EAI Endorsed Transactions
on Security and Safety
Research Article EAI.EU

Applying Machine Learning Techniques to Understand User Behaviors When Phishing Attacks Occur★

Yi Li[1], Kaiqi Xiong[1,*], and Xiangyang Li[2]

[1]University of South Florida, Tampa, Florida 33620, USA
[2]Johns Hopkins University, Baltimore, MD 21218, USA

Abstract

Emails have been widely used in our daily life. It is important to understand user behaviors regarding email security situation assessments. However, there are very challenging and limited studies on email user behaviors. To study user security-related behaviors, we design and investigate an email test platform to understand how users behave differently when they read emails, some of which are phishing. Specifically, we conduct two experimental studies, where participants take part in our experiments on site in a lab contained environment and online through Amazon Mechanical Turk that are referred to on-site study and online study, respectively. In the two experimental studies, we design questionnaires for the two studies and use a set of emails including phishing emails from the real world with some necessary modifications for personal information protection. Furthermore, we develop necessary software tools to collect experimental data include participants' basic background information, time measurement, mouse movement, and their answers to survey questions. Based on the collected data, we investigate what factors, such as intervention, phishing types, and an incentive mechanism, play a key role in user behaviors when phishing attacks occur. The difficulty of such investigation is due to the *qualitative* analysis of user behaviors and the limited number of data in the on-site study. For these reasons, we develop an approach to quantify user behavior metrics and reduce the number of user attributes by evaluating the significance of each attribute and analyzing the correlation of attributes. Moreover, we propose a machine learning framework, which contains attribute reduction, to find a critical point that classifies the performance of a participant into either 'good' or 'bad' through 10-fold cross-validation with randomly selected attributes cross-validation models. The proposed machine learning model can be used to predict the performance of a user based on the user profile. Our data analysis shows that intervention and an incentive mechanism play a significant role while phishing type I is more harmful to users compared to the other two types. The findings of this research can be used to help a user identify a phishing attack and prevent the user from being a victim of such an attack.

Received on 21 November 2019; accepted on 13 January 2020; published on 29 January 2020

Keywords: User behavior, phishing emails, machine learning, security attacks

doi:10.4108/eai.13-7-2018.162809

1. Introduction

Attackers usually send out phishing emails, which is an online identify theft, to deceive victims into providing their personal information and login credentials [2].

It is prevalent today since current growing Internet techniques heavily involve the sensitive information of users. Therefore, more and more personal computers and mobile device users are exposed to phishing attacks. Many researchers have studied phishing attack problems where many solutions have been proposed to detect phishing attacks at different levels [3–5]. However, there are only a very few studies on understanding how users' behavior can contribute to

★This paper is the extension version of the conference paper appeared in [1].

*Corresponding author. Email: xiongk@usf.edu

susceptibility to phishing. By understanding users' behaviors on phishing attacks, we can determine how to educate users so that they can better be prevented from phishing attacks.

Because of the non-homogeneity of users' network security education levels, users are susceptible to phishing attacks at different degrees [6]. Although security and usability experts claim that computer system should not rely on users' behavior, researchers found that phishing attack are directly correlated with user behavior factors [7]. Thus, an important security prevention method is to educate users to adapt better security behaviors, where user behavior education refers to teaching Internet users about phishing awareness and defense techniques. Education-based approaches usually offer online information or educational games [8, 9].

In this research, we aim at studying user behavior factors, such as intervention, phishing types and a monetary incentive, to understand how a user behaves during phishing email attacks and what mechanism may prevent a user from being a victim of such attacks. Our understanding of user behaviors will help us design a guideline to educate users how to identify phishing emails, thus reducing the chance to be a victim, although user education study is out of the scope of this study. Here, intervention is defined as a mechanism that helps users be aware of the phishing attacks more easily by modifying phishing types to make them appear more obvious [10]. A monetary incentive is introduced to motivate users to pay attention to phishing attacks [11].

Specifically, in our experiments, we recruit participants to conduct email sorting tasks. The emails used in the research consist of both phishing emails and non-phishing or normal emails. There are three kinds of phishing types in the phishing emails: (1) Suspicious sender's email address; (2) Suspicious links or attachments; (3) Malicious email contents. Performance of each participant, such as sorting correctness and time as well as mouse movement, is recorded in each experiment.

The goal of this study is first to understand how user behaviors are correlated to phishing victims through an analysis of the collected experimental data and then to develop a model to predict how likely a user will be a victim based on the user's profile and behaviors.

For this purpose, we explore to answer the following challenging questions in this research:

1. How intervention can affect user behaviors?
2. Which phishing type is more harmful than others?
3. How can a monetary incentive affect a user's behavior and sorting?
4. How accurately can we predict the performance of a user on email sorting based on user profiles and behaviors?

To answer the above questions, we propose two study designs, on-site study design and online study design. We start with an on-site study design that is carried out in a contained lab environment. In the lab, participants are asked to conduct a pre-setup experiment on our testbed, where each participant first read a number of emails and then sort them into either "phishing" or "non-phishing." We introduce a performance score to record the total number of the correctnesses of a participant's sorting.

In this research, our *first* main challenge is how to quantitatively answer the above questions. To address it, we first quantify user information and behaviors and then analyze the data obtained from participants' performance as well as participants' basic information from the questionnaires shown in Appendices A and B. The two questionnaires are designed for on-site and online, respectively, so they are not identical as their experimental environments are different. We also design a mouse tracking mechanism to trace their mouse movement. Particularly, this first challenge becomes very difficult to be addressed in the on-site study. This is because the number of experimental data is typically small in the on-site study. The small dataset constraint is due to the limitation of budget, resources, and participant diversity, resulting in the the limited number of people to be recruited. Actually, such limitations are very typical in many human subject studies. In this research, only 40 participants are recruited in the on-site study. Thus, our *second* main challenge to answer the above questions is how to extract useful information from a relatively small number of collected data to build a machine learning framework for predicting the performance score of each participant accurately.

Furthermore, to increase the diversity and scalability of recruitment, we design an online study through Amazon mechanical turk, where participants attends the study online. In the online study, we also collect the profile, performance and mouse movement of each participant similar to the case of the on-site study. Based on the collected data, we develop a comprehensive approach to building a machine learning framework for predicting participants' susceptibility to phishing. In order to better evaluate the performance of participants, we divide their performance into two classes, 'Good' and 'Poor,' based on their performance scores. Thus, it is important to setup a threshold, which we call *a critical point*, to divide participant performance scores into two classes. We evaluated the critical point in the online study to find the best division method. Our machine learning models are developed by a use of

the 10-fold cross-validation where we apply the similar idea of cross-validation to select attributes.

Our main contributions are summarized as follow:

1. We propose two study designs, on-site and online, to understand how a user behaves when phishing attacks occur and determine how we can help a user identify phishing attacks or prevent a user from being a victim of phishing attacks. The on-site study is conducted in a lab environment, while the online study is carried out online only. The on-site study is easily controlled as it is done in a contained environment, but recruiting a large number of participants become difficult. Conversely, the online study is easily scaled up, but the recruitment of online participants makes difficult to ensure the data and profile of participants to be worthiness.

