even suggest that males may be more susceptible in specific scenarios [2,5,15]. Similarly, inconsistencies in age-related findings were observed, with most studies indicating that younger individuals (ages 18 to 25) are more prone to clicking on phishing emails compared to older age groups [13,16,17]. However, conflicting findings also exist, including studies suggesting that the highest age group (over 59) is most susceptible [2] or reporting no significant age differences in phishing susceptibility among university students and faculty [15,18].

Psychosocial factors encompass psychological aspects, such as personality traits or interpersonal behaviors, that may impact an individual's vulnerability to phishing attacks [19]. Within the framework of the Big Five personality traits [20], Halevi et al. [6] and Alseadoon et al. [21] observed that individuals with higher levels of Openness were more susceptible to social engineering. Workman [22] identified a positive association between phishing susceptibility and Agreeableness, with higher levels of normative commitment, trust, and obedience to authority linked to a greater likelihood of falling for social engineering attacks. Halevi et al. [6], in their study on Facebook privacy settings and phishing susceptibility, found a positive correlation between Neuroticism and susceptibility to phishing emails. For a more in-depth review of human factors related to phishing susceptibility, interested readers are referred to two recent comprehensive surveys by Desolda et al. [23] and Zhuo et al. [24], along with the references therein.

2.2. ML and XAI in Phishing Study

Several research works have applied machine learning techniques for phishing studies. Abbasi et al. [25] 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. [26] 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. [27] 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. [28] 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. [29] 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 [30–33], textual information like URLs [34–43], and webpage content [44,45]. 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 [46] and Singh et al. [47] 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. [48] 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 im-

portant 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. [49] 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. [50] 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 [51] proposed a user-focused anti-phishing measure that leverages XAI to improve users' understanding of the cues that contribute to the suspicion of phishing and uncover the words and phrases in an e-mail that are most relevant for identifying phishing attempts. The summary of the ML and XAI-based phishing studies is depicted in Table 1.

Table 1. Summary of the ML and XAI-based phishing studies.

Authors	Year	Method	Description
Abbasi et al. [25]	2016	K-mean clustering	Using clustering method to identify user segments with high susceptibility, focusing on perceptions, demographics, and website traversal behavior.
Yang et al. [26]	2022	Logistic regression, boosting, and support vector machine	Using various supervised learning methods to identify the relation between demographics, personality, knowledge experience, security behavior, cognitive processes, and susceptibility.
Yang et al. [27]	2022	Hybrid LSTM and LightGBM	Combining static and dynamic features, together with a hybrid algorithm to predict phishing susceptibility.
Rahman et al. [28]	2022	C-GAN	Investigate the relationship between phishing susceptibility and personality traits using C-GAN
Cranford et al. [29]	2022	Multi-armed bandit	Using a restless multi-armed bandit framework to strategically target users for intervention in phishing email detection.
Bozkir and Aydos [30]	2020	Max margin object detection	Employing a histogram of oriented gradients and a max-margin object detector to localize and classify brand logos in phishing web page screenshots.
Chiew et al. [31]	2015	SVM	Utilizing SVM for logo extraction and Google image search for identity verification.
Panda et al. [33]	2022	Random forest	Developed a logo-based phishing detection mechanism using hue value distribution as a feature.
Liu et al. [34]	2022	CNN and RNN	Developed three multi-scale semantic deep fusion networks using URLs to identify phishing websites.
Yang et al. [35]	2021	Extreme learning machine	Proposed a non-inverse matrix extreme learning machine for phishing website detection, together with denoising autoencoder and adaptive synthetic sampling.
Sahingoz et al. [36]	2019	Random forest	Developed a real-time anti-phishing system using a random forest classifier with NLP-based features.
Akinyelu et al. [37]	2014	Random forest	Using random forest to classify phishing attacks with hand-crafted URL-based features.
AlEroud et al. [38]	2020	GAN	Utilizing a generative model to generate URL-based phishing examples that can deceive Blackbox detection models
Yerima et al. [39]	2020	CNN	Developed a 1D CNN-based detection model using website URLs as features.
Fang et al. [40]	2019	Recurrent CNN	Developed a phishing email detection model based on recurrent CNN with multilevel vectors and attention mechanisms.

