Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 2884769, 10 pages
https://doi.org/10.1155/2022/2884769

Research Article

Prediction of Phishing Susceptibility Based on a Combination of Static and Dynamic Features

Rundong Yang,[1] **Kangfeng Zheng**,[1] **Bin Wu**,[1] **Chunhua Wu**,[1] **and Xiujuan Wang**[2]

[1]*School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*School of Computer Science, Beijing University of Technology, Beijing 100124, China*

Correspondence should be addressed to Kangfeng Zheng; kfzheng@bupt.edu.cn

Received 13 March 2022; Accepted 13 April 2022; Published 10 May 2022

Academic Editor: Man Fai Leung

Phishing is a very serious security problem that poses a huge threat to the average user. Research on phishing prevention is attracting increasing attention. The root cause of the threat of phishing is that phishing can still succeed even when anti-phishing tools are utilized, which is due to the inability of users to correctly identify phishing attacks. Current research on phishing focuses on examining the static characteristics of the phishing behavior phenomenon, which cannot truly predict a user's susceptibility to phishing. In this paper, a user phishing susceptibility prediction model (DSM) that is based on a combination of dynamic and static features is proposed. The model investigates how the user's static feature factors (experience, demographics, and knowledge) and dynamic feature factors (design changes and eye tracking) affect susceptibility. A hybrid Long Short-Term Memory (LSTM) and LightGBM prediction model is designed to predict user susceptibility. Finally, we evaluate the prediction performance of the DSM by conducting a questionnaire survey of 1150 volunteers and an eye-tracking experiment on 50 volunteers. According to the experimental results, the correct prediction rate of the DSM is higher than that for individual feature prediction, which reached 92.34%. These research experiments demonstrate the effectiveness of the DSM in predicting users' susceptibility to phishing using a combination of static and dynamic features.

1. Introduction

With the continuous development of the Internet and the increasing popularity of electronic payments, the problem of online fraud has gradually become a focus of attention. Phishing refers to a category of fraudulent behaviors in which attackers use social engineering techniques to guide users to visit fake websites that appear to be real through SMSs, emails, fake advertisements, and live chat tools, among other ways, to fraudulently obtain users' private information, such as private account passwords, payment passwords, and credit card information.

Phishing is currently one of the biggest threats in cybersecurity and is widely used. Many expert researchers and government authorities give great attention to the phishing problem and expect the phishing threat to become more serious in the future [1, 2]. The reason for this phishing threat is simple: many phishing incidents succeed because users are unable to recognize phishing [3]. Historically, many disruptive security attacks have occurred, and one of the more threatening security attacks was a phishing security attack called the Locky ransomware attack [4]. When a user with enough access to an organization's server opens a phishing e-mail from an attacker, the attacker extorts the organization by obtaining a large amount of encrypted data. Researchers in fields that are related to cybersecurity [5, 6] and in governments [1, 7] have warned businesses and individual users about the growing phishing threat.

If anti-phishing tools and phishing detection techniques alone cannot effectively prevent phishing, the user is the ultimate defense against phishing [8]. According to an APWG report, 1,520,832 phishing websites and 1,031,347 phishing emails were detected in 2020, and the number of phishing websites peaked in October 2020, with 369,254 phishing attacks occurring in January alone; phishing activity is still at an all-time high [9].

Several recent studies have found that Internet users are unable to effectively distinguish legitimate sites from phishing sites and are unable to avoid transactions on phishing sites [10–12]; summarizing previous work, studies have found that 40–80% of users are unable to correctly identify phishing sites [10, 13, 14] and more than 70% of transactions are made on phishing sites [10, 11]. Many studies have used anti-phishing tools to prevent users from visiting phishing websites, which mainly include web browser security toolbars and plug-ins [15–17]. Although these tools are highly accurate in detecting phishing websites, phishing still has a high success rate, and these software tools fail to attract users' attention or warnings are ignored [18]. Internet users always ignore the warnings of anti-phishing tools because users believe that the warnings are not directed to them [19].

Internet users are unable to recognize the threat of phishing, and they click on links because they lack security awareness, which depends mainly on their static awareness. Static awareness is formed by a user before performing a security action, so it does not correctly reflect the real situation of the user at the time of performing the action [12, 20, 21].

The central issue for phishing prevention and detection is security awareness. Is the user thinking through the process of performing actions when he or she is subjected to a phishing attack? To solve this problem, we mainly consider the influence of personal characteristics (e.g., demographic characteristics, personality traits, experience, and knowledge) and e-mail characteristics (e.g., sender address and outgoing connections) on the person [22].

Static awareness cannot be addressed in the case where the user is performing the action and the static role is separated. Since human behavior when subjected to phishing attacks is determined by the interactions with human perceptions of phishing attacks, awareness in this interactive scenario is called dynamic awareness. Dynamic awareness, which is derived from situational awareness, is represented by a cyclical model (PCM) [15], which considers both knowledge states and processes.

In contrast to previous studies that treat static aspects of consciousness as separate from the actual situation, we adopt an approach that is based on the assumption that safety-related behaviors are generated by the interaction between the person and the perception of the situation. Thus, our basic premise is that security-related behaviors are based on the context of the environment in which they occur and require an assessment of the encountered situation. To conceptualize the actual awareness of individuals in security-related situations, we introduce the construct of situational information security awareness. It is derived from the perceptual cycle model (PCM) of situational awareness [15], which considers both products (knowledge states) and processes (how knowledge is created through intelligent interactions).