2. To help users, we introduce intervention, which is a mechanism used in our study to help participants be aware of their weakness areas related to phishing. We specifically address the type of phishing attacks that they are unaware of and help them to recognize that type of phishing attacks. Furthermore, we introduce a monetary incentive to test how the incentive impacts participants' security decision making. To conduct the monetary incentive, we divide the participants into two groups, a control group and an incentive group.

3. Beside participants' basic background information, we develop software tools to collect experimental data including time measurement, mouse movement, and their answers to the survey questions that we carefully design for the above two study designs. The collected experimental data in our two study designs help us answer all the questions raised before.

4. To understand the collected data, we propose and develop a machine learning framework to predict the performance score of each participant based on his/her profile. The proposed machine learning framework consists of four different models; all of them are developed with a 10-fold cross-validation and cross-validation based feature selection. We also perform attribute reduction by analyzing the data obtained from participants' performance as well as the participants' basic information from the survey to select the best attributes for our machine learning framework.

5. In order to better evaluate the performance of participants, we introduce two classes of performance, Good and Poor, based on their performance scores. In this research, we find the best critical point to divide the participants' performance scores into the two classes by using collected experimental data, through the proposed machine learning model.

The rest of the paper is organized as follows. In the section 2, we present related work on phishing emails and why people fall for phishing. In the section 3, we introduce the designs of our two studies. In the section 4, we present the dataset and attributes as well as our machine learning framework. We evaluate the results of our studies in the section 5. Finally, we discuss the implications of our findings int the section 6.

2. Related Work

As phishing becomes a more and more popular attack vector, email has been the most common way to conduct phishing attacks [12–15]. In 2011, Vishwanath et al. [16] discovered that most phishing emails are peripherally processed and the decisions made by individuals are usually based on simple clues embedded in an email message. They also found that if the email contains urgent information, the user will typically ignore other clues that could potentially help detect the deception. Furthermore, these findings suggest that the users who have more experience with emails are more likely to be phished.

Based on Vishwanath et al.'s observation [16], Angela Sasse and Kirlappos [9] claimed that the direction of security awareness and training against phishing attacks needed to be changed. They argued that user education needed to focus on challenging and correcting the misconceptions that guide current user behaviors. To better understand user's perspective, decision-making strategies is an effective way of implementing security awareness applications.

Dhamija et al. [17] conducted an experiment for better understanding why phishing worked. They first analyzed the large dataset of phishing attacks and hypotheses about the reasons of phishing attack feasibility. They then assessed those hypotheses by showing 20 web sites to 22 participants and asked them to determine which ones were deceptive. Their results showed that 23% of the participants did not attend to security indicators, leading to incorrect choices 40% of the time.

Vishwanath et al. [18] later conducted an experiment to examine the factors for phishing susceptibility and they found that an individual email habit was an important factor for phishing susceptibility. They found that those people with entrenched email habits tended to be more susceptible to phishing attacks. This is due to their habits that as soon as a notification arrives, they are more likely to open it even though they do not realize that they are opening it.

Interventions can be utilized for better understanding user behavior in phishing susceptibility when existing studies also have consequently focused on training individuals to better detect fraudulent emails [19, 20]. Liang et al. [21] demonstrated the effectiveness of warning interfaces with two groups, one control group that had no warnings for phishing attack, and another group that had warnings. They recruit nine participants in total, where eight of them are fell for the attack. After experiments, most of the participants claim that they did not notice the warning and some don't even know what it means. Further, many of the participants admit that they don't know the meaning of phishing.

A lot of studies have been done to show that people are vulnerable to phishing for the following reasons. Many users do not trust security indicators on the websites [22]. Attackers can easily replicate legitimate websites since people usually judge a website by how a website looks and how they feel about it [17]. Although some users are aware of phishing, the information does not contribute to detect or prevent phishing attacks [23, 24]. Nowadays, machine learning techniques have been applied to detect the phishing emails [25–29].

User education, we can also think it as an intervention, about security has made a significant impact on preventing phishing attacks [30]. There is an evidence to show that a well-designed user security education can be very effective [31]. Many forms of security education, such as, interactive games, can be utilized to improve user's knowledge to prevent phishing attacks [32–35].

Supriya et al. [36] has recently studied on user behaviors in phishing attacks with incentive and intervention. They conducted a three-round experiment where participants distinguish phishing emails from normal emails. In our study, we follow closely from their experience but do more analysis. We not only study how user behaviors will affect phishing attack outcomes but also try to predict how users will perform based on their behaviors and background.

The above studies suggest the importance of understanding user behaviors in phishing attacks in order for us to efficiently avoid such attacks. In the on-site study, we specifically focus on phishing indicators to test if users can differentiate various phishing attacks and to study which type of phishing attack has more impact to user. We introduce intervention in both incentive and control groups. The intervention is to tell the user to pay attention to a certain type of phishing attack. Participants were challenged with the intervention of a phishing type where they are weak in the first round by making the phishing type easier for them in the second round. We had the incentive group to test whether or not a monetary incentive impacts the decision-making of participants, i.e., whether or not participants perform better with the presence of a monetary incentive. In the online study, We build a machine learning framework to predict the performance of a user based on their behavior and background. Compare to existing studies in the literature, our approach is more comprehensive in understanding user behaviors when phishing attacks occurs. We proposed two study designs and investigated multiple factors, such as intervention and money incentive. Furthermore, we also proposed a machine learning framework to classify user performance regarding phishing emails.

3. Study Design

Nowadays, emails have been widely used throughout the world via the Internet. Many people, especially employees in a work environment and students at colleges, read and respond emails daily. Emails become an integral part in daily life for most people. Thus, it is very likely that many people might have experience to wrongly click on a request link seemingly to be legitimate, but actually a phishing link.

In order to thoroughly understand user behaviors when phishing attacks occur and provide better user education, we present two study designs, on-site and online. The on-site study design has experiments carried out in the lab environment while the online study was carried out online. The on-site study is designed to answer the first three questions given in section 1, and the online experiment is designed to answer the last question in that section. It is important to set up our study to be correspond to user behaviors when a user read emails in the real-world. Checking emails in our daily life can be viewed as an email sorting task because when we look at an email, we will first decide whether or not it is a legitimate email. If it looks suspicious, we will not open it. Even we open it, we will look at some keywords and make a decision on whether or not the email is trustworthy and useful. In both study designs, we mimic an email opening, reading, and decision atmosphere for participantswho are asked to act as an administrative assistant to help the department chair, Dr. Jane Smith, to sort her emails while she was on vacation. Therefore, we set up an email testbed to allow users to sort a bunch of emails for Dr. Jane Smith's email accounts. Those emails consist of both legitimate and phishing ones. Participants do not need to respond to any of the emails, only sort them into either a "phishing" or "non-phishing" folder based on the information within the email and email interface. In our study, we use emails obtained from the real world with some necessary modifications for personal information protection. Phishing emails were derived from a semi-random sample of emails in "Phish Bowl" database [37]. Legitimate emails were derived from legitimate emails received by the research team. In

this section, we will give a detailed description of these two study designs.