Table 1. *Cont.*

Authors	Year	Method	Description
Wang et al. [41]	2023	Transformer	Designed a Transformer model with an expert-mixture mechanism for phishing website detection, utilizing website URLs, attributes, content, and behavioral information.
Roy et al. [42]	2022		Using LSTM, bidirectional LSTM, and gated recurrent unit to identify malicious URLs.
Butnaru et al. [43]	2021	Random forest	Using random forest for blocking phishing attacks using URLs.
Wen et al. [44]	2023	LSTM	Developed a hybrid detection model that integrates LSTM and a fully convolutional network to detect phishing scam accounts on the Ethereum blockchain.
Alhogail et al. [45]	2021	Graph convolutional network (GCN)	Using GCN and NLP over an email body text to identify phishing emails.
Hernandes et al. [48]	2021	LIME	Using LIME to detect phishing websites, aiming to provide insights into the categorization of phishing URLs
Chai et al. [49]	2022	Hierarchical attention	Developed a multi-modal hierarchical attention model for phishing website detection, jointly learning deep fraud cues from three modalities.
Lin et al. [50]	2021	Faster RCNN object detection	A hybrid system designed to address technical challenges in phishing identification by accurately recognizing identity logos on webpage screenshots and matching logo variants.

3. Materials and Methods

In this section, we will initially present the dataset employed in the present study in Section 3.1. Subsequently, in Section 3.2, a brief introduction to methodologies associated with deep neural networks will be provided. In Section 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 a simulated phishing experiment performed by a research team that includes several authors of this paper. The detailed design of the phishing campaigns is thoroughly outlined in the studies by Li et al. [2] and Greitzer et al. [10]. Additionally, the survey questionnaire we used is included in the appendices of Greitzer et al. [10]. 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 6938 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. It is important to note that all collected data was de-identified to safeguard personally identifiable information. 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 2.

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 (6938).

Table 2. 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 the pre-campaign survey

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, wage staff, and other staff. The department type was grouped into administration, technical college (science and engineering-related fields), and another college (inclusive of non-administrative employees). Table 3 below summarizes the name, value type, and descriptions for the demographic variables.

Table 3. 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 [2,10]. Table 4 summarizes the variable name, value type, and descriptions for the pre-campaign survey data.

Table 4. 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 questions 11–14 of Section 1. Range from 1.0 to 5.0 to measure conscientiousness score
Emotional Stability (emo)	Numeric	Averaged over questions 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 questions 23–32 of Section 2. Range from 1.0 to 5.0 to measure the perceived stress score

Table 4. *Cont.*

Variable Name	Value Type	Description
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.

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.

Deep learning models, in comparison to traditional models in phishing susceptibility studies, offer a more complex and potentially higher-performing alternative. The model can capture nuanced patterns in phishing tactics that simpler models might miss. The strength of the model lies in automatic feature detection and creation, allowing them to adapt effectively to varied data types. This adaptability is particularly beneficial in the evolving landscape of phishing, where new threats constantly emerge. While they do require substantial computational resources, the potential for higher accuracy and better generalization to new, unseen data makes deep neural networks a powerful tool in the arsenal against phishing attacks, especially in scenarios where dataset size and diversity are sufficient.

We trained a multi-layer perceptron (MLP) neural network to make predictions. The neural network architecture used in the 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.