The approach that is used in this study differs from previous approaches in that, instead of predicting a single phishing attack, we predict the user's phishing susceptibility. In this paper, we define the user's susceptibility as the degree of interaction between the user and the phishing attack. Our approach can truly reflect the user's susceptibility level, analyze the factors of susceptibility, and provide personalized warning messages to the user.

We propose a hybrid phishing susceptibility prediction model that is based on static and dynamic features. The model integrates the key features of static and dynamic features. These static features mainly include individual-level features such as experience, personal attributes, and other features. To obtain the static features, we use questionnaires to collect the personal static features of 1150 volunteers. In addition, to more realistically reflect the real situation of users facing phishing, it is necessary to obtain the dynamic features of users in the scenario. We conduct eye-tracking experiments and questionnaires on 50 volunteers. We use a hybrid model for static features, we mainly use an LSTM model for feature extraction, and we apply the LightGBM model for dynamic feature prediction, in which the LightGBM algorithm is utilized to predict the user's susceptibility. Finally, after the prediction stage, the prediction results of the two models are combined. The combined model integrates the respective characteristics of the two models, which can not only explore the intrinsic connections between the time-series data but also avoid the influence of discontinuous features on the prediction results. The test results show that the combined model can realize lower error than a single model in special scenarios and has more stable prediction results.

Although phishing has been studied for many years, predicting users' susceptibility to phishing is a new challenge in proactive defense against phishing attacks. Many experts and scholars have emphasized that anti-phishing tools are not effective in preventing users from being phished [23–25]. Therefore, according to the scientific guidelines [26], our work can be regarded as an improvement to previous work, and the main contributions of this paper are as follows:

(1) Previous studies have not analyzed the prediction of users' susceptibility to phishing emails but have focused on developing or testing behavioral models, and the present model demonstrates the feasibility of susceptibility prediction.

(2) Previous phishing studies have analyzed user susceptibility with a static set of personal characteristics in the model. The present model combines static personal characteristics while taking into account interactivity and incorporates eye movement by using eye movement data as a dynamic response to threat perception for dynamic feature extraction. This model yields more accurate prediction results.

The remainder of the paper is organized as follows: Section 2 reviews the series of phishing susceptibility studies that have been conducted, Section 3 discusses the proposed work, Section 4 presents the detailed results of the data analysis and discusses the limitations of the model, and Section 5 discusses and summarizes the full paper.

2. Related Work

Traditional anti-phishing efforts have focused on designing and developing anti-phishing tools [26] and designing better algorithms to improve the accuracy of detection [27]. Despite researching anti-phishing techniques, phishing is still able to attack successfully, and phishing attacks have become one of the most threatening digging attacks for network security. To solve the problem of phishing attacks, more and more researchers have started to shift the focus of their research to user susceptibility analysis.

In recent years, a survey of Dutch cybercrime victims [28] has been conducted to determine which user behaviors increase the risk of being phished, using a multivariate risk analysis approach. The results of the analysis were used to determine several behaviors that increase the user's risk.

Leukfeldt [28] used risk perception theory, theory of planned behavior, and decision theory to construct a model to analyze the factors influencing spear phishing attacks on users. The experiments were conducted mainly in Middle Eastern countries, and the final results analyzed the behavioral factors that increase cyber network susceptibility and help to evaluate and select the use of phishing tools.

According to the research findings [20], some users are more susceptible to phishing attacks due to attack scenarios and personality factors, and experiments were conducted through individuals and companies to analyze how personality traits affect human behavior.

Lin et al. [29] investigated the factors affecting users' susceptibility by conducting a study on phishing emails and analyzing the impact of psychological, demographic, and cultural characteristics on susceptibility to phishing, and the experimental results showed that age had the greatest impact on phishing.

Luo et al. [30] designed a heuristic system model to investigate the impact of users' psychology and behavior on phishing susceptibility through a qualitative study. The model can explain the factors of user number victimization.

Lillo et al. [31] showed that the factors influencing susceptibility to phishing include demographics, knowledge, experience, and self-efficacy, which affect the user's access to the browser and transactional behavior on the phishing site.

Sheng et al. [32] proposed a phishing susceptibility assessment model (DRKM) that analyzes multiple demographic variables including age, education, and gender as well as knowledge and experience aspects mainly considering awareness of phishing, technical skills, etc. Based on the analysis of the model, the final implementation results showed that the susceptibility of users to phishing can be effectively predicted by gender, age, and risk propensity.

Abbasi et al. [33] proposed a phishing funnel model (PFM) for phishing susceptibility analysis, which divides the user's susceptibility into multiple stages and predicts the user's susceptibility by performing susceptibility analysis for each stage separately.

Yang et al. [34] proposed a model for phishing susceptibility analysis, which mainly extracts features of pairs of dimensions, including demographics, personality, knowledge experience, and computer knowledge, and uses these features to predict the susceptibility of users.

3. Proposed Method

In this section, we propose a user phishing susceptibility prediction model that is based on static and dynamic features. Based on the combined LSTM and LightGBM model, the prediction of phishing susceptibility using static features and dynamic features is realized. A flowchart of the model is presented in Figure 1.

The model has three main components. The first component obtains data for preprocessing. A lot of these static data are obtained using a normalization method. The original static features and dynamic features are obtained for preprocessing. The converted data, which contain a variety of factors that affect susceptibility, are conducive to the prediction of user susceptibility. Then, we design an LSTM model, set the LSTM parameters, use the LSTM model to predict susceptibility for static features, and use the LightGBM model to predict susceptibility for dynamic features. Finally, the prediction of user phishing susceptibility is realized by combining the prediction results of the static susceptibility prediction model and the dynamic susceptibility prediction model.

