Towards Personalized Anti-Phishing: Counterfactual Explanation Approach

Zhengyang Fan

zfan3@gmu.edu

George Mason University

Wanru Li

George Mason University

Kathryn Laskey

George Mason University

Kuo-Chu Chang

George Mason University

Research Article

Keywords: Phishing Susceptibility, Cyber Security, Explainable Artificial Intelligence, Counterfactual Explanation, Personalized Intervention

Posted Date: May 3rd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4335902/v1

Additional Declarations: No competing interests reported.

Towards Personalized Anti-Phishing: Counterfactual Explanation Approach

Zhengyang Fan[1*†], Wanru Li[1†], Kathryn B. Laskey[1], Kuo-Chu Chang[1]

[1*]Department of Systems Engineering and Operations Research, George Mason University, Fairfax, 22030, VA, USA.

*Corresponding author(s). E-mail(s): zfan3@gmu.edu;
Contributing authors: wli15@gmu.edu; klaskey@gmu.edu; kchang@gmu.edu;
†These authors contributed equally to this work.

Abstract

In today's digital landscape, phishing attacks persist as a formidable challenge, highlighting the need for robust strategies to mitigate individual risk. While advanced machine learning techniques have excelled in identifying those most susceptible to phishing, existing research has primarily focused on refining prediction accuracy rather than leveraging this understanding to mitigate risk. To bridge this gap, we present a novel counterfactual explanation approach aimed at identifying the specific traits that heighten an individual's vulnerability to phishing. Our approach integrates uncertainties and causal insights from the data generation process, producing actionable intelligence to effectively lower individual susceptibility. This enables us to tailor personalized recommendations to reduce individual's vulnerability. Through experimentation, we assess the efficacy of our methodology and demonstrate capacity to reduce susceptibility to phishing. These findings emphasize the importance of personalized interventions, arming individuals with the knowledge necessary to improve their online security protocols.

Keywords: Phishing Susceptibility, Cyber Security, Explainable Artificial Intelligence, Counterfactual Explanation, Personalized Intervention

1 Introduction

Phishing attacks have become a pervasive threat, with scammers exploiting email and other communication platforms to deceive victims into disclosing sensitive information. These fraudulent activities have led to significant financial losses for both commercial organizations and government entities. According to the recent FBI Internet Crime Annual Report, phishing attacks have reached their highest level since 2019, surpassing personal data breaches in terms of victim count in 2022. Furthermore, the financial impact of internet crimes, including phishing, soared to a staggering $10.3 billion in 2022, nearly doubling the previous year's figures.

To combat phishing, various efforts have been made, including law enforcement interventions, automated detection systems, and user education initiatives. However, to effectively enhance cybersecurity awareness and develop protective measures, it is crucial to understand the factors that contribute to an individual's susceptibility to phishing attacks. Previous research in this field has primarily relied on statistical tests to identify

these factors, such as age, demographics, and user behaviors [1–4]. However, these findings only provide insights at a collective scale and do not offer tailored guidance or help individuals in mitigating their vulnerability, as each person possesses distinct characteristics that may not align with population-level findings.

Alternatively, other studies have focused on building accurate prediction models for phishing behavior using machine learning techniques. Although these models can achieve high performance, their interpretability poses challenges, hindering the development of targeted educational and awareness campaigns to prevent and mitigate the impact of phishing attacks [5–7].

To address these limitations, our paper takes a step beyond merely predicting phishing behavior accurately and utilizes counterfactual explanation (CE), a form of explainable artificial intelligence (XAI), to analyze the rationale underlying a learned prediction model, such as deep neural networks (DNN). By employing CE, we aim to provide personalized suggestions and interventions to reduce individuals' susceptibility to phishing attacks. Our research contributes to bridging the gap between machine learning models' predictions and providing actionable guidance for individuals. By leveraging CE and personalization, we can empower users with tailored recommendations to enhance their ability to recognize and defend against phishing attempts. The contributions of this paper can be summarized as follows:

1. Novel Counterfactual Explanation Framework: In order to generate meaningful counterfactual explanations, we integrate uncertainties and causal insights of the data generation process and frame the search for relevant counterfactual instances as an optimization problem. By considering causal relationships and uncertainty, we are able to produce more actionable and informative counterfactual explanations that shed light on the underlying factors influencing an individual's susceptibility to phishing attacks.
2. Novel Application of Counterfactual Explanation: This study represents the first attempt to utilize counterfactual explanation techniques in the analysis of susceptibility to phishing attacks. By applying this approach, we can gain insights into the specific factors and behaviors that contribute to an individual's vulnerability. Furthermore, we leverage counterfactual explanations to provide personalized recommendations and interventions, enabling individuals to reduce their susceptibility to future phishing attacks.

By combining the application of counterfactual explanation techniques with our developed optimization model, we enhance the understanding of individual vulnerability to phishing attacks and provide personalized recommendations and interventions. These contributions significantly advance research in the realm of phishing susceptibility analysis and offer practical insights for enhancing individual cybersecurity measures.

The rest of the paper is organized as follows: Section 2 presents a review of the relevant literature, focusing on previous studies related to phishing susceptibility analysis and counterfactual explanations. Section 3 briefly describes the dataset used in our experiments. Section 4 introduces the counterfactual explanation framework and our proposed optimization model. Section 5 presents the experimental results and provides discussions based on these findings. Finally, Section 6 concludes the paper.

2 Related Literature

This section is organized as follows: Subsection 2.1 focuses on the review of counterfactual explanation methodology, discussing its applications and relevance to our research. In Subsection 2.2, we review the pertinent applications of machine learning and explainable artificial intelligence (XAI) methodologies in the context of phishing-related research.

2.1 Counterfactual Explanations

Counterfactual explanation, a recent advancement in XAI, has emerged as a method to understand the decision-making process of machine learning models [8]. Introduced by Wachter et al. [9], counterfactual explanations tackle the optimization problem of generating a new instance that closely resembles the original instance while producing a different outcome. By analyzing this new instance, a counterfactual explanation can be derived to elucidate the model's decision process. This type of

explanation proves particularly valuable in critical domains like healthcare [10, 11] and finance [12,13], where comprehending the rationale behind decisions is crucial.

Albini et al. [14] proposed a method for generating counterfactual explanations tailored to various types of Bayesian Network Classifiers, focusing on the influential relationships between variables rather than probabilistic information. To address the optimization problem, Mahajan et al. [15] developed a generative model using variational auto-encoders (VAE) to predict counterfactual explanation points without explicitly solving the optimization problem. Downs et al. [16] proposed a variant of VAE called conditional sub-space VAE, which generates counterfactual explanations while considering correlations between features, causal relations between features, and personal preferences. Similarly, Verma et al. [17] employed reinforcement learning techniques to generate counterfactual points for model explanation.

