A Machine Learning Surrogate Approach for Scalable Design Optimization in Cybersecurity Simulation

Siyuan Zhai, Jeongkeun Shin, L. Richard Carley, Kathleen M. Carley

CASOS Center, Carnegie Mellon University, Pittsburgh PA 15213, USA
{siyuanzh,jeongkes,lrc,carley}@andrew.cmu.edu

Abstract. Cybersecurity has traditionally focused on exploiting system vulnerabilities within computing devices in targeted organizations. However, as organizations have built increasingly robust defensive infrastructures, attackers have shifted their strategies to exploit human vulnerabilities, or the weaknesses of end users who operate those systems using social engineering techniques. In response, many organizations have adopted user training and awareness programs to reduce the impact of such threats. However, quantitative research that analyzes the mechanisms by which behavioral and psychological vulnerabilities lead to actual security incidents remains underdeveloped. In particular, due to practical constraints, few studies have examined how dynamically changing psychological states such as fatigue and job performance affect a user's phishing susceptibility. To overcome these limitations, simulation researchers have attempted to combine various empirical findings to model dynamically fluctuating human factors and their effects on phishing susceptibility within simulation environments. However, as such models become more sophisticated, the underlying equations also become more complex, and repeatedly updating them at every simulation time step can result in substantial computational costs. To address the resulting performance degradation and improve the simulation speed, this study introduces a surrogate modeling approach based on a multilayer perceptron (MLP). The proposed method replaces the complex phishing susceptibility model with an efficient MLP surrogate that minimizes accuracy loss while significantly improving the computational efficiency and scalability of the simulation.

Keywords: Cybersecurity, Human Factors, Simulation Scalability

1 Introduction

In the past, cybercriminals primarily achieved their objectives by identifying and exploiting technical vulnerabilities in the computing devices operated by organizations, relying on purely technical methods such as Distributed Denial of Service (DDoS) attacks or web shell injections. However, as defensive technologies such as antivirus software, intrusion detection systems, and firewalls have

advanced, it has become increasingly difficult for attackers to breach targeted organizations and achieve their goals using technical methods alone. As a result, attackers have shifted their strategies toward social engineering techniques that exploit human error within organizations, using these methods to bypass sophisticated technical defenses and infiltrate target organizations in order to accomplish their objectives.

As this trend continues to grow, it is now reported that, according to IBM, more than 95% of all security incidents are caused by human error, with the most common examples being the downloading of malicious attachments or clicking on unsafe URLs [1]. As cyberattacks have evolved, cybersecurity researchers have increasingly focused on identifying which human factors make individuals more vulnerable to attacks that target human weaknesses, such as social engineering and spearphishing. In this context, research on phishing susceptibility has become more active, aiming to analyze how easily individuals fall victim to phishing attacks and how various human characteristics influence this susceptibility. As a result, many empirical studies have examined the relationship between phishing susceptibility and a variety of factors, such as personality traits, age, gender, educational background, cybersecurity-related experience and beliefs, and work commitment style [2]. However, most previous research has primarily focused on relatively static human factors that do not change easily. While these static factors certainly play an important role in determining an individual's susceptibility to phishing attacks, dynamic human factors that fluctuate over time, such as fatigue and job performance, can also have a significant impact in practice. Nevertheless, these dynamic factors have been largely overlooked in empirical research, as it is inherently difficult to control and capture their moment-to-moment variability, making it challenging to clearly analyze their relationship with phishing susceptibility.