3.1. On-Site Study Design

In the on-site study design, its email sorting task consists of three rounds and each round is preloaded with 20 emails, where 15 are phishing emails and 5 are legitimate emails. Among those 15 phishing emails, there are 3 different phishing types and each type includes 5 emails. Thus, each type of emails contains 5 emails in each round. In the second round, the intervention is introduced to the participants based on their performance of the first round. We recruit 40 participants to perform this task. During the experiment, the participants are asked to differentiate the phishing emails from legitimate emails. After the tasks in three rounds, participants are required to take a survey in the lab, where they are asked their backgrounds and their feelings about the task.

Environmental Setup. Our email testbed has three main components: RoundCube email client, Postfix virtual mail server, and BurpSuite proxy listener. The RoundCube email client is a browser-based IMAP client. It is used as an interface for users to preview and make the decision of emails in our study. Postfix mail server provides the ability of hosting multiple virtual domains. The emails preloaded in the RoundCube client are sent through Postfix virtual mail server.

We utilized the HTTP Proxy feature of BurpSuite, which serves as a man-in-the-middle between the browser and the destination web servers. This allows the interception, inspection and modification of the raw traffic passing in both directions. Therefore, both HTTP request and response sent between RoundCube client and Postfix mail server can be captured by BurpSuite. The logs obtained from BurpSuite after each round are saved as an XML format. We then parse the XML file to extract useful information for later analysis.

The email testbed is set up in the environment of Ubuntu 16.04 Long Term Support (LTS). The testbed architecture is shown in Figure 1. It consists of RoundCube Email client, BurpSuite Proxy Listener, and Postfix Virtual Mail Server where there are HTTP requests and responses among them.

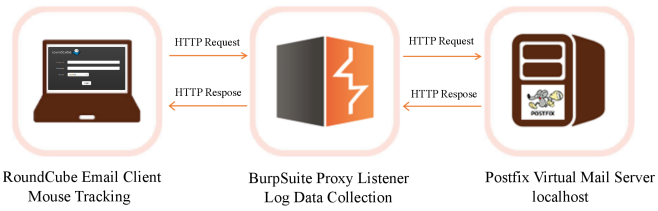

Figure 1. Email testbed architecture

In addition, we developed a Python code to track the movement of a mouse including the mouse's locations and staying durations at those locations. In our study, the developed Python code captures the time and location of the mouse during the experiment. If a participant's mouse stays in the same location for a certain long period of time, we will calculate and record the time interval t (in second). We then set up a threshold a to determine the hesitation times h. If $h > a$, we will increase h by 1. This helps us to estimate how hesitation will affect the performance.

Participant Recruitment. The IRB had been approved before we started to recruit participants (The approval number is: Pro00026240.) In the on-site study, participants are students as we recruited them on campus. They were recruited through flyers posted on the campus or announcements via mailing lists in different departments. Participants are asked to sign the Informed Consent Form before they start the experiment. We have recruited 40 participants at our university to perform this user study. To increase the diversity of participants in this study, we chose most of the participants from different majors and education backgrounds, where both undergraduate and graduate students were recruited.

The average age of the participants is about 23 years old while the participants' ages range from 18 to 38. Among 40 participants, 18 are female and 22 are male. The distribution of the participants is shown in Table 1.

Table 1. Participants Basic Information Distribution

Gender		*Age*	
Female	45%	18~ 20	25%
Male	55%	20~30	67.5%
		30~38	7.5%
Education			
Undergraduate	65%	Ph.D.	20%
Graduate	10%	Faculty	5%

We introduce a monetary incentive in our study. It is designed to answer the third question in section 1. We want to study whether the monetary incentive will affect a user's performance or not. In our another research for education purpose, we can decide if we will use this monetary incentive factor to motivate the users to pay more attention to phishing attacks. Each participant has a chance to receive \$10 to \$25 payment. To see how a monetary incentive can affect the performance, we assign them into two groups: a control group and a monetary incentive group. Each participant in the control group will get \$15 payment regardless of his/her performance. The base amount for incentive group is \$10, but participants will have a chance to earn \$5 extra from each round if they get accuracy above 80%.

Phishing Types. One purpose of this research is to study which type of phishing attacks is more malicious to user. There are three types of phishing attacks used in our study:

1. A suspicious sender's email address

This type of phishing contains a suspicious sender' email address. Nowadays, people are flooded with emails and tend to pay less attention of the sender's email address. They usually only look at the sender's name, neglect of the email address, or just catch a glimpse of the sender's email address. This information gives the scammer a high chance to replicate the email address. Some of the phishing email addresses are really hard to be distinguished from the authentic email addresses if users do not pay much attention. For example, the letter 'l' and the number '1' are very similar. Therefore, the scammer could utilize this feature to create a fake 'we11sfargo' domain name rather than 'wellsfargo.'

2. Suspicious links or attachments

Suspicious links can be very similar to suspicious sender's email addresses. These links could be manipulated through using similar characters or misspelling issues. For example, a link contains the word 'directdeposit' could be misspelled as 'directdepost.' The suspicious attachments can be disguised as the pdf file, exe file, or other types of files. A suspicious exe file may be easier to spot than a suspicious pdf file. Usually, people will not consider that a pdf file could be malicious until they open it.

3. Malicious Email Contents

This type of phishing is quite tricky. At first glance, the email content seems normal to most people. However, this kind of phishing attacks contains suspicious contents. For example, the contents may have several grammar issues or the icon of popular social networks are faked. They are very hard to notice if the user is not familiar with those popular social medias or if the user is not a native English speaker.

Experimental Rounds. In the on-site study, we let participants perform three rounds of email sorting tasks. We collected data of each participant from each round. The average time spend in each round is about 15 minutes. After each round, the participant can take few minutes rest while we save the data captured from experiments.

First Round: This round contains 20 emails in total. Among them, 15 are phishing emails and 5 are legitimate emails. Those 15 phishing emails consist of three types of phishing attacks we introduced above. Each type of phishing attack has 5 emails. The task for participants is to classify 20 emails into two folders, suspicious or keep, based on their knowledge and experience. Participants were not told how many phishing emails and legitimate emails were given.

Second Round: The second round has the same procedure as the first round except that we introduce the intervention in this round. The intervention is to make the phishing type that emails become more obvious to participants, so they will pay more attention to this certain type of phishing attack. This could be an useful factor in user education to prevention phishing attack that we can educate them about different types of phishing attacks. After a participant finished the first round, we calculate the score of first round for the participant. The score is calculated based on the correctness of sorting each email. If a participant moves the email to the correct folder, he/she will get 1 point; otherwise, 0 point is granted. The score then be added up together. The performance score is the total score of sorting all 60 emails. We separated the score for different phishing types and checked for the lowest score among three phishing types. Therefore, the phishing type with the lowest score was used as an intervention in the second round. Before the second round started, we pointed out the type of phishing attacks with the lowest score to the participant and made this type of phishing attack easier for participant to spot in the second round. The reason to introduce intervention is that we want to examine whether a participant will perform better in this round with the knowledge of the certain type of phishing attack. We also want to see whether the intervention action will affect the overall performance of each participant or not.

Third Round: This is the last round in our experiment. In this round, a participant will continue to sort 20 emails. The procedure of the third round is the same as the previous two rounds. But we will not give any intervention in this round. We will compare the performance score between round three and the other two rounds, to see if the intervention from last round still have an effect on the third round.