3.1. Data Preprocessing. When the LSTM and LightGBM hybrid model is applied for network phishing susceptibility prediction, the input feature vectors of the component models consist of questionnaires and eye-tracking experiment data, and feature values with different magnitudes and large differences in values are obtained. According to the characteristics of the network model, the data are directly input into the model because the weighted accumulator data will become abnormally large, thereby resulting in the network failing to converge; therefore, the input data vector needs to be normalized before the data are input. The commonly used method is the min–max normalization method because it is easy to use and fast, so it is applied in this paper, and the formula is presented below.

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{1}$$

In the above equation, $x_{\max}$ is the maximum value of the sample data, and $x_{\min}$ is the minimum value of the sample data.

In this paper, the data obtained by questionnaires and eye tracking, in which the input and output values vary widely, are normalized using the min–max normalization method, which can reduce the error of the model and map the data into the interval [0, 1]. The data are obtained from static features such as personal attributes (e.g., gender, age, income, experience, and knowledge) and dynamic feature data (e.g., the time of eye gaze and the number of gazes during user interaction) through questionnaires and eye tracking. Table 1 presents the acquired data.

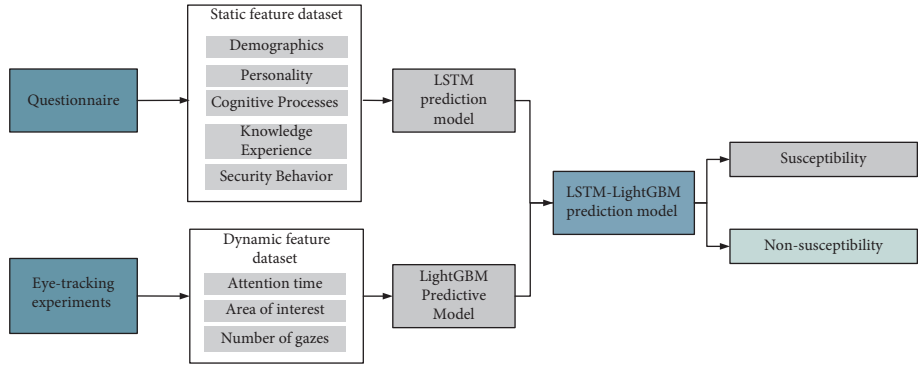

Figure 1: DSM flow.

Normalization is performed using the extreme values via (1) for the above data, and the normalized data are presented in Table 2.

3.2. LSTM. An LSTM is a special RNN network. RNNs have the problems of gradient disappearance and gradient explosion. An LSTM, through a cell gate switch to achieve temporal memory function and prevent gradient disappearance, can learn the long-term dependence on information, preserve the information, and overcome the disadvantages of RNNs [22]. Its algorithm structure is illustrated in Figure 2. The current loss x^t of an LSTM and the last state that is passed down h^{t-1} are used to obtain four states by splicing training, and the state formula is as follows:

$$z = \tanh\left(w_h^{(z/2)t-1}\right), \tag{2}$$

$$z^i = \sigma\left(w_h^{t^{t^t}}\right), \tag{3}$$

$$z^f = \sigma\left(wx^t h^{t-1}\right), \tag{4}$$

$$z^\rho = \sigma\left(w_h^{a^{x^t}}\right), \tag{5}$$

z^i, z^f, and z^0 in (2) and (3) are converted to values between 0 and 1 by a sigmoid activation function after multiplying the splicing vector by the weight matrix as a gating state. In (2), z is the result converted to a value between −1 and 1 by the tanh activation function. $\odot$ denotes multiplication of the corresponding elements in the operation matrix, which requires that the two multiplied matrices be of the same type [12]. + denotes matrix addition. c^t, h^t, and y^t are calculated as follows:

$$c^t = z^f \odot c^{t-1} + z^i \odot z, \tag{6}$$

$$h^t = z^0 \odot \tanh\left(c^t\right), \tag{7}$$

$$y^t = \sigma\left(W^f h^t\right). \tag{8}$$

The computational process of the LSTM model is divided into three main phases: the forgetting phase, the input phase, and output phase. In the forgetting phase, the previously unused information is forgotten. The forgetting gate decides which information will be forgotten based on the input x^t of the current node, the state c^{t-1} of the previous node, and the output h^{t-1} of the previous node. The input phase determines which information will be left behind for the current

Table 1: Collected data.

ID	Age (years)	Annual income (yuan)	Duration of gaze (s)
1	23	100,000	0.3
2	24	50,000	0.0087
3	54	80,000	0.2
4	12	300,000	0.0374
5	38	200,000	0.3551
6	21	400,000	0.008

Table 2: Data normalization.

ID	Age (years)	Annual income (yuan)	Duration of gaze (s)
1	0.26	0.11	0.33
2	0.28	0.04	0.00
3	0.85	0.08	0.22
4	0.06	0.36	0.03
5	0.55	0.24	0.39
6	0.23	0.49	0.00

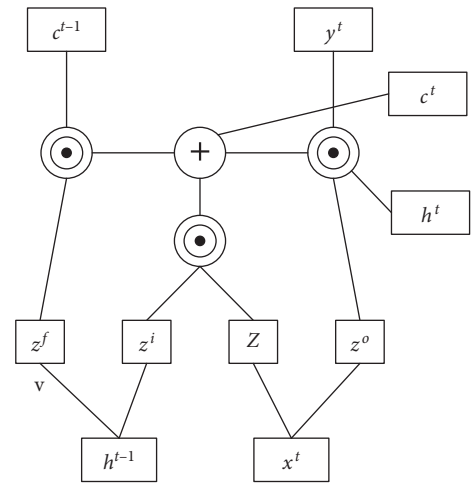

Figure 2: LSTM structure.

input data. The main memory selection is made for the input x^t. The current input content is expressed by (2).

z is derived, and the selection gating signal is controlled by z^i. From (2) to (4), the c^t value that is transmitted to the next state is obtained by (5). The output stage determines the output value. After obtaining the latest node state c^t, the LSTM combines the output h^{t-1} of the previous node and the input x^t of the current node to determine the output y^t of the current node, which is obtained by changing h^t in (6) and (7). z^o in (6) is used to perform control and scaling of the state c^t (which is transformed by the tanh activation function) [12].