To address this limitation in real-world scenarios, cybersecurity researchers have attempted to realistically model dynamic human factors such as memory recency, fatigue, perceived vulnerability, and job performance, and their influence on phishing susceptibility, by computationally modeling different human factors using regression models derived from various empirical studies [3]. However, as the model becomes more complex, involving more equations and rules for human factor modeling and requiring more frequent updates of dynamic human factors, the computational cost inevitably increases, degrading the simulation speed and scalability. To overcome this issue, we propose a surrogate approach [5] based on a multilayer perceptron (MLP) neural network [4], which predicts phishing susceptibility as a function of both static (age, gender, personality traits) and dynamic (fatigue and job performance) human factors. In this study, we first generate a large-scale synthetic dataset by systematically varying fatigue and job performance within a simulation framework based on established empirical models. Using this dataset, we train and evaluate an MLP model designed to flexibly respond to incremental changes in dynamic human factors. Our results demonstrate that this approach significantly improves simulation speed and scalability while incurring minimal loss in prediction accuracy.

2 Related Works

Empirical researchers in the field of cybersecurity have consistently explored how various human factors affect phishing susceptibility [2]. For example, Eftimie et al. conducted empirical phishing simulations to analyze the impact of age, gender, and Big Five personality traits on susceptibility to spearphishing attacks [6]. Ribeiro et al. also empirically investigated how factors such as age, gender, technological competence, education level, income, routine internet activities, and knowledge of phishing influence phishing susceptibility [7]. However, most previous empirical studies have mainly focused on the relationship between phishing susceptibility and relatively static human factors.

Simulation researchers have also actively studied how to model the phishing susceptibility of virtual humans in simulation environments. For example, Burns et al. leveraged social science theories to determine agent security levels and their influence on phishing rates [8]. Shin et al. proposed a method that assigns each agent unique phishing susceptibility scores before and after training (PSBE and PSAE) using regression models derived from empirical studies, particularly the findings of Eftimie et al. [6], with simulation outcomes calibrated to empirical results [9, 10]. Their framework was further extended to incorporate the effects of dynamic human factors such as job performance, fatigue, and perceived vulnerability [11].

However, their modeling approach had two significant limitations. First, the behavioral model fundamentally relied on switching between two extreme values, PSBE and PSAE, based on the state of human factors. While the probability of using each value could be adjusted, small changes in dynamic human factors such as fatigue or job performance were not reflected in a continuous or proportional manner in phishing susceptibility. Second, as their model incorporated an increasing number of regression models and rules, the simulation became substantially slower and scalability issues emerged due to the high computational cost. To overcome these limitations, this paper proposes a novel approach for modeling phishing susceptibility using a multilayer perceptron (MLP) neural network. The proposed model is designed to (1) respond more flexibly to gradual changes in human factors, enabling more realistic and interpretable simulation of phishing risk, and (2) improve simulation speed and scalability.

3 Data Farming

The objective of this study is to develop an accurate surrogate model [5] for predicting phishing susceptibility using a multilayer perceptron (MLP) neural network [4]. The proposed model is designed to estimate the phishing susceptibility of individual user agents within a simulation by comprehensively incorporating both static and dynamic human factors, including age, gender, Big Five personality traits, fatigue, and job performance. However, to the best of our knowledge, currently there is no empirical data set that contains all of these human factors together with actual measurements of phishing susceptibility. To address this

limitation, we adopted the behavioral framework proposed by Shin et al., which models the effects of fatigue and job performance on phishing susceptibility [11]. Based on this framework, we applied a data farming technique [12], commonly used in simulation-based research to generate the synthetic data required for model training.

We first imported the dataset of 235 virtual end user agents used by Shin et al. [10]. Each agent is characterized by demographic information (age and gender), psychological attributes (the Big Five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism), and two empirically estimated probabilities: phishing susceptibility before education (PSBE) and after education (PSAE). These values were calculated based on regression models developed in previous phishing studies [6]. To simulate the influence of dynamic human factors, we defined two continuous variables: fatigue and job performance. Both fatigue and job performance values ranged from 1.0 to 5.0 in increments of 0.2, resulting in a total of 441 unique fatigue & job performance combinations. For each user and each of these 441 combinations, we generated 100 synthetic observations, resulting in 44,100 simulation data points per user and a total of 10,363,500 (44,100 $\times$ 235) simulated data points.