Survey. The survey was carried out after three rounds of email sorting tasks. We used an online survey platform to record the answers from participants, where they were required to complete it in the lab. This survey contains 30 questions and is mainly about the background of participants, such as, age, gender, and some general questions about their experience and habits of using social medias. There were also some questions related to the email sorting tasks they just took. The examples of the survey questions are shown as follow:

- Have you taken any cybersecurity courses?
- I believe I was successful in the email sorting task.
- I briefly looked at the sender/source of the emails.

- I ignored the message content of the emails.

Here, participants are given multiple choices, “Yes/No" for the above first question and “Strongly Disagree/Disagree/Neutral/Agree/Strongly Agree" for the above rest of three questions, where the participants are required to choose one of the answers in the multiple choices.

Besides the data we collected throughout their experiment, this survey can better help us understand participants behaviors and background regarding to phishing attacks. A complete list of survey questions is shown in Appendix A.

3.2. Online Study Design

Although by performing the on-site study, we can sufficiently answer the first three questions mentioned in the section 1, it is not sufficient for us to thoroughly understand user behaviors regarding phishing attacks. This is due to the limitation of demographic diversity and the number of participants recruited, etc. Therefore, we propose the online study design developed by the project team at our university. The online study can sufficiently help us to answer the question of what kind of groups are more vulnerable to phishing attacks and how accurately can we predict the performance based on user behavior. Since the online study design is an extension of the on-site study design, we will only introduce the new components of the online study design and compare both designs afterwards.

Environmental Setup. Online study has an environment setup similar to the on-site study, except that we are not using BurpSuite proxy listener to capture the data since the experiment is carried out online. To collect user’s input, we use the JavaScript-Based Data Capture and to communicate the captured data to the server, we use the AJAX-Based Data Sender. The PHP Listener is used to receive the data sent from AJAX, and the Logger is used to log the data. On the server side, both Listener and Logger are installed. Both Data Capture and Data Sender are on the client side browser. In order to see how confident participants are while they are sorting an email, we add a rating module in the Roundcube email client so that they can rate their confidence level of each email. The rating is ranging from 1 to 10.

Participant Recruitment. In the online study, participants are recruited from Amazon Mechanical Turk (MTurk) [38]. We recruited 90 participants in total for this online study. The average age of the participants is about 34 years old while the participants’ ages range from 20 to 61. Among 90 participants, 35 are female and 55 are male. There are 8 participants are currently students. There is one participant who is not an English native speaker. Nine participants previously completed a network engineering or cybersecurity course/certificate. The distribution of the participants is shown in Table 2.

Table 2. Participants Basic Information Distribution

Gender		*Age*	
Female	39%	20~ 35	65.6%
Male	61%	36~50	26.7%
		51~61	7.7%
Highest Education			
High School	13.3%	Master	6.7%
College	77.8%	Doctorate	2.2%

We still keep the monetary incentive mechanism in the onlinse study because it is a useful feature/attribute for predicting the performance of user behavior when encountered with phishing attacks. The base amount is $4 for non-incentive group. For the incentive group, participants could earn additional payment (up to $8.00) for their performance if it is greater than 75% accuracy.

Phishing Types and Experimental Round. The online study utilizes the same phishing types as we used in the on-site study. However, there is only one experimental round in our online study. Since the on-site study is sufficient for exploring the effect of intervention, to make it simpler, we only use one round of email sorting task and the participants are asked to sort 40 emails as well as to rate their confidence level for each email. They are asked to complete the task within 30 minutes. Among those 40 emails, 20 emails are legitimate and 20 are phishing. Participants are not aware of this distribution when they take the experiment.

Survey. Because the study is carried out online, we designed two types of surveys, pre-survey and post-survey. We use the pre-survey to investigate the basic information and background of participants, such as age, gender, education background, cybersecurity background, habits of using social media, etc. When we carry out the on-site study, we include the Informed Consent Form and email sorting instructions whose information is similar to the questions in the pre-survey in the online study. The post-survey asked questions related to the email sorting task they took. The pre-survey and post-survey questions contain all the questions from the on-site study survey and we add more questions because the online study is designed differently from the on-site study. For example, we add the confidence rating in the online study, so in the post-survey from the online study, we have a question: “How confident are you in your assessment of the number of correctly sorted email?” which is not in the on-site survey. The complete survey question is shown in Appendix B.

Table 3. Example of Attributes Used In Each Study.

Attribute Name	*Description*	*Study*
Performance Score	Overall performances for the participants	Both
Rx_P1_Score	Performance for the phishing type 1 emails in each round	On-site
Rx_P1_Time	Time used for sorting the phishing type 1 emails in each round	On-site
Rx_Nr_Score	Performance for 5 legitimate emails in each round	On-site
Rx_Nr_Time	Time that used to sort 5 legitimate emails in each round	On-site
Rx_Phish_Score	Performance for sorting all 15 phishing emails in each round	On-site
Rx_Phis_Time	Time that used to sort all 15 phishing emails in each round	On-site
Avg_Rating	The average confidence rating of all emails	Online
Sort_Correct	How many emails the participant thought they sorted correctly	Online
Sort_Confidence	How confident the participant feels after the task	Online
num_sorted	How many emails have been sorted in the task	Online
Phish_Score	Performance for sorting all phishing emails	Both
Phish_Time	Time that used to sort all phishing emails	Both
Native_Spk	Whether the participant is native English speaker	Both
Edux	Education background for the participants	Both

3.3. Similarity and Comparison

Since the online study is designed based on the on-site study, these two study designs are similar to a certain extent but also have differences because they are focusing on different aspects of user behavior study.

Similarity. Both of the studies contain email sorting tasks and aim to study user behaviors when users encountered with phishing attacks. Both of them divide participants into two groups, monetary incentive group and control group. In these two studies, their phishing types are the same. Both conduct surveys about the email sorting task afterwards.

Comparison. First, the on-site study is carried out in the lab environment. Participants are asked to show up and perform the experiment on the testbed setup in the lab, while the online study is carried out totally online including recruit participants and email sorting task, and so on. Second, participants recruited online are more diverse than the on-site study and the number of participants recruited for the online study is a lot more than the one in the on-site study. Third, for the on-site study we design the intervention mechanism and three-round email sorting task sorting 60 emails in total. Our collected data is sufficient for doing the analysis regarding the intervention question, so, for simplicity, we only design one round email sorting task for sorting 40 emails. Fourth, in the online study we add the rating module to allow user to rate their confident level for each email, which is turned out to be a useful feature in our machine learning framework. Fifty, for the survey questions in the on-site study, we only ask participants to do them after the experiment and we do not have pre-survey. However, in the online study we have both pre-survey and post-survey due to the form of online experiment. There are more survey questions than the on-site study.

4. Data Analysis Methodologies

The goal of this research is to thoroughly understand the behavior of a user encountering phishing attacks and to identify which factor plays a significant role in phishing attack outcomes. We raised the four questions in the introduction section. In order to answer these questions effectively, we propose two data analysis methodologies. Especially, in order to answer the first three questions, we propose a statistic method to analyze each factor such as intervention, phishing types and incentive. To answer the fourth question, we propose machine learning techniques. In this section, we first discuss our dataset and attributes and then propose a machine learning framework.