3.3. LightGBM. LightGBM is a distributed gradient boosting GBDT framework that is based on a decision tree algorithm, and the GBDT algorithm faces substantial challenges in terms of performance and accuracy in a data environment with large training samples and high-dimensional features. To overcome these problems, LightGBM was developed. LightGBM has the features of fast training speed, low memory occupation, high accuracy, and support for parallelized learning, and it can handle large-scale data [11].

LightGBM mainly uses some optimization algorithms in the gradient algorithm.

(1) Gradient-based one-sided sampling algorithm (GOSS): LightGBM uses the GOSS algorithm to optimize the training sample sampling. The basic strategy of the GOSS algorithm is to first sort the training set data according to the gradient, apply a preset proportion, and keep the data samples with gradients that are higher than the proportion among all samples; the data samples with gradients that are lower than the proportion are not discarded directly but are sampled according to a sampling proportion. To compensate for the impact on the sample distribution, the GOSS algorithm calculates the information gained by multiplying the data with smaller gradients by a factor to amplify them. The algorithm can give more attention to the "undertrained" sample data when calculating the information gain.

(2) EFB (exclusive feature bundling) algorithm: The LightGBM algorithm not only optimizes the sampling of training samples by the GOSS algorithm but also performs feature extraction to further optimize the training speed of the model. In the algorithm, a table of nonzero-valued features can be created for each feature. By scanning the data in the table, the time complexity of creating the histogram can be effectively reduced.

(3) Histogram algorithm: LightGBM uses a histogram-based algorithm that discretizes continuous feature values into K integers, constructs a histogram of width K, traverses the training data, and counts the cumulative statistics of each discrete value in the histogram. In selecting the splitting points of the features, only the discrete values of the sorted histogram need to be traversed. The histogram algorithm reduces the computational cost and memory consumption.

(4) Grow-by-leaf (leaf-wise) algorithm: Most decision tree learning algorithms for tree generation use the grow-by-level (level-wise) strategy loss, as illustrated in Figure 3.

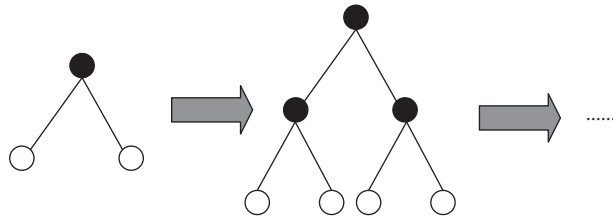

Figure 3: Level-wise strategy map.

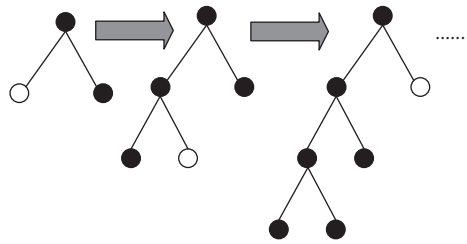

Figure 4: Leaf-wise strategy map.

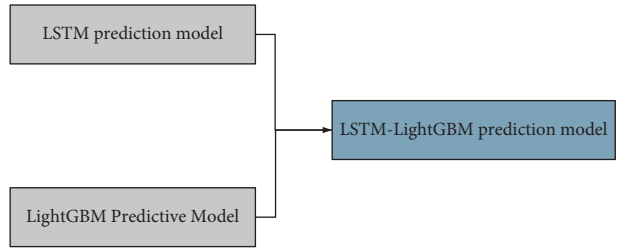

Figure 5: LSTM-LightGBM prediction model.

LightGBM uses a more efficient leaf-wise strategy algorithm, as illustrated in Figure 4. This strategy finds the leaf node with the largest splitting gain among all the leaf nodes of the current decision tree to split in each round. This mechanism reduces the splitting computation for the leaf nodes with lower gain. Compared with the level-wise strategy, the leaf-wise strategy can reduce the error and yield higher accuracy with the same number of splits. The disadvantage of the leaf-wise algorithm is that it may generate a deeper decision tree. Therefore, LightGBM adds a parameter to limit the maximum depth on the leaf-wise decision tree to prevent overfitting while ensuring the efficiency of the algorithm.

3.4. LSTM-LightGBM Model. Because the two models have different advantages for data processing, we linearly combine the two prediction results by applying a weighting factor α as follows:

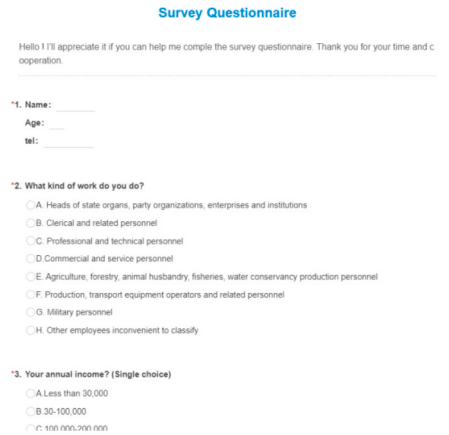

Figure 6: Questionnaire experiment.