Karimi et al. [18] and Lash et al. [19] categorized features into immutable, mutable, and actionable types and highlighted the distinction between features that can be changed and features that can be acted upon by individuals. For instance, in a financial application, credit score may not be directly modifiable but can be influenced by changes in other features such as income.

Mothilal et al. [20] proposed a framework based on determinantal point processes for generating and evaluating diverse counterfactual explanations while considering user context and constraints. The framework aimed to satisfy the feasibility and diversity of counterfactual actions, and experimental results on real-world datasets demonstrated its superiority over previous approaches. Similarly, Poyiadzi et al. [21] introduced a method focuses on providing feasible and actionable paths for transformation. They proposed the FACE algorithm, which generates coherent and achievable counterfactual instances based on density-weighted metrics.

For additional references on counterfactual explanations, see [22], [23], and [24].

2.2 Machine Learning and XAI in Phishing Study

Abbasi et al. [25] utilizes cluster analysis and a controlled experiment to identify user segments with high susceptibility to phishing based on their perceptions, demographics, and behavior on phishing websites. Their findings have important implications for training programs, anti-phishing tool usability, and security policies. Cranford et al. [26] 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 incorporate a cognitive model of phishing susceptibility to inform the bandit model's parameters. Yang et al. [27] focuses on predicting phishing victims by proposing a multidimensional phishing susceptibility prediction model (MPSPM). Through an experiment involving 1105 volunteers, the study collects demographic, personality, knowledge, experience, security behavior, and cognitive process data to classify users into susceptible and nonsusceptible categories using supervised learning methods. The findings show high accuracy in predicting user phishing susceptibility, highlighting the importance of machine learning techniques in combating phishing attacks. Yang et al. [28] addresses the persistent threat of phishing attacks and the limitations of existing anti-phishing tools in accurately identifying user susceptibility. The proposed user phishing susceptibility prediction model combines static features (experience, demographics, knowledge) and dynamic features (design changes, eye tracking) to predict susceptibility. Through questionnaire surveys and eye-tracking experiments, the model achieves a high prediction accuracy of 92.34 %, demonstrating its effectiveness in assessing user susceptibility to phishing by considering a comprehensive set of static and dynamic features.

Machine learning techniques have also been employed for phishing detection. Different approaches have been used to extract phishing classification information from various sources, including visual information like logos [29–31], textual information like URLs [32–34], and webpage content [35, 36].

In the field of explainable artificial intelligence (XAI), several studies have focused on applying

XAI techniques to understand and combat phishing attacks. Hernanders et al. [37] addresses the increasing threat of phishing attacks in the context of the expanding internet and online transactions. While Artificial Intelligence (AI)-based protection methods have shown efficiency, they often lack explanations for their categorization decisions. The study aims to detect phishing using explainable techniques, specifically Local Interpretable Model-Agnostic Explanations (LIME) and Explainable Boosting Machine (EBM), highlighting recent advancements and future directions in the field. Chai et al. [38] address the limitations of existing methods for detecting phishing websites and propose a multi-modal hierarchical attention model (MMHAM). This model learns deep fraud cues from three major modalities: URLs, textual information, and visual design. The MMHAM incorporates a shared dictionary learning approach in the attention mechanism to align representations from different modalities, improving phishing detection performance. The model not only enhances deep cue learning but also provides a hierarchical interpretability system, enabling the development of phishing threat intelligence for detecting phishing websites at different levels. In another study, Lin et al. [39] present Phishpedia, a hybrid deep learning system designed to overcome technical challenges in phishing identification. It focuses on accurately recognizing identity logos on webpage screenshots and matching logo variants of the same brand. Phishpedia achieves high accuracy and low runtime overhead without the need for training on phishing samples. Experimental results demonstrate its superiority over baseline identification approaches, and its deployment with the CertStream service led to the discovery of a significant number of new real phishing websites, including those not reported by other engines in VirusTotal. Kluge and Eckhardt [40] introduce a user-focused anti-phishing measure that incorporates Explainable Artificial Intelligence (XAI) techniques. By leveraging advanced phishing detectors, this approach identifies key words and phrases in an email that are crucial for detecting phishing attempts. Empirical results demonstrate the effectiveness of the approach in extracting relevant text segments to discriminate between genuine and phishing emails, highlighting the potential of XAI methods in the field of phishing prevention and beyond. Most recently, Fan et al. [7] present a novel machine learning approach that incorporates XAI techniques to investigate the impact of human and demographic factors on susceptibility to phishing attacks. The research explores the influence of psychological factors and online security habits on individuals' vulnerability to phishing scams using Shapley additive explanations (SHAP) [41]. The findings provide practical insights and personalized recommendations to help individuals mitigate the risk of falling victim to phishing attacks based on their specific circumstances.

3 Data Description

The data for this study was derived from a simulated phishing experiment conducted by a team that includes several authors of this paper. The experiment aimed at determining the characteristics of susceptilble users. The study involved 6,938 faculty and staff at George Mason University. Over a three-week span, from October 30 to November 21, 2018, participants received one of three types of simulated phishing emails, simulating IT/tech support, finance/banking, or e-commerce/package delivery scenarios. Individuals who clicked on the phishing link in these emails are labeled as PHISHED for subsequent analysis and were randomly directed to one of three specifically designed landing pages: one displaying a "page not found" error, one showing a message that disclosed the email as part of a phishing study, and one presenting the same disclosure along with a brief anti-phishing training video. This experiment was carefully structured using a Latin square design to evenly distribute the phishing simulations among participants and was accompanied by a detailed questionnaire to collect demographic, psychological, and behavioral data, which is used as features in our analysis. The comprehensive design of the phishing campaigns is extensively described in our previous works [1, 4]. Table 1 below shows the name, type and description of each variable used in our study.