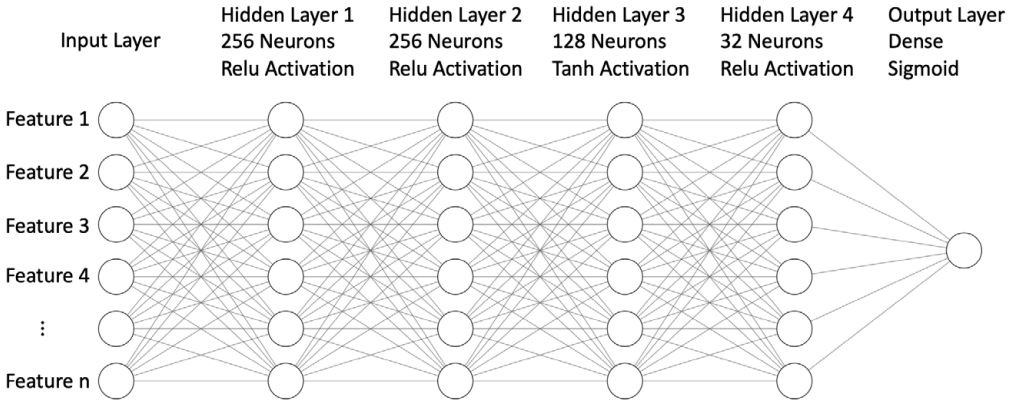

Figure 1. Network Architecture Used for the Neural Network Model for Prediction.

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.

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. The method is applied to a trained 'black box' model to unveil the contributions of each feature to individual predictions, offering valuable insights into the reasons for the model's decisions. 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 $X = \{X_1, X_2, \ldots, X_M\}$ be a set of M features and $f(X)$ be the model that needs to be explained. According to the Shapley value, given a sample point $x^* = \{x_1^*, x_2^*, \ldots, x_M^*\}$, the amount that player/feature j contributes at sample x is

$$\phi_j(\nu) = \phi_j = \sum_{S \subseteq M - \{j\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} (\nu(S \cup \{j\}) - \nu(S)) \quad (1)$$

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

$$\nu(S) = \mathbb{E}(f(X)|x_S = x_S^*) - \mathbb{E}(f(X)) \quad (2)$$

The resulting Shapley value ϕ_j adheres to the property of efficiency, expressed as $f(X) = \sum_{j=1}^{M} \phi_j$. This property signifies that the Shapley value ϕ_j represents the contribution of the j^{th} feature, with all interactions between this feature and others being averaged out.

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 0 s 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.

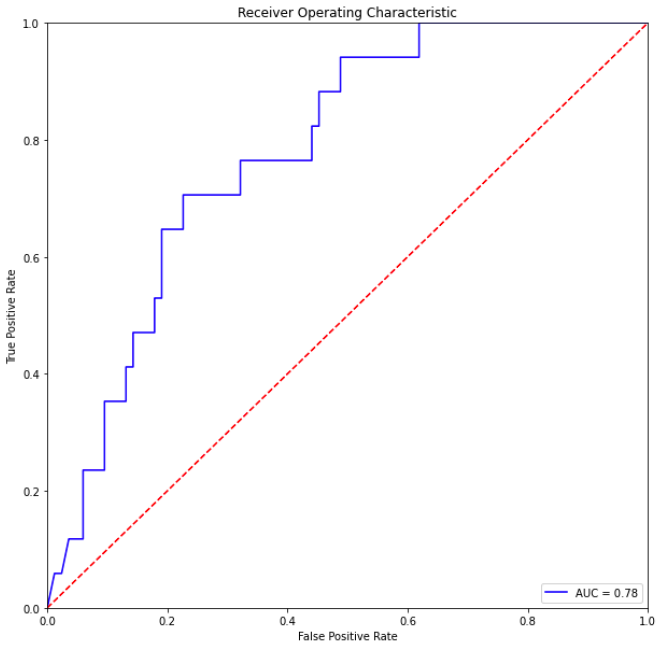

Figure 2. ROC Curve for the Learned Deep Learning Model. The aera under the ROC curve (AUC) is 0.78.

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. [10] 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 5 summarizes the model performance results.

Table 5. Model Performance.