(a) (b)

Figure 7: Experimental diagrams of eye tracking. (a) Trajectory diagram. (b) Heat diagram.

$$o = \alpha o_1 + (1 - \alpha) o_2, \tag{9}$$

where o_1 is the prediction probability of the LSTM model, o_2 is the prediction model of the LightGBM model, o is the final prediction result, and the value of α is determined by the final evaluation metric; namely, the best α value on the validation set is chosen. The structure of the combined model is illustrated in Figure 5.

4. Results and Discussion

To evaluate the effectiveness of the proposed model for predicting users' susceptibility to phishing based on a combination of static and dynamic features, the static feature dataset using LSTM and the dynamic feature dataset using LightGBM are analyzed and studied, and all experiments in this paper are conducted on a laptop with an Intel 8-core 2.8 GHz processor, 32 GB RAM, and a 1 TB hard disk using Python 3.6.

4.1. Experiment Setup and Dataset. In this study, to evaluate the effectiveness of the method that is proposed in this paper, the experimental dataset is mainly obtained by using questionnaires and eye-tracking methods. Two datasets are generated: a dynamic dataset and a static dataset. Static

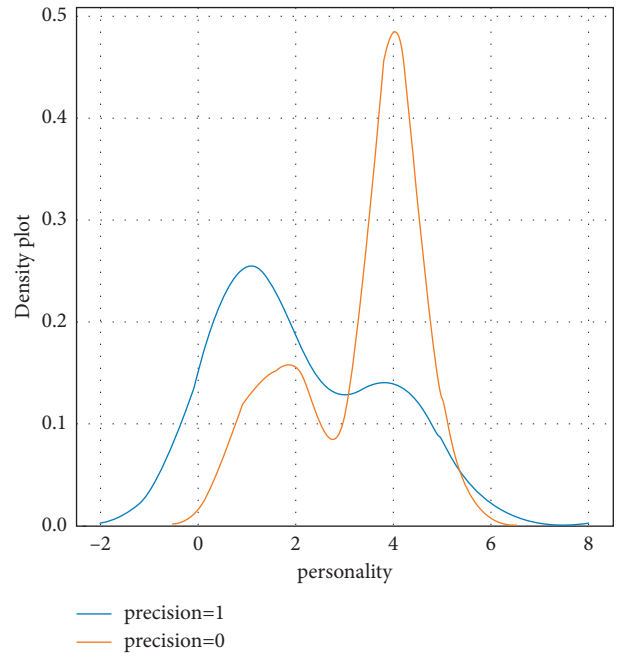

Figure 8: Distribution of susceptibility characteristics.

features are collected by using questionnaires; they consist mainly of users' characteristics, including name, gender, age, income, experience, and knowledge. Data are collected from 1150 volunteers. The susceptibility of users is judged by the phishing detection of users. One of the survey questionnaires is shown in Figure 6.

Fifty volunteers are selected from 1150 volunteers to conduct eye-tracking experiments to build a dynamic feature dataset, which is the user's interaction behavior when receiving phishing requests. The dynamic features are interaction behavior features, including gaze time, number of gazes, and area of interest. One of the eye-tracking experiments is shown in Figure 7. Because the testers are mostly Chinese, the phishing emails in the questionnaire and eye-tracking experiments are in Chinese, considering that volunteers will experience ambiguity when they are recognizing English.

In this experiment, two datasets are generated. For the static dataset, 19 characteristics, such as name, age, income, and knowledge, of 1150 volunteers are collected through 144 questions. Eighty percent of these are randomly selected as the training dataset, and the rest are selected as the test dataset. A dynamic feature dataset is obtained through an eye-tracking experiment on 50 people who are randomly selected from 1150 people for the experiment, and the dynamic features include six dimensions. Before the experiment, the data need to be preprocessed.

4.2. Feature Analysis. For static features, in this paper, 19-dimensional features, including gender, age, and knowledge, are constructed mainly from the data that are obtained from the questionnaire. To better make predictions, all the features are analyzed, and by analyzing the distributions of the categories and quantities of data, one can better understand the data and improve the model' correctness rate.

4.3. Analysis of Static Data. Personality is a very important characteristic that impacts the susceptibility of users; the distribution of personality characteristics in both categories is shown below. According to Figure 8, the neurotic personality is the most common in susceptible users, and the pleasant personality is the most common in nonsusceptible users.

Previous studies have shown that education level is an important factor that influences users' susceptibility, but according to the experimental results of this paper, education level has little effect on phishing susceptibility; the results are shown in Figure 9.

The cybersecurity knowledge scores of susceptible and nonsusceptible users are shown in Figure 10. The cybersecurity knowledge scores of users who are susceptible are distributed relatively evenly, but the scores of nonsusceptible users are mainly concentrated at the average level.

The distributions of other characteristic data, such as age, annual income, and gender, are shown in Table 3. The distribution of gender is relatively even. The number of users with annual incomes of less than 30,000 RMB is the highest, followed by the number of users with incomes of 30–100,000, and the number of users with incomes of more than 200,000 is the smallest. The ages are mainly distributed in the range of 20–30 years old, because the volunteers are mostly college students.