In summary, our study incorporates 17 features, including five psychological characteristics, eight behavior-related factors, and four demographic variables. Given that only 504 out of the 6938 participants completed the questionnaire,

Table 1 Variable Name, Type and Description for Data

Variable Name	Type	Value Type	Description
Impulsivity	Psychological	Numeric	Range from 1 to 5 to measure impulsivity score.
Conscientiousness	Psychological	Numeric	Range from 1.0 to 5.0 to measure conscientiousness score
Emotional Stability	Psychological	Numeric	Range from 1.0 to 5.0 to measure the emotional stability score
Agreeableness	Psychological	Numeric	Range from 1.0 to 5.0 to measure the agreeableness score
Perceived Stress	Psychological	Numeric	Range from 1.0 to 5.0 to measure the perceived stress score
Check Link	Behavior	Numeric	Range from 1.0 (never) to 5.0 (very often)
Privacy Setting	Behavior	Numeric	Range from 1.0 (never) to 5.0 (very often)
Check HTTPS	Behavior	Numeric	Range from 1.0 (never) to 5.0 (very often)
Click w/o Check	Behavior	Numeric	Range from 1.0 (never) to 5.0 (very often)
Phished Before	Behavior	Binary	Binary valued: Yes = 1, No = 0
Phished in Last 3 Months	Behavior	Binary	Binary valued: Yes = 1, No = 0
Lose Info Due to Phishing	Behavior	Binary	Binary valued: Yes = 1, No = 0
Download Malware	Behavior	Binary	Binary valued: Yes = 1, No = 0
Age	Demographic	Categorical	5 values: [19, 27), [27, 41), [41, 49), [49, 59), [59+)
Gender	Demographic	Categorical	2 values: Female, Male
Department	Demographic	Categorical	3 values: Technical college, Administrative, Other College
Position	Demographic	Categorical	4 values: Full-time faculty, adjunct faculty, wage staff, other staff
Click	Label	Binary	Individuals who clicked at least one simulated phishing scams are labeled as PHISHED

our subsequent analysis concentrates on this subset. Within this group, 121 individuals clicked on the simulated phishing emails.

4 Methodology

In this section, we will outline our methodology for generating counterfactual explanations that integrate causal knowledge, aiming to enhance the reliability and actionability of the resulting counterfactual instances. Figure 1 outlined the research framework.

Wachter et al. [9] introduced a novel approach to generating counterfactual explanations by formulating it as an optimization problem. The objective of this optimization problem, as stated in Equation 1, is to minimize the distance between the counterfactual point x' and the original data

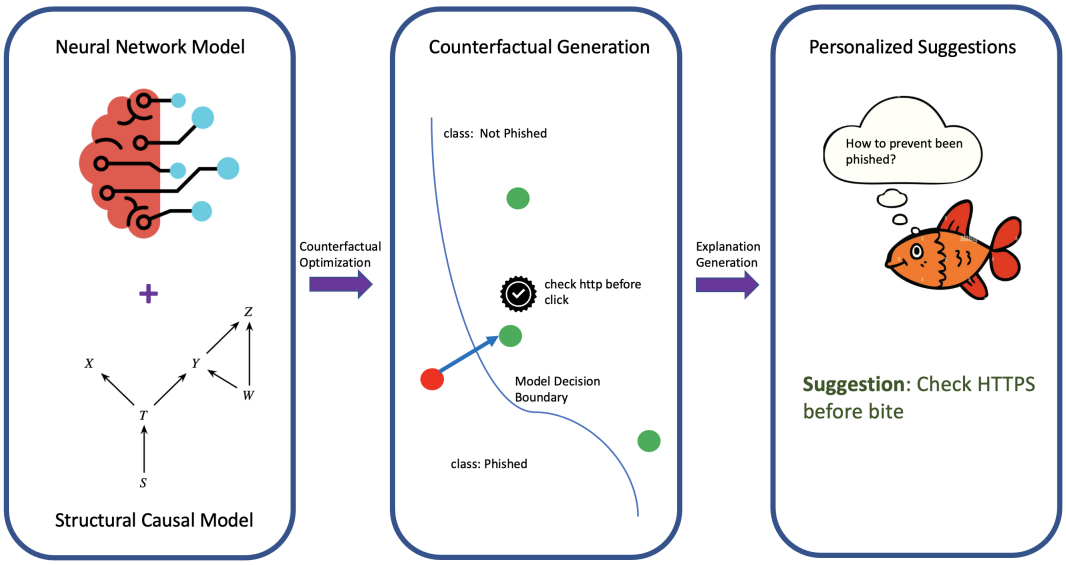

Fig. 1 Causality Aided Counterfactual Explanation Framework

point x, while ensuring that the classifier's output for the counterfactual point falls within the desired class (e.g., acceptance of a loan application). In simpler terms, they sought to find a new data point that is as close as possible to the original one, but with the desired outcome.

$$\begin{aligned} \min_{x' \in \mathcal{X}} \quad & d(x, x') \\ \text{s.t.} \quad & f(x') = y' \end{aligned} \tag{1}$$

The objective function in this optimization problem aims to ensure that the counterfactual point x' remains in close proximity to the original data point x. The measure of closeness is determined by a metric, such as the $\ell 1$ or $\ell 2$ distance, which is defined based on the feasible region of the data $\mathcal{X}$. If the machine learning model f is differentiable, it is possible to reformulate the given optimization problem (Equation 1) as a differentiable and unconstrained optimization problem (Equation 2), facilitating the use of gradient-based optimization techniques.

$$\min_{x' \in \mathcal{X}} \quad \lambda_1 \left(f(x') - y'\right)^2 + d(x, x') \tag{2}$$

In Equation 2, the parameter λ_1 is a user-defined value that can be set to a relatively large number. This choice of lambda encourages the output of the classifier for the counterfactual point to be in close proximity to the desired class, effectively emphasizing the importance of achieving the desired outcome. In this paper, the $\ell 2$ distance or its weighted version will be utilized as the metric to quantify the proximity between the counterfactual point x' and the original data point x:

$$d(x, x') = \sum_{k \in K} (x_k - x'_k)^2 \tag{3}$$

$$d^{weight}(x, x') = \sum_{k \in K} \frac{(x_k - x'_k)^2}{\text{std}(x_k)}$$

where $\text{std}(x_k)$ is the standard deviation for the k^{th} feature.

An important limitation of the model formulation in Equation 2 is that it allows for independent changes to the features of the input x when constructing the counterfactual point x'. However, neglecting the interdependence or causal relationships between features can lead to the generation of invalid counterfactuals that are not logically reasonable or actionable. For instance, in financial applications, credit scores are dependent on other feature values such as income, making it inappropriate to modify them independently. To tackle this issue, Mahajan et al. [15] proposed incorporating a structural causal model (SCM) and introducing an additional term in the objective function to ensure the preservation of causal requirements.