4.1. Data Set and Attributes

We developed a data collection infrastructure such that it automatically captured and monitored the detailed actions of each participant like clicks, navigation, timestamps, decisions, etc. We further processed and stored this information in a CVS file format for analysis. Since we proposed two study designs, the datasets, on-site dataset and online dataset, are separately stored and analyzed by different data analysis methods. The on-site study design mainly focuses on the factors that will affect user behaviors with phishing attacks, while the online study is designed to predict the user performance with phishing attacks. Although we also predicted the user performance by analyzing the on-site dataset based on machine learning approaches, the result is not as good as we expect because of the small dataset. Thus, we design the online study

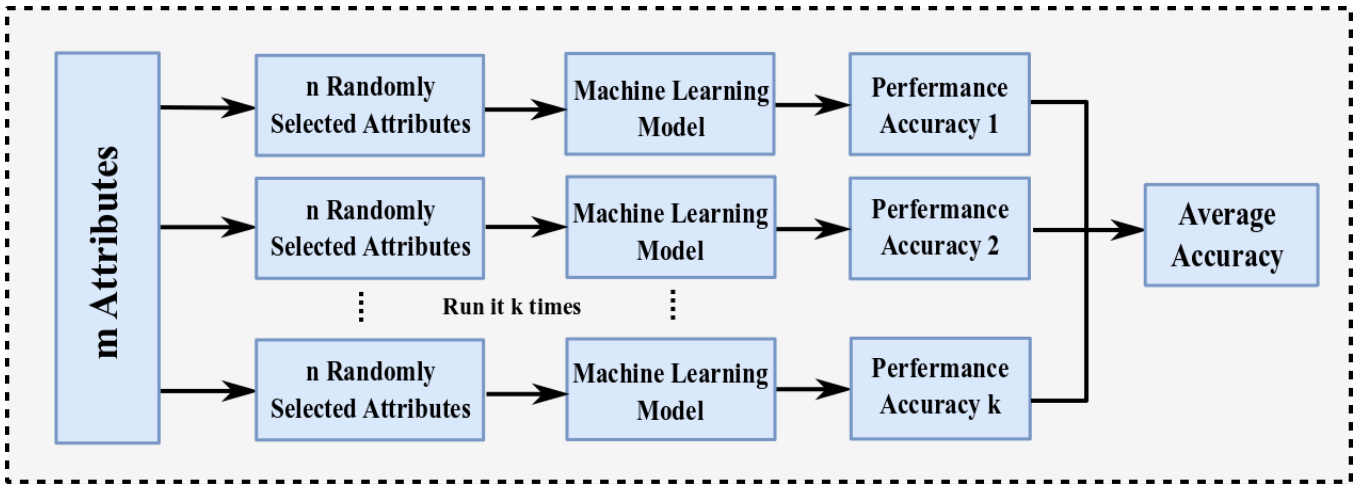

Figure 2. The proposed machine learning framework where we choose $m = 16$, $n = 4$, and $k = 4$ in our study.

whose collected data are efficient to predict the user performance as shown later. For this reason, we apply statistic models to analyze the on-site dataset in order to understand the contributing factors of user behaviors when phishing attacks occur. We will also briefly show our machine learning user performance prediction results by using on-site dataset and compare them with the ones based on online dataset.

The on-site data file stores data collected from 40 participants. The file includes the participants' detailed information, time used to sort email (processing time) and performance score, etc. There are 50 attributes in total for the on-site dataset. The online data file stores data collected from 90 participants. It includes the similar information as in on-site data file and additional information including confidence rating, new survey questions and so on, but it doesn't include intervention information. There are 119 attributes in total in the online dataset. A part of them are the processing time and performance score to evaluate the performance of a user. Besides these parts, a lot of attributes are coming from the pre and post surveys that provide basic information such as gender, age, education level, questions about the task and so on (see Appendices A and B). An example of the attributes can be seen in Table 3. In the study column, 'On-site' or 'Online' means that the attribute is only for the on-site study or the online study, respectively. 'Both' means that the attribute is for both studies. The online study has more attributes from the survey questions than on-site one. A complete list of attributes with their descriptions is given in Appendix C.

Performance score is one of the most important indicators in both of our studies. For the on-site dataset, we calculate the score for each participant in each round as well as the score of each phishing type. We can use a statistic method to analyze what factors, such as, intervention, processing time, and incentives, are closely related to the performance score. For the online dataset, we calculate the score of each participant and the score of phishing email and normal email. We can feed the online dataset into machine learning models to predict the performance of the participant for understanding the behavior of a user encountering with phishing attacks.

4.2. Proposed Machine Learning Framework

The goal of using a machine learning method is to predict a user's performance when the user encounters phishing attacks, whether or not the user can do well or poorly. Hence, we divided the performance score into two different classes, *Good* and *Poor* where a critical point, c, is used as the threshold in the division. If the performance score is greater or equal to c, then we label it as *Good*, otherwise, it is labeled as *Poor*. Thus, we have to deal with the classification problem. Some attributes are more significant than others, where some of those other attributes have minimal or no significant effects. Therefore, choosing attributes is critical in our machine learning framework, especially with a small dataset in the case of the on-site study. Moreover, correlation studies are conducted to illustrate the relationship of these independent attributes with the performance scores of participants. Since we have 119 attributes but only 90 datasets, a critical question has raised. That is, while the sample size of our dataset is relatively small, the number of attributes is relatively big. In order to prevent over-fitting, we require a sufficient number of datasets for a certain number of attributes in machine learning models [39, 40]. As we know, a non-over-fitting machine learning model usually requires at least P^2 datasets to train the model for P attributes. Clearly, our dataset does not meet the requirement. To resolve this problem, we introduce a stepwise attribute section to

reduce attributes and proposed our machine learning models in detail. Then, we present how to find the best critical point to classify user performance where a prediction accuracy is ensured.

Stepwise Attribute Selection. To select the best attributes for the following machine learning models, we first perform a Pearson-correlation coefficient analysis to observe the importance of each single attribute. Based on the data we collected, we then fit our data into a linear regression model to evaluate all the attributes. In order to select the most significant attributes to build the model, we use three ways which is stepwise, forward, and backward selections to select the attributes. The model entry significant level was set to 0.5 and the stay significant level was set to 0.2.

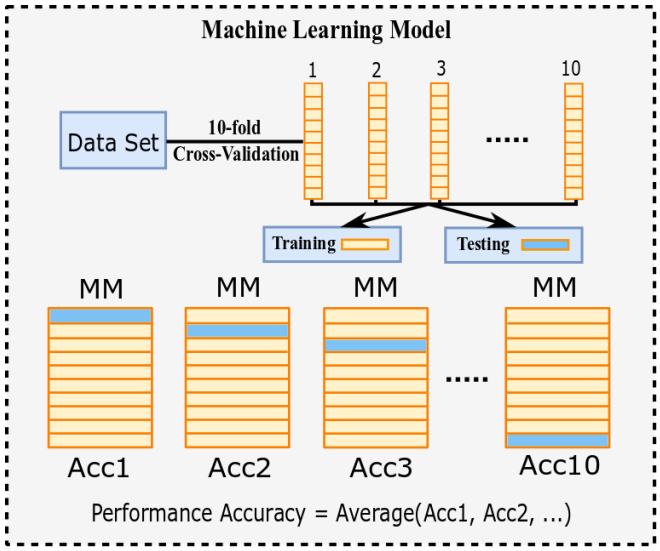

Figure 3. Machine learning model with 10 fold cross-validation

Machine Learning Models. After we get the reduced attributes, we now apply machine learning approaches to predicting the overall performance. We build 4 different machine learning models, Decision Tree-J48, Naive Bayes (NB), Support Vector Machine (SVM), and Multilayer Perceptron (MP). We first use Decision Tree-J48 based on the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. The Decision Tree classifier requires relatively little effort for data preparation. The Naive Bayes classifier works well for independent attributes based on the Bayes rule of conditional probability [41, 42]. It will consider each of the attributes separately when classifying a new instance. SVM is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. Multilayer Perceptron is a type of neural networks that usually consist of at least three layers of node. The node in each layer uses nonlinear activation function.