In each simulation instance, as in the simulation study by Shin et al. [11], we applied the regression model from the empirical research of Hassan and Morsy, which describes the effect of fatigue on job performance [13]. Both variables were converted to a scale from 1 to 5, and the following transformation was then used to calculate the final job performance score.

$$\textit{Final Job Performance} = \max\left(\textit{Job Performance} - 0.34 \times \textit{Fatigue},\ 1\right)$$

This formulation reflects the assumption that increased fatigue impairs job performance. The resulting value was bounded below by 1.0 to reflect a realistic minimum performance floor. We then computed a probabilistic decision threshold based on the normalized final performance:

$$\textit{Threshold} = \max\left(\left(\frac{\textit{Final Job Performance} - 1}{4}\right) \times 100,\ 0\right)$$

This threshold represents the probability (in percentage) that a user successfully recalls and applies their previous phishing training, as described by Shin et al. [11]. In every simulation tick, for each end user agent, a random integer from 0 to 100 was sampled. If this value was less than the computed threshold, the simulation applied PSAE, assuming the user retained the training effect. Otherwise, it used PSBE, modeling a lapse in training recall. Thus, the final phishing susceptibility score for each scenario was determined as a binary stochastic outcome based on the user's dynamic state.

Each row of the final dataset contains the static user characteristics, the dynamic fatigue and job performance values, and the assigned phishing susceptibility label for that simulation instance. For each end user agent, phishing susceptibility was generated 100 times for every specific combination of fatigue and job performance according to the formula above, with a higher final job

performance leading to a higher frequency of selecting the PSAE value. This synthesized dataset constitutes a large-scale, high-resolution behavioral resource that captures the subtle interactions between static and dynamic human factors, and serves as the foundation for training the machine learning model.

Finally, the entire data-farmed dataset was divided into training and test sets. Specifically, 80% of the data, corresponding to 188 agents (about 8,290,800 data points), was used for training, while the remaining 20%, corresponding to 47 agents (about 2,072,700 data points), was used for testing.

4 Phishing Susceptibility Prediction Model

In this study, we designed and implemented a neural network model based on a multilayer perceptron (MLP) [4] to simulate and predict individual susceptibility to phishing attacks. Inspired by the neural architecture of the human brain, this model can process multiple input features in parallel and effectively capture complex, nonlinear relationships among variables [18]. These characteristics make the model particularly well-suited for accurately modeling the impact of human behavioral traits on security-related outcomes.

4.1 Input Variables

The model takes as input a total of nine human factor variables, categorized as follows:

Demographic factors: Age (21-56), Gender (0-1)
Personality traits: Openness (20-80), Conscientiousness (20-80), Extraversion (20-80), Agreeableness (20-80), Neuroticism (20-80)
Dynamic factors: Fatigue (1-5), Job Performance (1-5)

4.2 Model Architecture

The MLP model is implemented as a three-layer fully connected feedforward neural network [15], defined as follows:

Input layer: 9 input features
Hidden Layer 1: 64 neurons, ReLU activation [14]
Hidden Layer 2: 32 neurons, ReLU activation [14]
Output Layer: 1 neuron, Sigmoid activation [15] (to produce probability between 0 and 1)

4.3 Loss Function & Training

The model was implemented using the PyTorch deep learning framework [16]. It was trained using the Binary Cross Entropy Loss (BCELoss) [19], a standard choice for binary classification tasks where outputs represent probabilities. Optimization was handled by the Adam optimizer [17] with a learning rate of 0.001. Training was conducted over 10 epochs, using mini-batches of size 256, and the dataset was shuffled in each epoch to ensure robustness.