To be self complete, we briefly review some concepts in the structural causal model. A structural causal model can be formally defined as a tuple (X, G, F, N), wherein X represents a set of real-valued random variables known as endogenous variables. The directed acyclic graph (DAG) G describes the causal dependence structure among the variables in X. F denotes a set of functions, where each function f_i is associated with a specific variable $X_i \in X$. Furthermore, N represents a set of real-valued random variables referred to as exogenous or noise variables [42]. Together, these components provide a formal framework for modeling and analyzing causal relationships among variables in a structured manner. With the above definition, each variables $X_i \in X$ can be expressed as

$$X_i = f_i(Pa(X_i), N_i)$$

where $Pa(X_i)$ represents the parameters of node X_i. one usually assume the exogenous variables are additive and are independent to endogenous variables X [43]:

$$X_i = f_i(Pa(X_i)) + N_i$$

Figure 2 below shows a simple example of a structrual causal model. In the figure, $X = \{X_1, X_2\}$, $N = \{N_1\}$. The valued of node X_1 is determined by a structural equation

$$X_1 = f_1(X_2) + N_1$$

With the structural causal model, Mahajan et al. [15] proposed the following causal proximity loss to preserve the causal restrictions:

$$d_{causal}(x') = \sum_{k \in K} d(x'_k, f_k(Pa(x'_k))), \qquad (4)$$

where $d(\cdot, \cdot)$ denotes the $\ell 2$ distance as defined in Equation 3. Notice that this distance metric does not include the original data point x and force the values of counterfactual points to be close to the predicted value of x_k using structural causal model. Therefore, given structural casual model among feature inputs, we can solve the following optimization problem to generate counterfactual

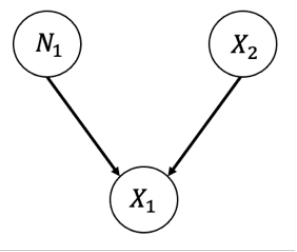

Fig. 2 Structural Causal Model Example

point that does not violate causal restrictions:

$$\min_{x' \in \mathcal{X}} \lambda_1 (f(x') - y')^2 + \lambda_2 d_{causal}(x') + d(x, x') \quad (5)$$

However, the above problem is not solvable in general due to the fact that the causal proximity loss is extremely complicated when the structural equation $f_k(\cdot)$ is highly non-linear. To address this issue, we assume the function $f_k(\cdot)$ is linear. There are two reasons we follow this assumption:

1. The assumption can lead to an optimization model that is solvable via traditional gradient based algorithms;
2. Preliminary data visualizations (see Section 5) suggest that only linear functional relationships occur in our phishing data set.

However, there is one drawback remaining in the optimization model (5): it forces the counterfactual point to be close to the pre-trained function prediction $f_k(Pa(x'_k))$, which ideally is the conditional expectation of X'_k given its parent $Pa(X'_k)$. In reality, the true realizations of data points can be noisy, and the counterfactual point generated according to some optimization algorithms should take this uncertainty into account. Figure 3 below depicts the idea: in the figure, black point $\hat{x}$ will induce $\hat{\delta}$ penalty when we use $d_{causal}(x')$ defined in Equation 4. However, due to randomness inside the data generating process, $\hat{x}$ should not be penalized due to the fact that it falls into the confidence band of the learned predictor $f_k(Pa(X'_k))$. Similarly, $\tilde{x}$ will be penalized by $\tilde{\delta}_1$ if we use distance $d_{causal}(x')$. Due to the uncertainty inside the data generating process, it is more reasonable to penalize the deviation of $\tilde{x}$ from its causal prediction by $\tilde{\delta}_2$.

Recall the structural causal model defines $X'_k = f_k(Pa(X'_k)) + N_k$. By assuming homogeneity

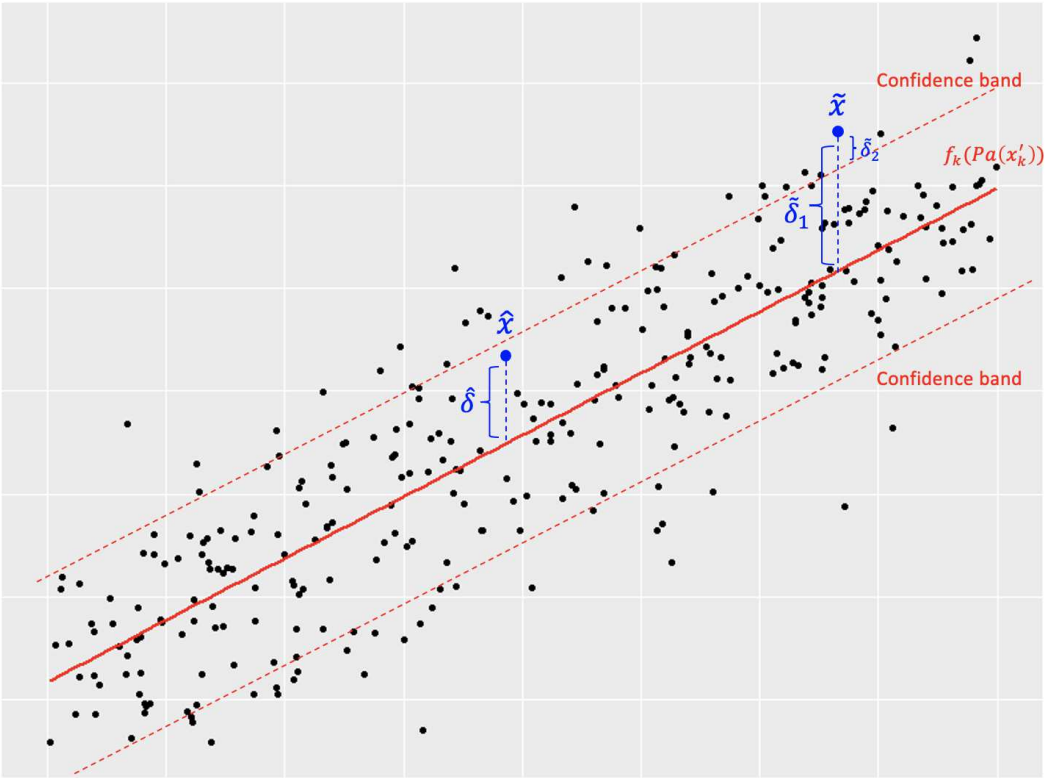

Fig. 3 Illustration of the Idea

of variance, we can first estimate sample standard deviation $\hat{\sigma}_k$ of N_k, and define new causal proximity loss as

$$\tilde{d}_{causal}(x') = \sum_{k \in K} \tilde{d}(x'_k, f_k(Pa(x'_k))) \triangleq \sum_{k \in K} \min_{-\lambda_3 \hat{\sigma}_k \leq z_k \leq \lambda_4 \hat{\sigma}_k} d(x'_k, f_k(Pa(x'_k)) + z_k)$$

where λ_3 and λ_4 are pre-defined parameters to quantify how wide the confidence band is. For simplicity, we let $\lambda_3 = \lambda_4 = 1$ in our case. With the newly defined causal proximity loss and positiveness of parameter λ_2, we can solve the following constrained optimization problem to generate counterfactual point that also incorporate causality requirements:

$$\begin{aligned} \min_{x' \in \mathcal{X}} \quad & \lambda_1 \left(f(x') - y'\right)^2 + d(x, x') + \\ & \lambda_2 \sum_{k \in K} d(x'_k, f_k(Pa(x'_k)) + z_k) \\ \text{s.t.} \quad & -\hat{\sigma}_k \leq z_k \leq \hat{\sigma}_k \quad \forall k \in K \end{aligned} \tag{6}$$

Notice that the above problem formulation can be solved using a classical non-linear programming algorithm such as projected gradient method.