We also propose to use the method of 10 fold cross-validation to precisely predict the performance of each participant, as shown in Figure 3, where MM is short for the machine learning model. The original dataset is randomly partitioned into 10 equal size subdatasets. Of the 10 subdatasets, a single subdataset is retained as the validation data for testing the model and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (i.e., 10 folds), with each of the 10 subdatasets used exactly once as the validation data. The 10 results from the folds can then be averaged to produce a single estimate. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. Besides the 10 fold cross-validation, we also utilize the similar idea of cross-validation for the attributes. Suppose we have m attributes and each time we randomly select n attributes to feed into our cross-validation machine learning model. This process will be running in k times, where $m = k \times n$.

The procedure of the proposed machine learning framework is shown in Figure 2. We have m attributes in total after stepwise attribute selection. Then we randomly choose n attributes to do the cross-validation training by applying our machine learning model. The next step is to calculate the performance accuracy. This process can be running k times. These k performance accuracies are averaged to form one final accuracy. In our proposed model, we use 10-fold cross-validation to do the training and testing. Each time we will obtain an accuracy, and this will be done 10 times. The performance accuracy is calculated by averaging all the accuracies.

Finding A Critical Point. To predict the performance of a user encountering phishing attacks, we divide the performance into two classes, *Good* and *Poor*, based on the performance score of a user, as discussed before. Let us recall that the critical point c is used to divide the performance score. Finding a critical point is very important step before we use our machine learning models to do the prediction. To find the critical point, we use a greedy method to go through each threshold and check to see if it is the preferred accuracy. In the evaluation section, we will show how to find the critical point in details.

5. Evaluation

In this section, we analyze and identify what factors may make a significant impact on a phishing attack outcome. Motivated by the questions introduced in the section 1, we are going to first evaluate the intervention factor and find the type of phishing attacks that is

more harmful to people. Then, we want to see if there is the difference of time spent on between phishing emails and normal emails. Also, we will study if a monetary incentive can improve the participants performance regardless of their backgrounds. Last, but not the least, the evaluation of our machine learning models will be presented. The evaluation of intervention, phishing types, and a monetary incentive are using the dataset from the on-site study, while the evaluation of performance prediction is using the dataset from the online study. We will also present the performance prediction results when using on-site dataset and compare it with the results by using online dataset.

5.1. Intervention Evaluation

To answer the first question that how intervention can affect the phishing attack, we calculate mean phishing score, mean total score, mean total processing time and mean phishing processing time. The result is shown in Table 4. The intervention is introduced in the second round and based on the performance of the participant from the first round. From Table 4, we can see both phishing scores where the full score is 15 and the total score is 20. As shown in the table, the second round has been slightly improved compare with the first round. The mean time used in the second round is also lesser than the first round. However, in the third round, the performance score has decreased and even worse compared with the first round.

5.2. Phishing Type Evaluation

We analyze the performance score and time for different types of phishing attacks. The question is what kind of phishing attacks are more harmful to people can be answered in Table 5. Type 1 phishing attack contains a suspicious sender's email address, type 2 phishing attack has suspicious links or attachments, and type 3 phishing attack contains malicious contents. The mean score (full score is 15) and mean time are calculated by taking average of all 40 participants' score and time of different phishing types. The intervention frequency describes the total times of a certain type phishing intervention introduced in the task. We can see from the table that type 1 phishing has the lowest score and it has been used the most as an intervention. This implies that the type 1 phishing is more harmful compared to the other two types. In addition, it is not hard to see that the score is in inversely proportion to intervention frequency. Thus, intervention is a suitable attribute that can be used in our neural network.

5.3. Monetary Incentive

The next question is whether a monetary incentive affects the performance and total processing time. We calculate the mean total performance score, mean phishing performance score, mean total processing time, and mean phishing processing time of all 40 participants. The result is shown in Table 6. Condition 0 means that there is no monetary incentive. That is, it is the control group, and condition 1 represents that this group will get a monetary incentive. We can see from the table that the group with incentive has a higher performance score than the group who doesn't. Furthermore, the incentive group tends to spend more time than the control group. Therefore, incentive is also a useful attribute regarding a phishing outcome.

5.4. Mouse Movement Evaluation

In our study, we also record the mouse movement from each participant and calculate the hesitation times as described in the section 3, where we pick the threshold $h = 10$. We analyze the relationship between hesitation and the total time used in this study as well as the relationship between hesitation and total score. The relationships between hesitation and total time and the relationship between hesitation and performance score are shown in Figures 4.

We can see from Figure 4 (a) that as hesitation times increases, the total time is also increasing. The orange line is representing the incentive group while the blue line is representing the control group. It is clear from this figure that the incentive group tends to spend more time and has more hesitation times. This is because the participants in the incentive group are more cautious when doing this task.

Figure 4 (b) shows the relationship between total score and hesitation times. For the control group the relationship is not so obvious. For the incentive group, the total score is decreasing as the hesitation times increases. It is interesting to see that if the participants get more cautious, they tend to be performing worse.

5.5. Time difference between phishing email and control email

The next question we want to know is whether users spend different time in normal or phishing emails. Since there are three rounds in total, we compare the time of each round as well as the total time of all three rounds. As shown in Figure 5, in round one, users spend lesser time in phishing email. In round two, User also spend lesser time in phishing email. However, in round three, user spend more time in phishing email. Thus, in total, there is no significant time difference between normal and phishing email.

5.6. Attribute Reduction

The first important step is to select the useful attributes that will be used in our machine learning models.