Evaluation Metric	Value
Accuracy	0.78
Precision	0.57
Recall	0.71
F-1 Score	0.64
True Positive Rate	0.75
True Negative Rate	0.79

4.2. SHAP Explanation

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

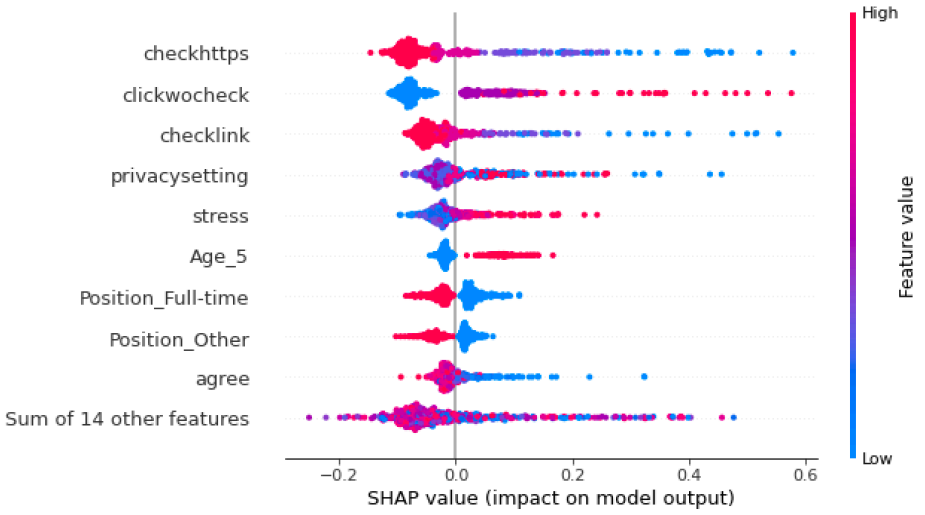

Figure 3. SHAP Values for Features: In this plot, each data point represents an observation in the dataset. Features are displayed along the *y*-axis, and the *x*-axis represents the corresponding SHAP values. The color bar on the right serves as a reference, with red indicating high feature values and 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. These findings align with established behavioral theories in cybersecurity, suggesting that individuals with greater technical knowledge and security awareness are less vulnerable [53,54].

It is noteworthy that these results are inconsistent with a previous study based on the same data set [10], 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.