5 Results and Discussion

In this section, we'll start with exploratory analysis. Following that, we'll employ a causal structure learning algorithm to uncover causal connections among the input features. Next, we'll showcase the predictive performance of a machine learning model, particularly a deep neural network that we used to predict phishing susceptibility. Lastly, we'll apply the proposed counterfactual explanation approach to conduct an in-depth examination of both the prediction model and the data, allowing us to offer tailored training recommendations.

5.1 Exploratory Analysis

A scatter plot matrix offers a concise overview of the relationships between multiple variables in a dataset. The diagonal of the matrix displays density estimates, depicting the distribution of each

variable, while the off-diagonal plots illustrate the pairwise relationships between variables.

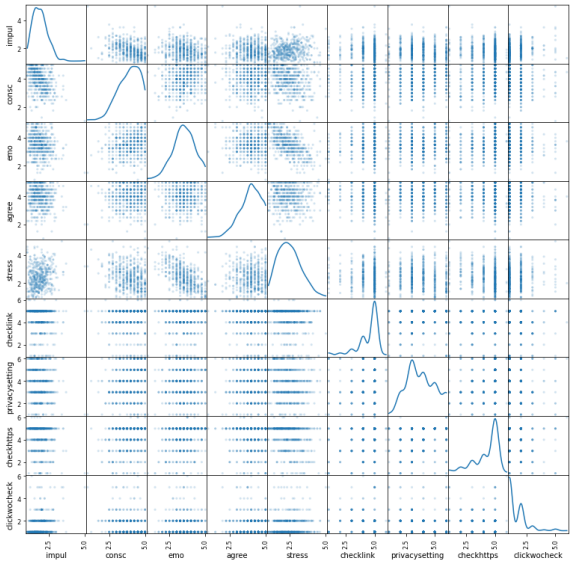

Fig. 4 Scatter Plot Matrix for Nine Actionable Factors

Figure 4 shows the scatter plot matrix depicting the relationships among the nine actionable factors: impulsivity, conscientiousness, emotional stability, agreeableness, perceived stress and cyber security related experience. The demographic factors such age, gender and position that cannot be changed are excluded from the figure as our focus lies solely on counterfactual explanations. Two observations can be found from this figure:

- Most variables are independent of each other. Therefore, a sparse causal relationship among these features is expected when learning the causal structures.
- For the variables that are correlated with each other, only linear relationships are observed, e.g., stress and emotional stability. Later, our residual plots further confirm that linear functions sufficiently capture the relationships among these variables.

A correlation heat map visually displays the relationships between variables using color gradients, providing a concise overview of their strength and direction. The correlation heat map depicted in Figure 5 indicates that the majority of variables exhibit independence or weak dependence, aligning with the findings presented in Figure 4.

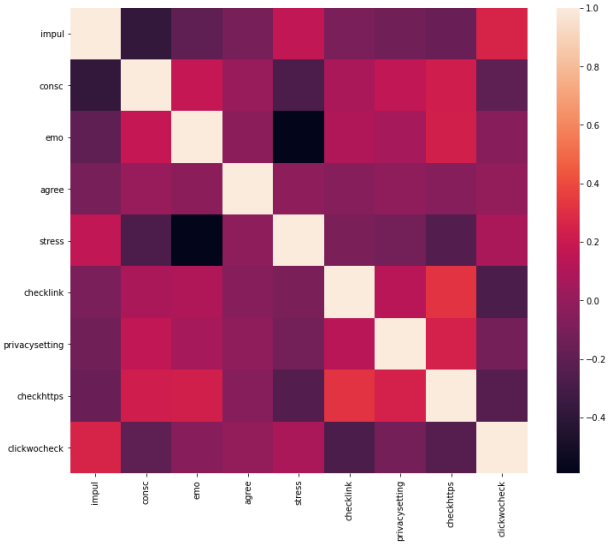

Fig. 5 Correlation Heat Map over Features

5.2 Learning Causal Structure

Based on our initial exploratory analysis, it was evident that the feature variables demonstrated a linear functional relationship. Consequently, we employed the DirectLiNGAM algorithm proposed by Shimizu et al. [44] to identify the causal structures underlying these features. The algorithm utilizes regression and independence tests iteratively to minimize estimation errors, assuming a linear model and non-Gaussian errors. The linear and non-Gaussian assumptions in DirectLiNGAM are necessary conditions for the algorithm to recovery the underlying true casual stricture. The linear model assumption in our data is relatively easy to be verified based on our exploratory analysis in previous sections. However, the non-Gaussian assumption cannot be directly verified since it requires the ground truth functional relationship among variables. To address this issue, we employ a post-hoc diagnostic approach based on residuals.

Two causal relationships are learned from our data using DirectLiNGAM algorithm:

- emotional stability is a cause of perceived stress
- impulsivity is a cause of conscientiousness

To verify whether the non-Gaussian assumption holds, we first fit two simple linear regression model to estimate structural functions using the least square method. The two fitted linear model is described as follows:

$$\text{Conscientiousness} = -0.555 * \text{Impulsivity} + 4.95$$
$$\text{Stress} = -0.614 * \text{Emotional Stability} + 4.606$$

Then, we calculate residuals by subtracting prediction from true value and Shapiro-Wilk test to test the non-Gaussian assumption. The test statistics and the corresponding p-value is reported in Table 2. Since the p value is less than 0.001 for both models, the non-Gaussian assumption is satisfied. Later, when we generating counterfactual explanations, we will include these two linear causal model into our optimization framework described in the methodology section. Moreover, the residual plots depicted in Figures 6-7 indicate that the linear function adequately models the relationships between variables, as no discernible patterns are evident in these plots.

Table 2 Testing Non-Gaussian Assumption

Model	Test Statistics	p value
impul-consc	0.981	<0.001
emo-stress	0.95	<0.001

Notice that these causal relationships are also consistent with literature such as [45] and [46] that observe emotional stability (impulsivity) is highly correlated with perceived stress (conscientiousness). Furthermore, our results complement the exact causal directions under the linear and non-Gaussian assumptions.

5.3 Deep Learning Predictor

We utilized a multi-layer perceptron (MLP) with 4 hidden layers as our predictor. To prepare the data for training and testing the performance of the MLP model, we divided our collected dataset into training and testing sets in a 4:1 ratio, with 80% of the data allocated to training and 20% to testing. Figure 8 to 13 shows that the distribution of the features is similar between the training and testing datasets, indicating homogeneity.

We implemented the NearMiss undersampling technique to address the class imbalance in our dataset [47]. Additionally, we utilized one-hot encoding to encode categorical and binary features like gender and position. These features cannot be directly used as input for the MLP model due to their nominal nature.