Table 4. Email Round Score and Time

Attributes (Mean)	*Round1*	*Round2*	*Round3*	*R2-R1*	*R3-R2*
Phish_Score	10.18	11.6	10.02	1.42	-0.15
Total_Score	14.2	15.23	14.25	1.025	0.05
Phish_Time(s)	437.58	413.45	433.25	-24.13	19.8
Total_Time(s)	630.88	600.4	568.35	-30.48	-32.05

Table 5. Different Types of Phishing Score and Time

Phishing Type	*Mean Score*	*Mean time(s)*	*Intervention Frequency*
Type 1	9.5	447.425	17
Type 2	11.35	431.875	8
Type 3	10.95	404.975	15

Table 6. Monetary Incentive Analysis

Condition	*Phish_Score*	*Total_Score*	*Phish_Time*	*Total_time*
0	30.1	42.65	1148.95	1580.5
1	33.5	44.7	1419.6	2018.75

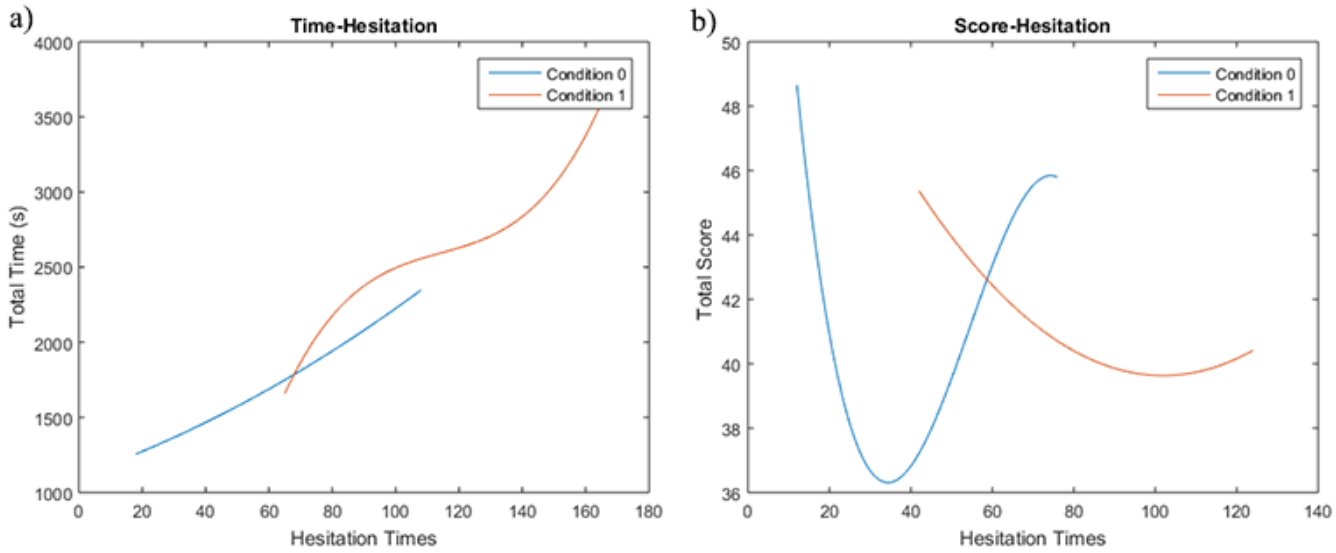

Figure 4. (a) Relationship between hesitation and total time. (b) relationship between hesitation and performance score.

We perform the Pearson-correlation coefficient analysis to observe the importance of each single attribute. From our observation, we could see that most of the attributes are not significant related to the total score. The detailed information of part of the attributes is shown in Table 7.

From the table, the order is sorted into most significant to less significant. We can see the attributes phishing_accuracy has $p < 0.0001$. Some attributes are significant, such as sort_agreement_4 and sort_correct_1, are from the survey questions. In particular, sort_agreement_4 is referred to the question in post survey: "I felt irritated and stressed while sorting emails." In this analysis, we select 16 attributes that will be used in our machine learning models.

5.7. Critical Point Evaluation

Before we apply machine learning model to predict whether a participant will perform well or not, we need to find a critical point to label the training data as Good or Poor. We test the critical points for each machine

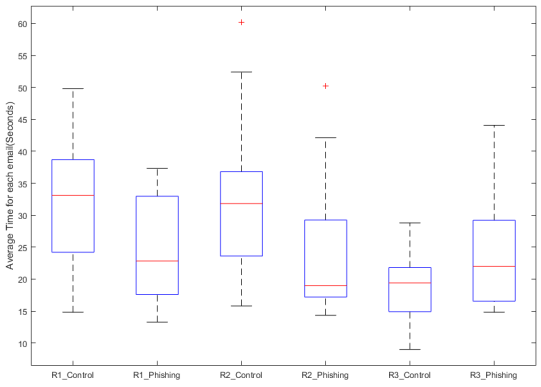

Figure 5. Time comparison for normal email and phishing Email

Table 7. Pearson Correlation Coefficients

Attributes Name	*Pearson Correlation Coefficient*	*Prob > |r|*
phishing_accuracy	0.71771	<.0001
pay	0.56868	<.0001
num_sorted	0.53976	<.0001
sort_agreement_4	-0.46401	<.0001
sort_correct_1	0.41037	<.0001
avg_rating	0.30178	0.0038
strat_cues2_5	-0.26026	0.0132
sort_confident_1	0.25666	0.0146
sort_agreement_1	-0.25249	0.0164
BFI_BFI_41	-0.24255	0.0213
BFI_BFI_30	0.19971	0.0591
incentive	0.19248	0.0691
beliefs_agreement_2	0.16783	0.1139
highest_education	0.15475	0.1453
cyber_experience	0.13459	0.2059
reg_sm_scale	-0.12917	0.2250
strat_cues3_6	-0.11807	0.2677
gender	-0.11538	0.2788
strat_cues3_3	0.09484	0.3739
beliefs_agreement_5	-0.09279	0.3844
BFI_BFI_9	-0.08385	0.4320
habits_sm_scale	-0.03005	0.7786
input_type	0.00439	0.9672

learning model. The result is shown in Table 8. We can see from the table, c is the critical point starts from 15 to 38, this is because the lowest performance score is 14 and the highest performance score is 39 in our online dataset. The items Good and Poor are the number of instances are labeled as Good or Poor based on the critical point c, respectively. Both accuracy and time are presented for each machine learning model, where the time is in second. From the table, we observe that when the critical point is 15, 16, 17, 18, 19, or 20, the four models have the highest accuracy. However, the distribution of the number of Good and the number of Poor are unevenly distributed. We need to choose the critical point that is able to divide the number of Good and the number of Poor more reasonably and meanwhile to achieve better accuracy. Therefore, the best critical point for J48 is $c = 32$ with the accuracy of 97.78%, and for Naive Bayes is 88.89% accuracy with $c = 31$, for SVM and Multilayer Perceptron it is also when $c = 30$, the accuracies are 92.22% and 96.67%, respectively.

5.8. Performance Prediction Evaluation

In our machine learning framework, we use four different machine learning models, J48, Naive Bayes, SVM and multilayer Perceptron, to predict the performance. We have presented a table of critical point in the above section, to further observe the preferred critical point, let's take a look at Figure 6 (a). It shows the accuracies of different machine learning models choosing different critical points. The critical points started from 26 because for the critical points smaller than 26, the division of two classes are unevenly distributed. We can see from this figure, when the critical point is 30, except for J48, all other three models, have the relatively highest accuracy compared with choosing other critical points. Therefore, we choose the critical point $c = 30$ to label the dataset into two classes, Good and Poor.