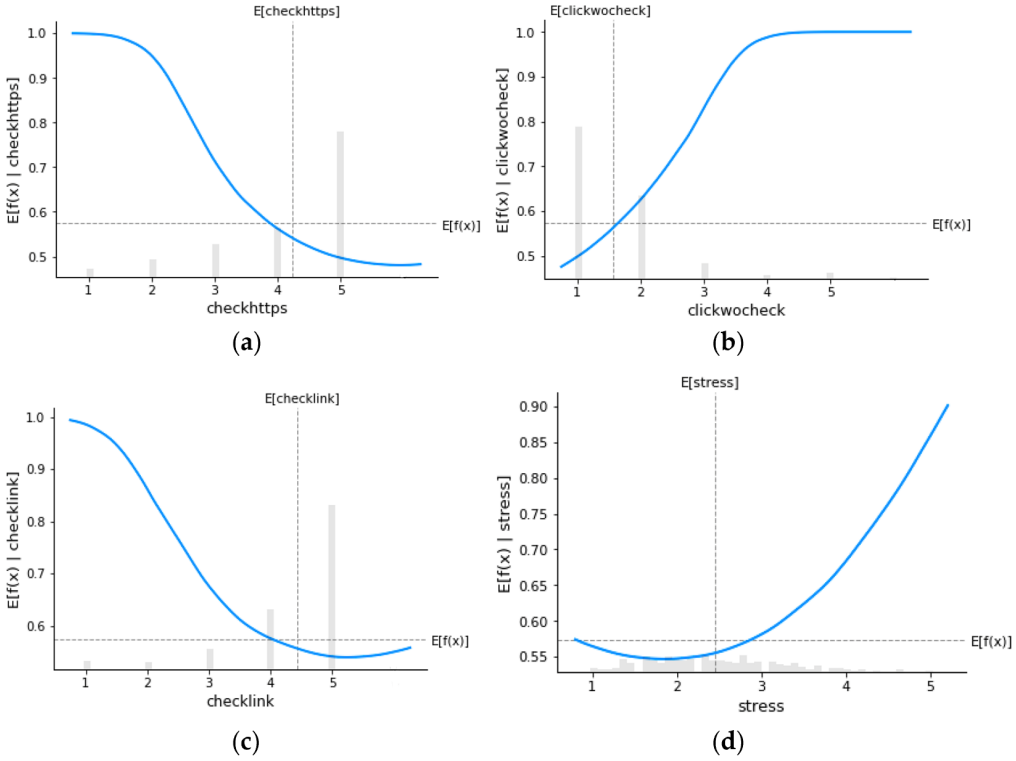

Figure 4. Partial dependence plots: (**a**) partial dependence plot for checkhttps; (**b**) partial dependence plot for clickwocheck; (**c**) partial dependence plot for checklink; (**d**) partial dependence plot for stress.

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–c. 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.

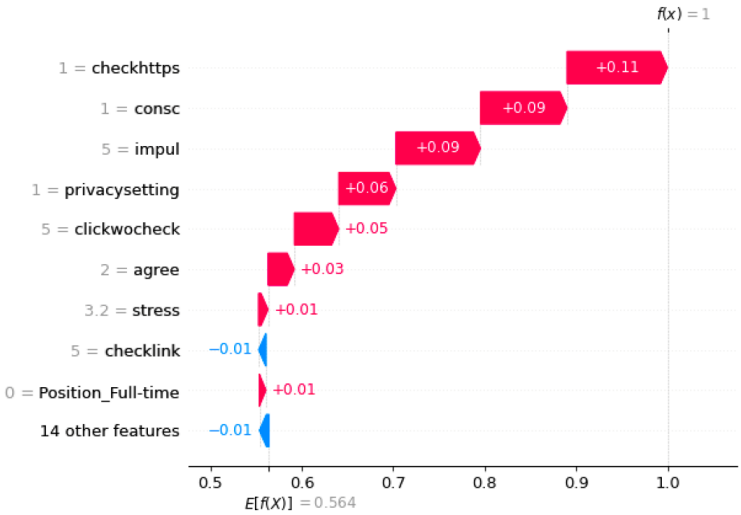

Figure 5. Waterfall plot for individual No. 1.

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).

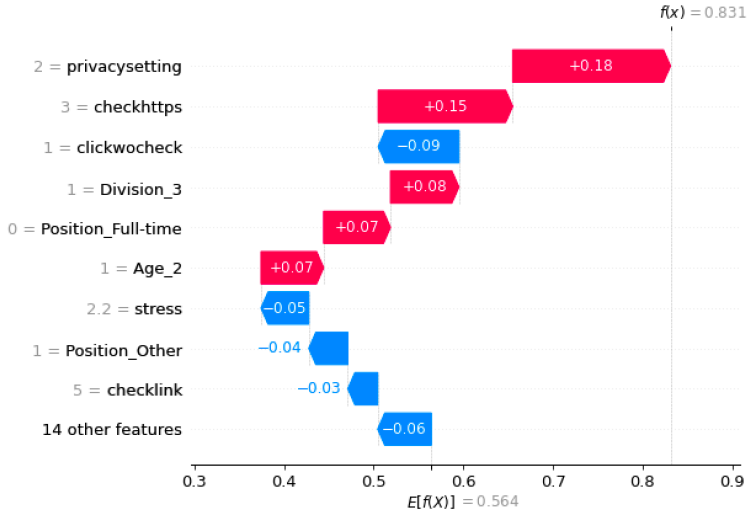

Figure 6. Waterfall plot for individual No. 2.