The performance of the trained MLP predictor was evaluated using the testing set. Table 3 summarizes the model performance results.

Table 3 Model Performance

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

These results indicate that the model exhibits reasonable predictive capabilities and can be valuable in identifying potential click risks in real-world scenarios.

5.4 Counterfactual Explanation and Personalized Recommendation

In this subsection, we provide examples of individuals classified as phishing victims to illustrate the process of reducing their vulnerability. It should be noted that certain categorical variables such as age and gender are not modifiable, and they are excluded from the calculation of counterfactual points. Table 4 presents the counterfactual points for the first individual who experienced phishing. In the table, the 'Original' row displays the observed values of each factor for this individual, whereas the 'CF' row shows the counterfactual explanation generated using the proposed framework. The differences between the values in the 'CF' row and the 'Original' row indicate the minimal changes this individual must make to reduce vulnerability to future phishing attacks. In other words, the counterfactual explanation provides a personalized actionable recommendation for this individual to reduce phishing susceptibility. The table reveals that 'Stress' is the most influential feature for reducing the individual's susceptibility

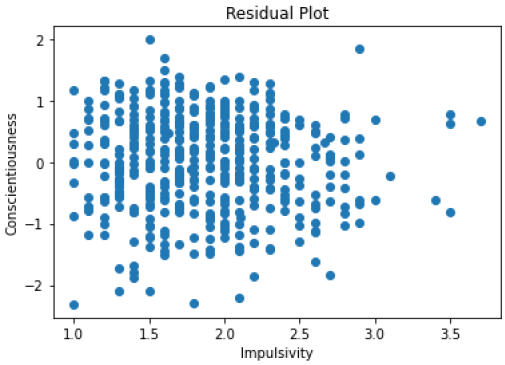

Fig. 6 Residual Plot for the First Linear Model

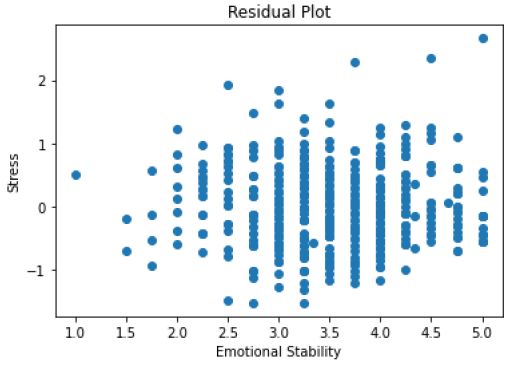

Fig. 7 Residual Plot for the Second Linear Model

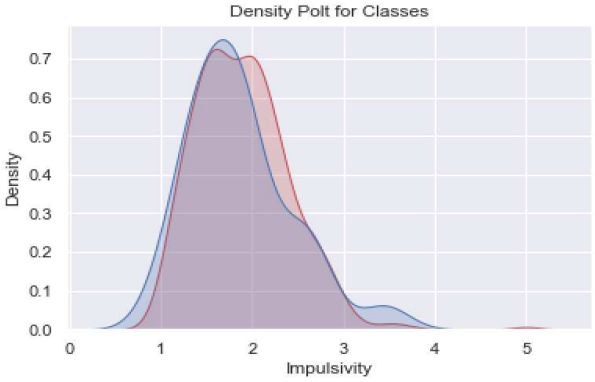

Fig. 8 Compare Density of Impulsivity

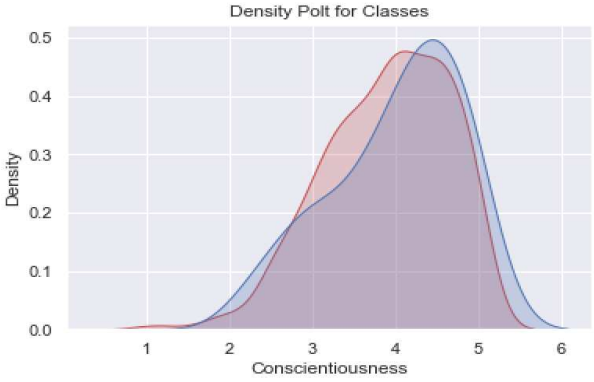

Fig. 9 Compare Density of Conscientiousness

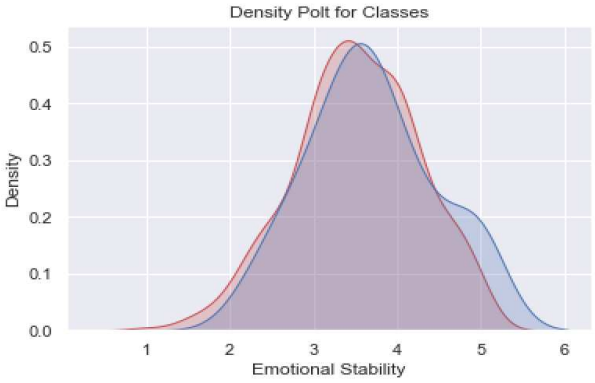

Fig. 10 Compare Density of Emotional Stability

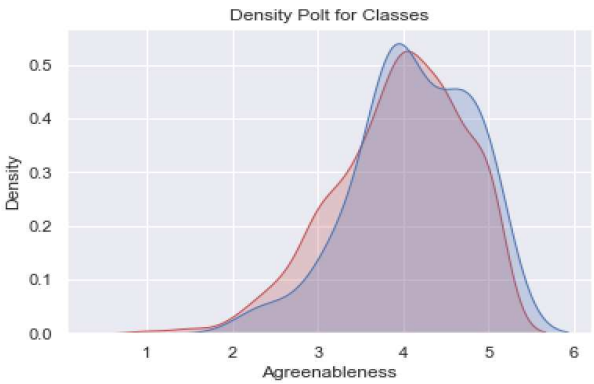

Fig. 11 Compare Density of Agreeableness

to future phishing attacks, suggesting the implementation of mind-body-stress-reduction (MBSR) techniques or other stress management interventions [48]. Additionally, there is a need to improve emotional stability for this individual, as it is identified as the cause of 'Stress' according to the causal information.