5 Results

To begin our results section, Figure 1 presents the phishing susceptibility prediction heatmap for a single end user agent. In the original test dataset, this user's pre-training phishing susceptibility (PSBE) is 0.149561, and the post-training value (PSAE) is 0.066252. Figure 1 illustrates how the trained neural network model responds to variations in two dynamic human factors: fatigue and job performance. Each grid cell in the heatmap represents the predicted phishing susceptibility for a specific combination of fatigue and job performance. The prediction values for this user range from 0.074821 to 0.151971, which is similar to the empirical range of the user's phishing susceptibility (PSBE to PSAE), indicating that the model provides generally reasonable predictions. In addition, the heatmap clearly shows that phishing susceptibility increases as fatigue increases and job performance decreases. In particular, this result demonstrates that, unlike the conventional method which relies on probabilistically switching between PSBE and PSAE values based on fatigue and job performance [11], the proposed model is able to continuously adjust the phishing susceptibility score in response to changes in these dynamic human factors. In other words, our neural network-based approach is significant in that it can flexibly respond to subtle variations in dynamic human factors, providing more precise and interpretable predictions of phishing susceptibility compared to previous methods.

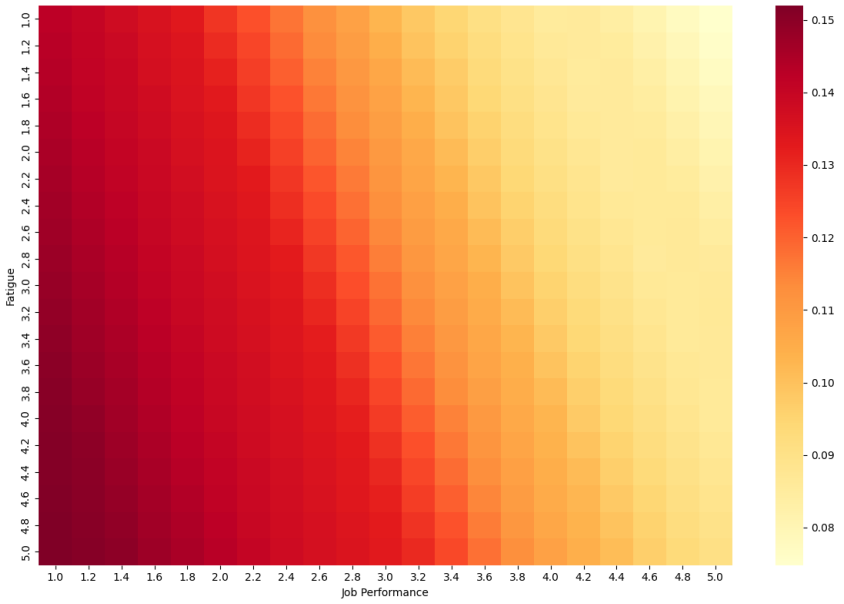

Fig. 1. Heatmap of predicted phishing susceptibility for a single end user agent.

Table 1. Prediction Performance of the Surrogate Model

Dataset	MSE	MAE	Pearson Correlation
Train	0.0002	0.0074	0.9976
Test	0.0021	0.0251	0.9862

Next, for all 235 end users in both the training and test sets, we generated phishing susceptibility predictions using our surrogate model for every possible combination of job performance and fatigue (441 in total). The model's prediction performance was evaluated by comparing the predicted phishing susceptibility scores to the actual values, where the actual value for each combination was computed as the mean of 100 simulation runs generated through data farming for each end user agent. Table 1 summarizes the prediction performance of the proposed surrogate model. The results demonstrate that the MLP-based surrogate accurately approximates the original simulation-based phishing susceptibility model, achieving a mean squared error (MSE) of 0.0002, a mean absolute error (MAE) of 0.0074, and a Pearson correlation coefficient of 0.9976 on the training set. On the test set, the MSE and MAE were 0.0021 and 0.0251, respectively, with a Pearson correlation of 0.9862. These results indicate that the surrogate model maintains high predictive accuracy and generalizes well to unseen data, as evidenced by the minimal drop in performance from the training to the test set.

Table 2. Average calculation time (in seconds) and standard deviation for the original simulation method versus the MLP-based surrogate method in varying number of N.