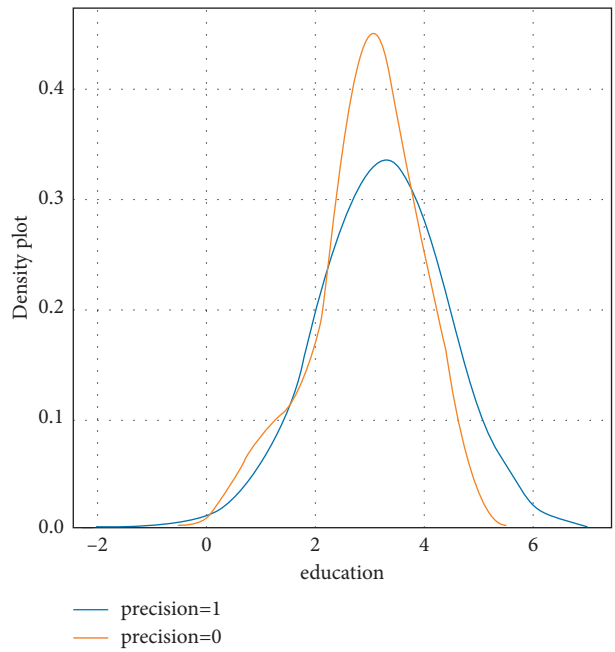

Figure 9: Distribution of educational level characteristics.

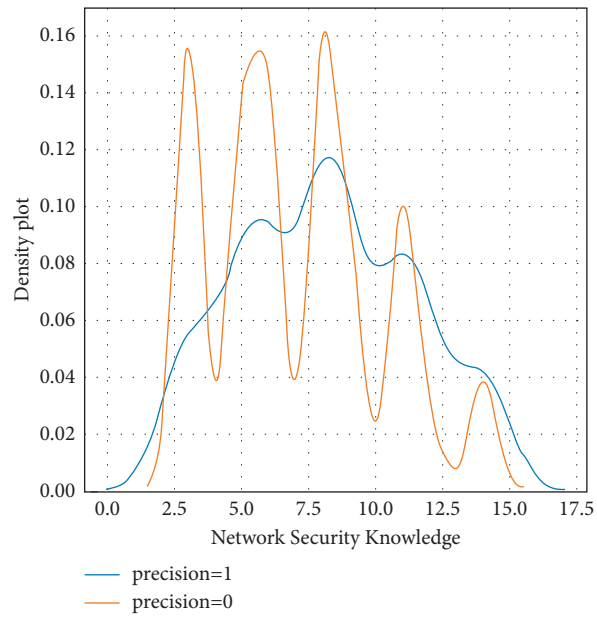

Figure 10: Network security knowledge feature distribution.

Experiment 1. Prediction of phishing susceptibility based on static features.

In this experiment, the LSMT model is mainly used to predict the susceptibility of users to network phishing using static features as input. It is also compared with several other machine learning models for analysis.

Table 3: Distributions of other characteristics.

Attribute	Feature	Category	Frequency
Demographics	Age	<20	67
		20–30	720
		30–40	107
		40–50	128
		>50	83
	Education level	Below high school	61
		Vocational high school/high school	109
		Undergraduate	610
		Graduate or above	325
	Gender	Male	555
		Female	550
	Annual income	<¥30,000	477
		¥30,000–¥100,000	369
		¥100,000–¥200,000	178
		>¥200,000	81

First, the input data are organized into a three-dimensional structure, as required for LSTM, which is denoted as (TrainX, SeqLen, Dim_in). The first dimension, TrainX, represents the corresponding sample; the second dimension, SeqLen, represents the sequence data (specified sequence length) that are collected for that sample; and the third dimension, Dim_in, represents the corresponding feature dimension of the static feature variables in the dataset of this paper. The dataset is divided into a training set and a test set according to this three-dimensional structure [20].

Second, the LSTM model is constructed. The model structure is represented as (Units, Input_Shape, Activation, Recurrent_dropout). Units is the number of neurons in the hidden layer, Input_Shape is the structural form of the input dataset, Activation is the activation function, and Recurrent_dropout is the learning rate. The sample size and characteristics of the dataset are analyzed to obtain the best model values. In this paper, Units is 45; Activation is the activation function, which is "ReLU"; and Recurrent_dropout is 0.01.

Third, the LSTM model is trained. The training model structure is (X_Train, Y_Train, Epochs, Batch_Size, Validation_Split). X_Train and Y_Train are the model training data; Epochs is the number of iterations; Batch_Size is the number of batch samples; and Validation_Split is the training validation set splitting ratio. The values of the training model parameters are based on a priori knowledge and many experiments. In this paper, Epochs is set to l00, Batch_Size to l6, and Validation_Split to 0.8.

To compare and analyze the static feature prediction models, a variety of machine learning models are selected, such as RF [35], SVM [36], and KNN [37], and their experimental results are presented in Table 4.

In this experiment, the main dataset used is a static feature dataset for the analysis of users' phishing susceptibility keys. The LSTM static feature prediction model was designed for comparison with other machine learning models. According to the experimental results, LSTM obtains the highest correctness rate of 89.71%, which is better than the rates of other models.

Table 4: Evaluation of LSTM on special static datasets with various classifiers.

Classifier	Accuracy (%)	Precision (%)	Sensitivity (%)	F-measure (%)
RF	87.00	89.00	88.00	87.98
SVM	88.04	88.42	88.23	88.25
KNN	88.59	83.31	85.88	84.26
LSTM	89.71	89.49	88.90	90.15

Experiment 2. Dynamic feature network-based phishing susceptibility prediction.

In this experiment, the main objective is to predict the user network phishing susceptibility using dynamic features as input to the LightGBM model. This model is also compared with several other machine learning models for analysis.