The counterfactual points for the second individual who experienced phishing are shown in Table 5. The table highlights that 'CheckHttps' is

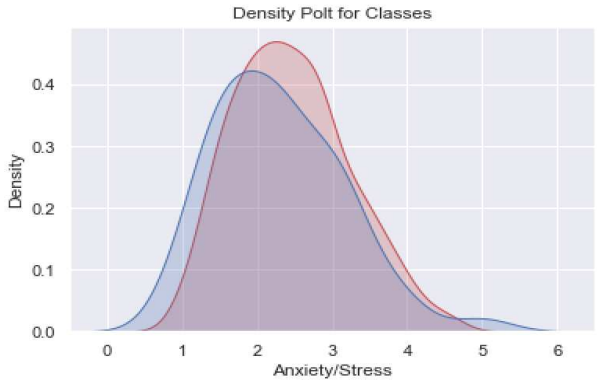

Fig. 12 Compare Density of Stress

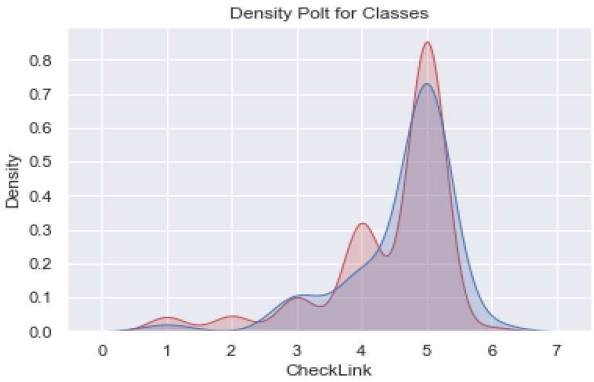

Fig. 13 Compare Density of Checklink

Table 4 Counterfactual Explanation for Victim No.1

	Impul	Consc	Emo	Agree	Stress	Link	Privacy	Http	woCheck
Original	1.2	3.5	2.75	5	3.9	4	3	4	3
CF	1.1	3.4	3.22	5	2.6	3.96	3.05	4.12	2

the most influential feature in reducing the individual's susceptibility to future phishing attacks. To enhance this individual's cyber security awareness, an anti-phishing training program could emphasize the use of tools like NoPhish, a smartphone app-based game designed for increasing cyber security related awareness [49].

Table 6 displays the counterfactual points for the third individual who experienced phishing. The results indicate that 'Impul' and 'Consc' are the most influential features in reducing the individual's susceptibility to future phishing attacks. To address these factors, an anti-phishing training program can concentrate on reducing impulsivity and increasing conscientiousness. Psychological intervention techniques introduced in [50] can be implemented to enhance the individual's conscientiousness and reduce impulsivity.

Based on the analysis of the three individuals who experienced phishing attacks, it is evident that personalized anti-phishing training programs should be customized to address their specific vulnerabilities in an efficient manner. For the first and third individuals, psychological interventions aimed at reducing stress levels and increasing conscientiousness can be effective in lowering their susceptibility to phishing. In contrast, the second individual would benefit from a cyber security-focused intervention to enhance their awareness and understanding of phishing threats. By tailoring interventions based on individual characteristics and needs, anti-phishing training programs can effectively mitigate the risk of future attacks. This emphasizes the importance of personalized approaches in combating phishing and highlights the need for targeted strategies to enhance security for individuals.

6 Conclusion

In this study, we propose a personalized anti-phishing training strategy based on counterfactual explanations. By utilizing a trained machine learning model and an optimization approach, we identify counterfactual points for each individual, considering their unique characteristics and vulnerabilities. Our findings indicate that while certain features may exhibit significant correlation with phishing susceptibility on a population scale, individual disparities necessitate personalized anti-phishing interventions.

This research emphasizes the significance of understanding the specific factors that contribute to an individual's susceptibility to phishing attacks. By tailoring anti-phishing training programs to address these individual vulnerabilities, we can effectively reduce future susceptibility.

Table 5 Counterfactual Explanation for Victim No.2

	Impul	Consc	Emo	Agree	Stress	Link	Privacy	Http	woCheck
Original	2.4	3.5	3.75	3	2.6	4	3	2	1
CF	2.4	3.6	3.89	2.78	2.6	4.2	3.13	5	1

Table 6 Counterfactual Explanation for Victim No.3

	Impul	Consc	Emo	Agree	Stress	Link	Privacy	Http	woCheck
Original	3.1	2.75	3.75	5	2.4	4	3	5	2
CF	2.0	5	3.93	5	2.4	4	3.12	5	1.88

Moreover, our method incorporates causal knowledge, enhancing the generation of counterfactual points and providing valuable insights into the underlying causal relationships.

However, the practical implementation of personalized training programs presents several challenges. For instance, when multiple factors such as impulsivity, conscientiousness, and stress are identified as crucial in decreasing susceptibility, determining the optimal sequence or approach for intervening in these factors without disrupting their causal relationships poses a dilemma. Future research should prioritize the development of comprehensive frameworks and guidelines aimed at effectively implementing personalized interventions.

References

[1] W. Li, J. Lee, J. Purl, F. Greitzer, B. Yousefi, K. Laskey, Experimental investigation of demographic factors related to phishing. Hawaii International Conference on System Sciences pp. 2240–2249 (2020)

[2] A. Diaz, A.T. Sherman, A. Joshi, Phishing in an academic community: A study of user susceptibility and behavior. Cryptologia **44**(1), 53–67 (2020)

[3] T. Halevi, J. Lewis, N. Memon, Phishing, personality traits and facebook. arXiv preprint arXiv:1301.7643. (2013)

[4] F.L. Greitzer, W. Li, K.B. Laskey, J. Lee, J. Purl, Experimental investigation of technical and human factors related to phishing susceptibility. ACM Transactions on Social Computing **4**(2), 1–48 (2021)

[5] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program. AI magazine **40**(2), 44–58 (2019)

[6] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)

[7] Z. Fan, W. Li, K.B. Laskey, K.C. Chang, Investigation of phishing susceptibility with explainable artificial intelligence. Future Internet **16**(1), 31 (2024)

[8] M. Gasse, D. Grasset, G. Gaudron, P.Y. Oudeyer, Causal reinforcement learning using observational and interventional data. arXiv preprint arXiv:2106.14421 (2021)

[9] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)

[10] Z. Wang, I. Samsten, P. Papapetrou, *Counterfactual explanations for survival prediction of cardiovascular ICU patients*, in *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings* (Springer, 2021), pp. 338–348

[11] D. Dave, H. Naik, S. Singhal, P. Patel, Explainable ai meets healthcare: A study on heart disease dataset. arXiv preprint arXiv:2011.03195 (2020)

[12] R.M. Grath, L. Costabello, C.L. Van, P. Sweeney, F. Kamiab, Z. Shen, F. Lecue, Interpretable credit application predictions with counterfactual explanations. arXiv preprint arXiv:1811.05245 (2018)

[13] M. Hashemi, A. Fathi, Permuteattack: Counterfactual explanation of machine learning credit scorecards. arXiv preprint

arXiv:2008.10138 (2020)

[14] E. Albini, A. Rago, P. Baroni, F. Toni, *Relation-Based Counterfactual Explanations for Bayesian Network Classifiers.*, in *IJCAI* (2020), pp. 451–457