N	Original Method		Surrogate Method	
	Avg Time	Std Dev	Avg Time	Std Dev
1	0.18	0.038	4.07	0.11
10	1.47	0.065	4.71	0.13
50	6.69	0.13	6.54	0.22
100	15.43	0.36	14.79	0.16
200	27.34	0.53	22.21	0.12
500	67.63	1.82	38.75	0.35
1000	135.04	2.90	72.91	1.27
10000	1436.68	18.31	734.58	6.46

Lastly, we measured the computation time required to update phishing susceptibility for each end user agent at every tick (second) over a simulated day (86,400 ticks) in the simulation environment. Two approaches were compared: the original method, which computes phishing susceptibility through a series of equations and conditional statements based on fatigue and job performance for each agent, and our proposed MLP-based surrogate model, which instantly predicts phishing susceptibility using the trained neural network. All MLP-based surrogate model predictions were executed on the CPU, and all experiments were

conducted on a DELL computer equipped with an Intel Core Ultra 7 Processor and 32 GB of RAM.

Table 2 summarizes the results of our experiments. For simulations involving small and medium-sized organizations ($N < 50$), the original method achieved faster computation times when calculating phishing susceptibility for all end user agents. This is likely because, for small N, the surrogate method incurs a fixed overhead for loading the neural network into memory, converting the input data into tensors, and initializing the inference context, which outweighs any gains from parallelized matrix operations. In contrast, the original method relies solely on simple arithmetic and conditional checks, so it remains faster when the number of agents is low. Starting from $N = 50$, however, the surrogate method increasingly outperformed the original method in terms of computation time. As N increased, the computational efficiency gap between the two methods widened, and for $N = 10,000$, the surrogate method was nearly twice as fast as the original method. This demonstrates the superior scalability and computational advantage of the MLP-based surrogate approach in large-scale simulation settings.

6 Discussion and Conclusion

In this study, we proposed a surrogate approach based on an MLP neural network to significantly improve the computation speed of phishing susceptibility calculations for human agents in cybersecurity simulations, while incurring only minimal loss in accuracy. This method also enhances the overall scalability of the simulation. Furthermore, our approach advances beyond previous methods that determined phishing susceptibility by probabilistically switching between PSBE and PSAE based on job performance and fatigue values. Instead, our model continuously and responsively adjusts phishing susceptibility, producing more realistic outputs as fatigue decreases and job performance increases.

The experimental results show that, although the surrogate model was slower than the original method for small values of N, a clear break-even point appeared around $N = 50$, after which the surrogate method became increasingly faster. Even with a relatively simple phishing susceptibility prediction model that considers only two dynamic human factors, fatigue and job performance, the surrogate model achieved nearly a twofold increase in computation speed compared to the original method for $N = 10,000$. This suggests that, when the surrogate approach is applied to more sophisticated models incorporating a greater variety of static and dynamic human factors and their complex interactions, even greater improvements in scalability can be expected. Additionally, since our experiments ran the surrogate model using only the CPU, further speed improvements are likely when employing a GPU. These results highlight the potential of the surrogate modeling approach to serve as a scalable, generalizable tool for a wide range of human factor–driven simulation studies in cybersecurity.

For future research, we plan to incorporate additional dynamic human factors that are likely to influence phishing susceptibility, such as stress level, social media usage patterns, emotional state, and perceived vulnerability. Including

these variables will allow the model to better capture the complexity of real-world human behavior, though it may also increase the overall complexity of the model. If the current neural network becomes insufficient to maintain high predictive accuracy under these more complex conditions, we intend to explore and compare more sophisticated neural network architectures to identify the most suitable structure for real-time phishing susceptibility prediction and for improving the scalability of cybersecurity simulations.

Finally, we plan to conduct further research using realistic cybersecurity simulation frameworks such as OSIRIS [20], which integrate diverse human factors [11], organizational culture [3], social networks, and a range of system and human vulnerabilities [21]. Through these platforms, we aim to investigate how the proposed surrogate approach can be effectively applied to simulation modules beyond phishing susceptibility prediction, and how it can further improve the computational speed and scalability of large-scale cybersecurity simulations.