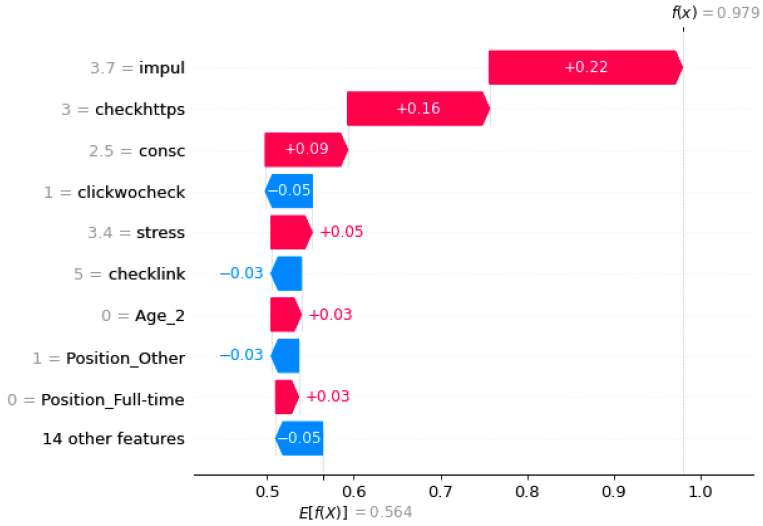

Figure 7. Waterfall plot for individual No. 3.

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, checkouts also contribute 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 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.

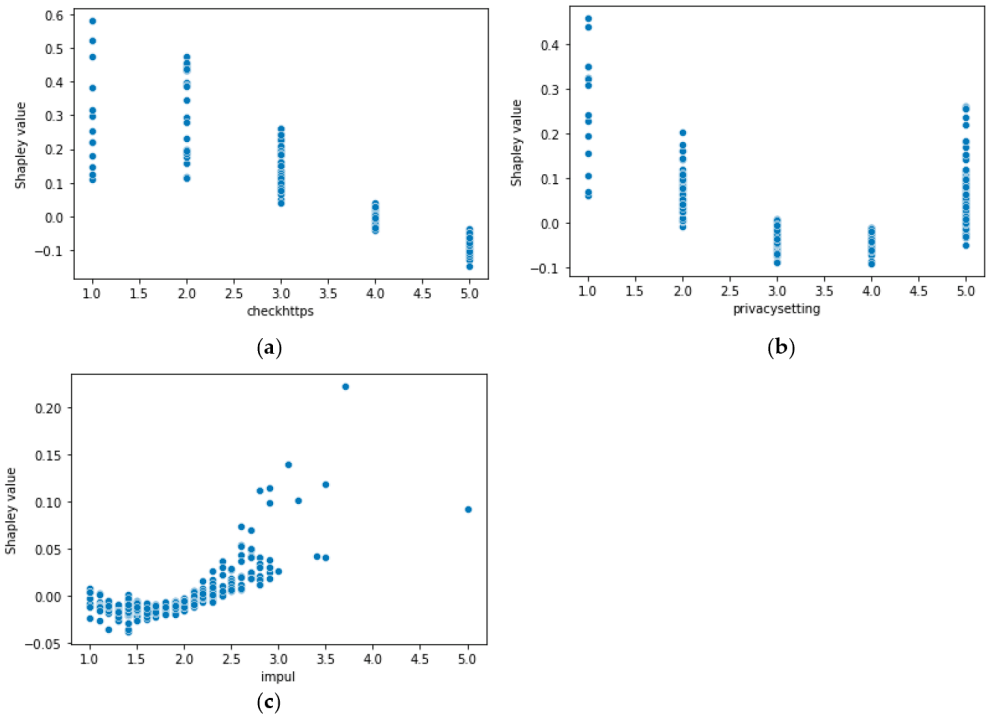

Figure 8. Scatter plots: (**a**) scatter plot for check https; (**b**) scatter plot for privacy setting; (**c**) scatter plot for impulsivity.

These examples illustrate the application of our SHAP explanation method for designing future anti-phishing interventions. It is important to note that these results are based on the data collected from our simulated phishing campaigns at GMU, and therefore, these recommendations may not be universally applicable across diverse populations. However, the methodology proposed in this study can be adapted to different datasets targeting diverse populations, enabling the generation of personalized anti-phishing training for varied demographic groups.