[15] D. Mahajan, C. Tan, A. Sharma, Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv preprint arXiv:1912.03277 (2019)

[16] M. Downs, J.L. Chu, Y. Yacoby, F. Doshi-Velez, W. Pan, Cruds: Counterfactual recourse using disentangled subspaces. ICML WHI **2020**, 1–23 (2020)

[17] S. Verma, K. Hines, J.P. Dickerson, Amortized generation of sequential counterfactual explanations for black-box models. arXiv preprint arXiv:2106.03962 (2021)

[18] A.H. Karimi, B. Schölkopf, I. Valera, *Algorithmic recourse: from counterfactual explanations to interventions*, in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (2021), pp. 353–362

[19] M.T. Lash, Q. Lin, N. Street, J.G. Robinson, J. Ohlmann, *Generalized inverse classification*, in *Proceedings of the 2017 SIAM International Conference on Data Mining* (SIAM, 2017), pp. 162–170

[20] R.K. Mothilal, A. Sharma, C. Tan, *Explaining machine learning classifiers through diverse counterfactual explanations*, in *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), pp. 607–617

[21] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, *FACE: feasible and actionable counterfactual explanations*, in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), pp. 344–350

[22] S. Verma, V. Boonsanong, M. Hoang, K.E. Hines, J.P. Dickerson, C. Shah, Counterfactual explanations and algorithmic recourses for machine learning: A review. arXiv preprint arXiv:2010.10596 (2020)

[23] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974–12001 (2021)

[24] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery pp. 1–55 (2022)

[25] A. Abbasi, F.M. Zahedi, Y. Chen, *Phishing susceptibility: The good, the bad, and the ugly*, in *2016 IEEE conference on intelligence and security informatics (ISI)* (IEEE, 2016), pp. 169–174

[26] E.A. Cranford, S. Jabbari, H.C. Ou, M. Tambe, C. Gonzalez, C. Lebiere. Combining machine learning and cognitive models for adaptive phishing training (2022)

[27] R. Yang, K. Zheng, B. Wu, D. Li, Z. Wang, X. Wang, et al., Predicting user susceptibility to phishing based on multidimensional features. Computational Intelligence and Neuroscience **2022** (2022)

[28] R. Yang, K. Zheng, B. Wu, C. Wu, X. Wang, Prediction of phishing susceptibility based on a combination of static and dynamic features. Mathematical Problems in Engineering **2022** (2022)

[29] A.S. Bozkir, M. Aydos, Logosense: A companion hog based logo detection scheme for phishing web page and e-mail brand recognition. Computers & Security **95**, 101855 (2020)

[30] K.L. Chiew, E.H. Chang, W.K. Tiong, et al., Utilisation of website logo for phishing detection. Computers & Security **54**, 16–26 (2015)

[31] K.L. Chiew, J.S.F. Choo, S.N. Sze, K.S. Yong, Leverage website favicon to detect phishing websites. Security and communication networks **2018** (2018)

[32] D.J. Liu, G.G. Geng, X.C. Zhang, Multi-scale semantic deep fusion models for phishing website detection. Expert Systems with Applications **209**, 118305 (2022)

[33] L. Yang, J. Zhang, X. Wang, Z. Li, Z. Li, Y. He, An improved elm-based and data preprocessing integrated approach for phishing detection considering comprehensive features. Expert Systems with Applications **165**, 113863 (2021)

[34] O.K. Sahingoz, E. Buber, O. Demir, B. Diri, Machine learning based phishing detection from urls. Expert Systems with Applications **117**, 345–357 (2019)

[35] T. Wen, Y. Xiao, A. Wang, H. Wang, A novel hybrid feature fusion model for detecting phishing scam on ethereum using deep neural network. Expert Systems with Applications

211, 118463 (2023)

[36] A. Alhogail, A. Alsabih, Applying machine learning and natural language processing to detect phishing email. Computers & Security **110**, 102414 (2021)

[37] P.R.G. Hernandes, C.P. Floret, K.F.C. De Almeida, V.C. Da Silva, J.P. Papa, K.A.P. Da Costa, *Phishing detection using url-based xai techniques*, in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (IEEE, 2021), pp. 01–06

[38] Y. Chai, Y. Zhou, W. Li, Y. Jiang, An explainable multi-modal hierarchical attention model for developing phishing threat intelligence. IEEE Transactions on Dependable and Secure Computing **19**(2), 790–803 (2021)

[39] Y. Lin, R. Liu, D.M. Divakaran, J.Y. Ng, Q.Z. Chan, Y. Lu, Y. Si, F. Zhang, J.S. Dong, *Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages.*, in *USENIX Security Symposium* (2021), pp. 3793–3810

[40] K. Kluge, R. Eckhardt, *Explaining the suspicion: Design of an XAI-based user-focused anti-phishing measure*, in *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues* (Springer, 2021), pp. 247–261

[41] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

[42] J.Y. Halpern, J. Pearl, Causes and explanations: A structural-model approach. part ii: Explanations. The British journal for the philosophy of science (2005)

[43] J. Pearl, *Causality* (Cambridge university press, 2009)

[44] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P.O. Hoyer, K. Bollen, P. Hoyer, Directlingam: A direct method for learning a linear non-gaussian structural equation model. Journal of Machine Learning Research-JMLR **12**(Apr), 1225–1248 (2011)

[45] M.Y. Ho, F.M. Cheung, J. You, C. Kam, X. Zhang, W. Kliewer, The moderating role of emotional stability in the relationship between exposure to violence and anxiety and depression. Personality and Individual Differences **55**(6), 634–639 (2013)

[46] T. Mao, W. Pan, Y. Zhu, J. Yang, Q. Dong, G. Zhou, Self-control mediates the relationship between personality trait and impulsivity. Personality and Individual Differences **129**, 70–75 (2018)

[47] I. Mani, I. Zhang, *kNN approach to unbalanced data distributions: a case study involving information extraction*, in *Proceedings of workshop on learning from imbalanced datasets*, vol. 126 (ICML, 2003), pp. 1–7

[48] S.B. Stillwell, A.L. Vermeesch, J.G. Scott, Interventions to reduce perceived stress among graduate students: A systematic review with implications for evidence-based practice. Worldviews on Evidence-Based Nursing **14**(6), 507–513 (2017)

[49] G. Canova, M. Volkamer, C. Bergmann, R. Borza, *NoPhish: an anti-phishing education app*, in *Security and Trust Management: 10th International Workshop, STM 2014, Wroclaw, Poland, September 10-11, 2014. Proceedings 10* (Springer, 2014), pp. 188–192

[50] K.N. Javaras, M. Williams, A.R. Baskin-Sommers, Psychological interventions potentially useful for increasing conscientiousness. Personality Disorders: Theory, Research, and Treatment **10**(1), 13 (2019)