The dynamic feature dataset is divided into training and validation sets, and the scaling factor between the training and validation sets is set to 0.8. The parameters of the model are set based on a priori knowledge and many experiments, and the core parameters of the LightGBM training model are set as follows for the experimental data.

objective: task type. The task type options are regression, binary, and multiclass, among others. In this paper, the task is to make predictions, and the type is set to regression.

num_leaves: the number of leaf nodes. This parameter determines the complexity of the tree model. The larger it is, the more accurate the model is; however, larger values may lead to overfitting. This parameter is set to 120.

max_depth: controlling the maximum depth of the tree. This parameter can explicitly limit the depth of the tree. It is generally set to a value no greater than log2(num_leaves) and is set to 7 in this paper.

min _data_in_leaf: the minimum number of samples per leaf node. It is an important parameter for dealing with overfitting of the leaf-wise tree. By setting it to a larger value, generation of an overly deep tree can be avoided, but a larger value may also lead to underfitting. In this paper, it is set to 16.

Table 5: Evaluation of LightGBM on the dynamic feature dataset with various classifiers.

Classifier	Accuracy (%)	Precision (%)	Sensitivity (%)	*F*-measure (%)
DT	82.66	84.66	83.66	83.64
LR	83.70	84.08	83.89	83.91
XGBoost	84.25	78.97	81.54	79.92
LightGBM	85.37	85.15	84.56	85.81

Table 6: Evaluation of LSTM-LightGBM on the feature dataset with various classifiers.

Classifier	Accuracy (%)	Precision (%)	Sensitivity (%)	*F*-measure (%)
LSTM	89.71	89.49	88.90	90.15
LightGBM	85.37	85.15	84.56	85.81
LSTM-LightGBM	92.34	91.26	92.29	92.26

learning_rate: the learning rate of the training model. A larger learning rate will accelerate the convergence but reduce the accuracy, and the default value is 0.1. The learning rate can be adjusted according to the size of the dataset and is set to 0.05 in this paper.

Experiments are conducted on dynamic datasets to predict susceptibility using the LightGBM model, and a performance comparison is conducted with several other machine learning models, such as DT [38], LR [39], and XGBoost [40]. The experimental results are presented in Table 5.

In this experiment, the dataset is a dynamic feature dataset, which is used to analyze the users' susceptibility to phishing. The designed dynamic feature prediction model, namely, LightGBM, is compared with other machine learning models. According to the experimental results, LightGBM obtains the highest correctness rate of 85.37% among the compared models.

Experiment 3. Susceptibility prediction of phishing based on hybrid features.

In the experiments in this section, users' susceptibility to phishing is predicted using a combined LSTM-LightGBM model, in which the static feature variables personal attributes, personality, knowledge, and experience data are input into the LSTM training model, and each static feature variable can be calculated to predict users' susceptibility to phishing. The dynamic feature multidimensional prediction values are fed into the LightGBM model as input variables, from which the user's phishing susceptibility prediction values can be derived.

In this paper, a static feature dataset and dynamic feature dataset are selected as test samples, and the test results of the LSTM model alone, LightGBM model alone, and combined model are presented in Table 6. According to Table 6, the combined model has the highest correctness value and the best result compared with the stand-alone models; not only is it sensitive to timing, but it can also handle large batch data to effectively predict the phishing susceptibility.

5. Conclusions

In this paper, based on static and dynamic data of users, a prediction algorithm was used to predict the susceptibility analysis of users to phishing. First, susceptibility prediction was performed using an LSTM model for 19-dimensional features, such as the annual income, occupation, age, knowledge, and experience of users. Then, we used the LightGBM model to predict the susceptibility using dynamic features. Finally, we used a hybrid LSTM-LightGBM model to predict the susceptibility of users to phishing. By comparing the susceptibility prediction results that were obtained using the LSTM model alone and the combined LSTM-LightGBM model, we concluded that the prediction accuracy of the combined LSTM-LightGBM model was higher, namely, 92.34%, and that its prediction results were closer to the real situation. In this paper, we combined dynamic features and static prediction of phishing susceptibility, and by predicting user susceptibility, we identified users who were susceptible to phishing, for whom a more secure phishing defense could be implemented.

Although the proposed model can predict the susceptibility of users, there are still several areas for improvement in future work. For example, the testers in this paper were mainly college students, and the sample data were not evenly distributed. More data on occupation and age groups can be collected in the future to improve the model's robustness and accuracy rate.

Data Availability

The processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded by the National Key R&D Program of China (2017YFB0802800) and Beijing Natural Science Foundation (4202002).

References

[1] K. Erb, *IRS Warns on Surge of New Email Phishing Scams*, 2018.

[2] M. Landewe, "Four phishing attack trends to look out for in 2019," 2019, https://www.forbes.com/sites/forbestechcouncil/2019/01/10/four-phishing-attack-trends-to-look-out-for-in-2019/#6b7a63d4ec20.

[3] J. Hong, "The state of phishing attacks," *Communications of the ACM*, vol. 55, no. 1, pp. 74–81, 2012.

[4] L. Mathews, "Massive ransomware attack unleashes 23 Million emails in 24 hours," 2017, https://www.forbes.com/sites/leemathews/2017/08/31/massive-ransomware-attack-unleashes-23-million-emails-in-24-hours/.

[5] M. Vergelis, T. Shcherbakova, T. Sidorina, and T. Kulikova, "Spam and phishing in 2018," *Secure List*, 2019, https://securelist.com/spam-and-phishing-in-2018/89701.

[6] R. Brewer, "Ransomware attacks: detection, prevention and cure," *Network Security*, vol. 2016, no. 9, pp. 5–9, 2016.

[7] A. Chanthadavong, "US, Canada issue alert on ransomware," 2017, http://www.zdnet.com/article/us-canada-issue-alert-on-ransomware/.

[8] J. Wang, Y. Li, and H. R. Rao, "Coping responses in phishing detection: an investigation of antecedents and consequences," *Information Systems Research*, vol. 28, no. 2, pp. 378–396, 2017.