Acknowledgments

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by the Minerva Research Initiative under Grant #N00014-21-1-4012 and by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the US Government.

References

1. IBM: IBM security services 2014 cyber security intelligence index (2014)
2. Tornblad, McKenna K., et al. "Characteristics that predict phishing susceptibility: a review." Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 65. No. 1. Sage CA: Los Angeles, CA: SAGE Publications, 2021.
3. Shin, Jeongkeun., et al. "Simulating Cyber Defense: The Impact of Phishing Training and System Updates on Mitigating Damage from Hybrid Phishing and Watering Hole Attacks" The Journal of Defense Modeling and Simulation, Forthcoming.
4. Rumelhart, David E., James L. McClelland, and PDP Research Group. Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations. The MIT press, 1986.
5. Alizadeh, Reza, Janet K. Allen, and Farrokh Mistree. "Managing computational complexity using surrogate models: a critical review." Research in Engineering Design 31.3 (2020): 275-298.
6. Eftimie, Sergiu, Radu Moinescu, and Ciprian Răcuciu. "Spear-phishing susceptibility stemming from personality traits." IEEE Access 10 (2022): 73548-73561.
7. Ribeiro, Liliana, Inês Sousa Guedes, and Carla Sofia Cardoso. "Which factors predict susceptibility to phishing? An empirical study." Computers & Security 136 (2024): 103558.

8. Burns, A. J., et al. "Organizational information security as a complex adaptive system: insights from three agent-based models." Information Systems Frontiers 19 (2017): 509-524.
9. Shin, Jeongkeun, L. Richard Carley, and Kathleen M. Carley. "Simulation-Based Study on False Alarms in Intrusion Detection Systems for Organizations Facing Dual Phishing and DoS Attacks." 2024 Annual Modeling and Simulation Conference (ANNSIM). IEEE, 2024.
10. Shin, Jeongkeun, et al. "Design, Modeling and Simulation of Cybercriminal Personality-Based Cyberattack Campaigns." 2024 Winter Simulation Conference (WSC). IEEE, 2024.
11. Shin, Jeongkeun, Richard Carley, and Kathleen Carley. "Simulation of Human Organizations with Computational Human Factors Against Phishing Campaigns." International Conference on Cyber Warfare and Security. Academic Conferences International Limited, 2025.
12. Sanchez, Susan M. "Data farming: Methods for the present, opportunities for the future." ACM Transactions on Modeling and Computer Simulation (TOMACS) 30.4 (2020): 1-30.
13. Kamal Hassan, Sabreen Mohammed, and Sahar Mohamed Morsy. "Effect of Work Conditions and Fatigue on Job performance of Staff Nurse's at Al Eman General Hospital." Assiut Scientific Nursing Journal 11.34 (2023): 317-327.
14. Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." Proceedings of the 27th international conference on machine learning (ICML-10). 2010.
15. Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." nature 323.6088 (1986): 533-536.
16. Paszke, A. "Pytorch: An imperative style, high-performance deep learning library." arXiv preprint arXiv:1912.01703 (2019).
17. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
18. Rosenblatt, Frank. "The perceptron: a probabilistic model for information storage and organization in the brain." Psychological review 65.6 (1958): 386.
19. Bengio, Yoshua, Ian Goodfellow, and Aaron Courville. Deep learning. Vol. 1. Cambridge, MA, USA: MIT press, 2017.
20. Shin, Jeongkeun, et al. "OSIRIS: organization simulation in response to intrusion strategies." International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Cham: Springer International Publishing, 2022.
21. Shin, Jeongkeun, et al. "Revelation of System and Human Vulnerabilities Across MITRE ATT&CK Techniques with Insights from ChatGPT." CASOS technical report (2023) (2023).