When applying the model's recommendations in real-world scenarios, especially in organizations or educational institutions, several considerations should be taken into account. Firstly, it is crucial to recognize that the model's recommendations are based on the features and behaviors observed in the specific dataset used for training. Therefore, generalization to diverse populations or evolving threat landscapes may require ongoing validation and adaptation.

Organizations and educational institutions should consider the ethical implications of using predictive models, ensuring that the recommendations are fair, unbiased, and do not disproportionately impact certain groups. Transparent communication about the model's purpose, limitations, and the potential consequences of following its recommendations is essential to build trust among users. Additionally, the model's recommendations should be integrated into broader cybersecurity awareness and training programs rather than being solely relied upon. Human judgment and expertise remain critical in evaluating and contextualizing the recommendations within the dynamic and complex nature of cybersecurity threats.

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. Employing this approach is crucial in offering personalized guidance, enabling individuals to proactively mitigate their susceptibility to phishing attacks. This is achieved through an understanding of factors, such as browsing behavior, that significantly contribute to being a target for 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 the current results, 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. Another potential avenue for future research involves exploring susceptibility to phishing across various phishing attacks. This entails analyzing the impact of different attacks on susceptibility levels and developing adaptive models capable of evolving with the changing landscape of phishing techniques. Such investigation could provide valuable insights to inform the development of customized anti-phishing interventions tailored to distinct types of phishing attacks. Moreover, the employed SHAP method is specific to the model, necessitating a recalibration process for any modifications or updates to the prediction model, which can be time-consuming. Therefore, it is essential to investigate the development of new explanation methods such as structural causal models that can adapt more efficiently to changes in the data. Furthermore, the SHAP method used in the current study does not consider actionable considerations such as causality restrictions and the stereotypical nature of certain features like age and gender. Hence, it is crucial to incorporate these actionable considerations into our model to enhance its practical applicability.

Author Contributions: Conceptualization, Z.F., W.L. and K.-C.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.C.; supervision: K.B.L. and K.-C.C.; writing—original draft, Z.F. and W.L.; writing—reviewing and editing, K.B.L. and K.-C.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 conflicts of interest.

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, San Jose, CA, USA, 17–18 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*; University of Hawaii Manoa Library: Honolulu, HI, USA, 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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
6. Halevi, T.; Lewis, J.; Memon, N. Phishing, Personality Traits and Facebook. *arXiv* **2013**, arXiv:1301.7643.
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. [CrossRef]
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. [CrossRef]
9. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
10. 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*, 1–48. [CrossRef]
11. James, P.J.; Bailey, J.; Courtney, J. A Personality Based Model for Determining Susceptibility to Phishing Attacks. In Proceedings of the Southwest Decision Sciences Institute Annu. Meeting (SDSI '09), Oklahoma, OK, USA, 24–28 February 2009; pp. 285–296.
12. Jagatic, T.N.; Johnson, N.A.; Jakobsson, M.; Menczer, F. Social Phishing. *Commun. ACM* **2007**, *50*, 94–100. [CrossRef]
13. 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 SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 10 April 2010; Association for Computing Machinery: New York, NY, USA, 2010; pp. 373–382.
14. Blythe, M.; Petrie, H.; Clark, J.A. F for Fake: Four Studies on How We Fall for Phish. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 7 May 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 3469–3478.
15. 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), Abu Dhabi, United Arab Emirates, 18–20 March 2012; pp. 249–254.
16. 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*, 1–28. [CrossRef]
17. 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. [CrossRef]
18. Downs, J.S.; Holbrook, M.B.; Cranor, L.F. Decision Strategies and Susceptibility to Phishing. In Proceedings of the Second Symposium on Usable Privacy and Security, New York, NY, USA, 12 July 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 79–90.
19. Canham, M.; Posey, C.; Strickland, D.; Constantino, M. Phishing for Long Tails: Examining Organizational Repeat Clickers and Protective Stewards. *SAGE Open* **2021**, *11*, 2158244021990656. [CrossRef]
20. Digman, J.M. Personality Structure: Emergence of the Five-Factor Model. *Annu. Rev. Psychol.* **1990**, *41*, 417–440. [CrossRef]
21. Alseadoon, I.; Chan, T.; Foo, E.; Nieto, J.G. Who Is More Susceptible to Phishing Emails?: A Saudi Arabian Study. In Proceedings of the 23rd Australasian Conference on Information Systems, Geelong, Australia, 3–5 December 2012; pp. 1–11.
22. 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. [CrossRef]
23. 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*, 1–35. [CrossRef]
24. Zhuo, S.; Biddle, R.; Koh, Y.S.; Lottridge, D.; Russello, G. SoK: Human-Centered Phishing Susceptibility. *ACM Trans. Priv. Secur.* **2023**, *26*, 1–27. [CrossRef]
25. 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), Tucson, AZ, USA, 28–30 September 2016; pp. 169–174.
26. 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. [CrossRef]
27. 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. [CrossRef]