[9] S. Grazioli and S. L. Jarvenpaa, "Perils of Internet fraud: an empirical investigation of deception and trust with experienced Internet consumers," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 30, no. 4, pp. 395–410, 2000.

[10] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Communications of the ACM*, vol. 50, no. 10, pp. 94–100, 2007.

[11] L. Li, E. Berki, M. Helenius, and S. Ovaska, "Towards a contingency approach with whitelist- and blacklist-based anti-phishing applications: what do usability tests indicate?" *Behaviour & Information Technology*, vol. 33, no. 11, pp. 1136–1147, 2014.

[12] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 581–590, Montréal, Canada, 2006.

[13] A. Herzberg and A. Jbara, "Security and identification indicators for browsers against spoofing and phishing attacks," *ACM Transactions on Internet Technology*, vol. 8, no. 4, pp. 1–36, 2008.

[14] L. Li and M. Helenius, "Usability evaluation of anti-phishing toolbars," *Journal in Computer Virology*, vol. 3, no. 2, pp. 163–184, 2007.

[15] A. Abbasi, F. Zahedi, and Y. Chen, "Impact of anti-phishing tool performance on attack success rates," in *Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics*, pp. 12–17, IEEE, Washington, DC, USA, June 2012.

[16] D. Zhang, Z. Yan, H. Jiang, and T. Kim, "A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites," *Information & Management*, vol. 51, no. 7, pp. 845–853, 2014.

[17] D. Akhawe and A. P. Felt, "Alice in warningland: a {Large-Scale} field study of browser security warning effectiveness," in *Proceedings of the 22nd USENIX Security Symposium (USENIX Security 13)*, pp. 257–272, Washington, DC, USA, August 2013.

[18] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," *MIS Quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012.

[19] M. Alsharnouby, F. Alaca, and S. Chiasson, "Why phishing still works: user strategies for combating phishing attacks," *International Journal of Human-Computer Studies*, vol. 82, pp. 69–82, 2015.

[20] G. D. Moody, D. F. Galletta, and B. K. Dunn, "Which phish get caught? an exploratory study of individuals' susceptibility to phishing," *European Journal of Information Systems*, vol. 26, no. 6, pp. 564–584, 2017.

[21] A. R. Dennis and R. K. Minas, "Security on autopilot: why current security theories hijack our thinking and lead us astray," *ACM SIGMIS-Data Base: The DATABASE for Advances in Information Systems*, vol. 49, no. SI, pp. 15–38, 2018.

[22] K. Smith, P. A. Hancock, and A. Peter, "Situation awareness is adaptive, externally directed consciousness," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, no. 1, pp. 137–148, 1995.

[23] L. Musthaler, *Security Analytics Will Be the Next Big Thing in IT Security. Network World (May 31)*, 2013.

[24] B. Taylor, *How Big Data Are Changing the Security Analytics Landscape. TechRepublic (January 2)*, 2014.

[25] S. Gregor, A. R. Hevner, and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Quarterly*, vol. 37, no. 2, pp. 337–355, 2013.

[26] Y. Zhang, S. Egelman, L. Cranor, and J. Hong, "Phinding phish:Evaluating anti-phishing tools," in *Proceedings of the 14th Annual Network and Distributed System Security Sympos*, pp. 1–16, San Diego, CA, USA, 2007.

[27] S. Li and R. Schmitz, "A novel anti-phishing framework based on honeypots," in *Proceedings of the 2009 eCrime Researchers Summit*, Tacoma, WA, USA, September 2009.

[28] E. R. Leukfeldt, "Phishing for suitable targets in The Netherlands: routine activity theory and phishing victimization," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 8, pp. 551–555, 2014.

[29] T. Lin, D. E. Capecci, D. M. Ellis et al., "Susceptibility to spear-phishing emails: effects of internet user demographics and email content," *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 5, pp. 1–28, 2019.

[30] X. Luo, W. Zhang, S. Burd, and A. Seazzu, "Investigating phishing victimization with the Heuristic-Systematic Model: a theoretical framework and an exploration," *Computers & Security*, vol. 38, pp. 28–38, 2013.

[31] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri, "Bridging the gap in computer security warnings: a mental model approach," *IEEE Security & Privacy Magazine*, vol. 9, no. 2, pp. 18–26, 2011.

[32] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "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*, pp. 373–382, ACM, New York, NY, USA, 2010.

[33] A. Abbasi, D. Dobolyi, A. Vance, and F. M. Zahedi, "The phishing funnel model: a design artifact to predict user susceptibility to phishing websites," *Information Systems Research*, vol. 32, no. 2, pp. 410–436, 2021.

[34] R. Yang, K. Zheng, B. Wu, D. Li, Z. Wang, and X. Wang, "Predicting user susceptibility to phishing based on multi-dimensional features," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 7058972, 11 pages, 2022.

[35] J. G. Bartlett, R. F. Breiman, L. A. Mandell, and T. M. File, "Community-acquired pneumonia in adults: guidelines for management," *Clinical Infectious Diseases*, vol. 26, no. 4, pp. 811–838, 1998.

[36] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004*, pp. 32–36, IEEE, Cambridge, UK, August 2004.

[37] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in *Proceedings of the OTM Confederated International Conferences On the Move to Meaningful Internet Systems*, pp. 986–996, Berlin, Heidelberg, Germany, 2003.

[38] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[39] A. DeMaris, "A tutorial in logistic regression," *Journal of Marriage and Family*, vol. 57, no. 4, pp. 956–968, 1995.

[40] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd acm SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco, CA, USA, August 2016.