28. 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. [CrossRef]
29. Cranford, E.; Jabbari, S.; Ou, H.-C.; Tambe, M.; Gonzalez, C.; Lebiere, C. Combining Machine Learning and Cognitive Models for Adaptive Phishing Training. In Proceedings of the 20th Annual Meeting of the International Conference on Cognitive Modeling, Toronto, ON, Canada, 23–27 July 2022.
30. 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. [CrossRef]
31. 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. [CrossRef]
32. 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. [CrossRef]
33. 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. [CrossRef]
34. 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. [CrossRef]
35. 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. [CrossRef]
36. Sahingoz, O.K.; Buber, E.; Demir, O.; Diri, B. Machine Learning Based Phishing Detection from URLs. *Expert Syst. Appl.* **2019**, *117*, 345–357. [CrossRef]
37. Akinyelu, A.A.; Adewumi, A.O. Classification of Phishing Email Using Random Forest Machine Learning Technique. *J. Appl. Math.* **2014**, *2014*, e425731. [CrossRef]
38. AlEroud, A.; Karabatis, G. 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, New York, NY, USA, 16 March 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 53–60.
39. 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), Riyadh, Saudi Arabia, 19–21 March 2020; pp. 1–6.
40. 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. [CrossRef]
41. 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. [CrossRef]
42. 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. [CrossRef]
43. Butnaru, A.; Mylonas, A.; Pitropakis, N. Towards Lightweight URL-Based Phishing Detection. *Future Internet* **2021**, *13*, 154. [CrossRef]
44. 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. [CrossRef]
45. Alhogail, A.; Alsabih, A. Applying Machine Learning and Natural Language Processing to Detect Phishing Email. *Comput. Secur.* **2021**, *110*, 102414. [CrossRef]
46. Divakaran, D.M.; Oest, A. Phishing Detection Leveraging Machine Learning and Deep Learning: A Review. *arXiv* **2022**, arXiv:2205.07411. [CrossRef]
47. 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), Coimbatore, India, 6–7 March 2020; pp. 398–404.
48. 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), Orlando, FL, USA, 5–7 December 2021; pp. 1–6.
49. 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. [CrossRef]
50. 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. In Proceedings of the 30th USENIX Security Symposium (USENIX Security), Vancouver, BC, Canada, 11–13 August 2021; pp. 3793–3810.
51. Kluge, K.; Eckhardt, R. Explaining the Suspicion: Design of an XAI-Based User-Focused Anti-Phishing Measure. In *Innovation through Information Systems*; Ahlemann, F., Schütte, R., Stieglitz, S., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 247–261.
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.

53. Heartfield, R.; Loukas, G.; Gan, D. You Are Probably Not the Weakest Link: Towards Practical Prediction of Susceptibility to Semantic Social Engineering Attacks. *IEEE Access* **2016**, *4*, 6910–6928. [CrossRef]
54. Wright, R.T.; Marett, K. The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived. *J. Manag. Inf. Syst.* **2010**, *27*, 273–303. [CrossRef]