Next, we use four machine learning models with 10 fold cross-validation to do the classification. Figure 6 (b) shows the accuracy of each fold when using four different machine learning models. The final accuracy result is the average of all 10 folds. For each fold, fold NO.1 to 10, the accuracy ranges because each fold is using different training and testing subdataset as we discussed in the last section. For J48, the accuracies ranging from 55.56% to 100%, only fold 5 and fold 9 reaches 100% accuracy and the worst accuracy is 55.56% from fold 10. For Naive Bayes, the accuracies ranging from 77.78% to 100%, fold 2, 3, 10 has the lowest accuracy and fold 5, 6, 7, 9 has the highest accuracy. SVM has the accuracies ranging from 77.78% to 100%, but it is better performance than Naive Bayes. Multilayer Perceptron has the accuracies ranging from 88.89% to 100%. It is better compared with other four machine learning models. Figure 7 (a) shows the Mean Squared Error (MSE) of each fold for four different machine learning models. The results of MSE show the correspondence with the accuracies of each fold. As accuracy increases, the MSE decreases. Among them, J48 of fold 10 has the highest MSE because accuracy of J48 with fold 10 is the lowest.

Then, we evaluate the performance accuracy as described in our machine learning framework in Figure 2. As we described in section 4.2.2, we also apply the similar idea of cross-validation to attributes.

Table 8. Critical Point (c)

c	Good	Poor	J48		NB		SVM		MP	
			Accuracy	Time(s)	Accuracy	Time(s)	Accuracy	Time(s)	Accuracy	Time(s)
15	89	1	98.89%	2.889	98.89%	0.024	98.89%	0.063	98.89%	2.889
16	89	1	98.89%	2.76	98.89%	0.011	98.89%	0.045	98.89%	2.76
17	89	1	98.89%	2.745	98.89%	0.01	98.89%	0.023	98.89%	2.745
18	89	1	98.89%	2.74	98.89%	0.007	98.89%	0.055	98.89%	2.74
19	89	1	98.89%	2.754	98.89%	0.01	98.89%	0.018	98.89%	2.754
20	89	1	98.89%	2.717	98.89%	0.006	98.89%	0.019	98.89%	2.717
21	88	2	95.56%	2.731	97.78%	0.007	98.89%	0.016	97.78%	2.731
22	88	2	95.56%	2.73	97.78%	0.006	98.89%	0.049	97.78%	2.73
23	87	3	94.44%	2.721	97.78%	0.007	97.78%	0.047	96.67%	2.721
24	86	4	92.22%	2.75	97.78%	0.008	96.67%	0.015	95.56%	2.75
25	85	5	91.11%	2.728	96.67%	0.007	96.67%	0.017	94.44%	2.728
26	80	10	85.56%	2.743	86.67%	0.004	90.00%	0.027	92.22%	2.743
27	78	12	90.00%	2.748	84.44%	0.004	88.89%	0.022	88.89%	2.748
28	70	20	87.78%	2.732	85.56%	0.003	87.78%	0.019	87.78%	2.732
29	63	27	88.89%	2.788	83.33%	0.004	91.11%	0.032	93.33%	2.788
30	53	37	86.67%	2.835	**88.89%**	0.004	**92.22%**	0.069	**96.67%**	2.835
31	47	43	96.67%	2.781	88.89%	0.005	91.11%	0.067	93.33%	2.781
32	40	50	**97.78%**	2.833	82.22%	0.007	91.11%	0.071	94.44%	2.833
33	37	53	97.78%	2.86	81.11%	0.004	88.89%	0.031	91.11%	2.86
34	27	63	85.56%	2.851	82.22%	0.004	86.67%	0.019	93.33%	2.851
35	16	74	94.44%	2.809	84.44%	0.004	86.67%	0.013	90.00%	2.809
36	11	79	96.67%	2.835	84.44%	0.003	87.78%	0.015	88.89%	2.835
37	8	82	96.67%	2.802	92.22%	0.003	91.11%	0.014	91.11%	2.802
38	3	87	93.33%	2.802	96.67%	0.003	96.67%	0.022	96.67%	2.802

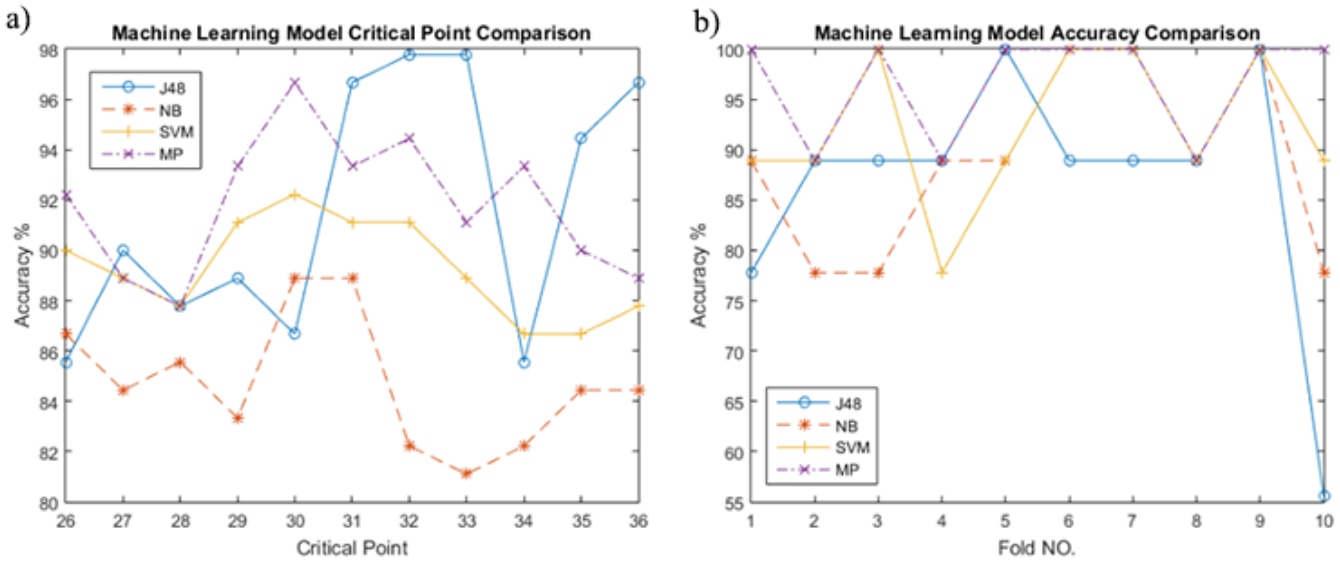

Figure 6. (a) Accuracy of each machine learning model with different critical points. (b) Evaluation of 10-fold cross-validation accuracy of each fold for different machine learning models.

Figure 7 (b) shows the accuracy result of random selected attributes. In our study, we choose $N = 4$, so we have each time there are 4 attributes used for testing and rest are used for training, and this process is done for 4 times, as shown in the x-axis, 1st, 2nd, 3rd, and 4th. The accuracy of each time is the performance accuracy after doing the 10-fold cross-validation and the final accuracy is the average of the four performance accuracies. The accuracies range because the different attributes are chosen each time. We can see from the figure, we can see the final accuracy for J48, Naive Bayes, SVM and Multilayer Perceptron is 86.67%, 88.89%, 92.22%, and 96.67%, respectively.

After analyzing the 10-fold cross validation, the accuracy in the following analysis is the final accuracy by averaging 10 folds accuracies. Figure 8 (a) shows the relationship between accuracy and number of instances, which means the number of participants because we treat each participant as an instance. We can see as the number of instances increases, the accuracy is also increasing. Among them, Multilayer Perceptron has the best accuracy, which is 93.84% in average. When using all 90 instances, the accuracy reaches 96.67%

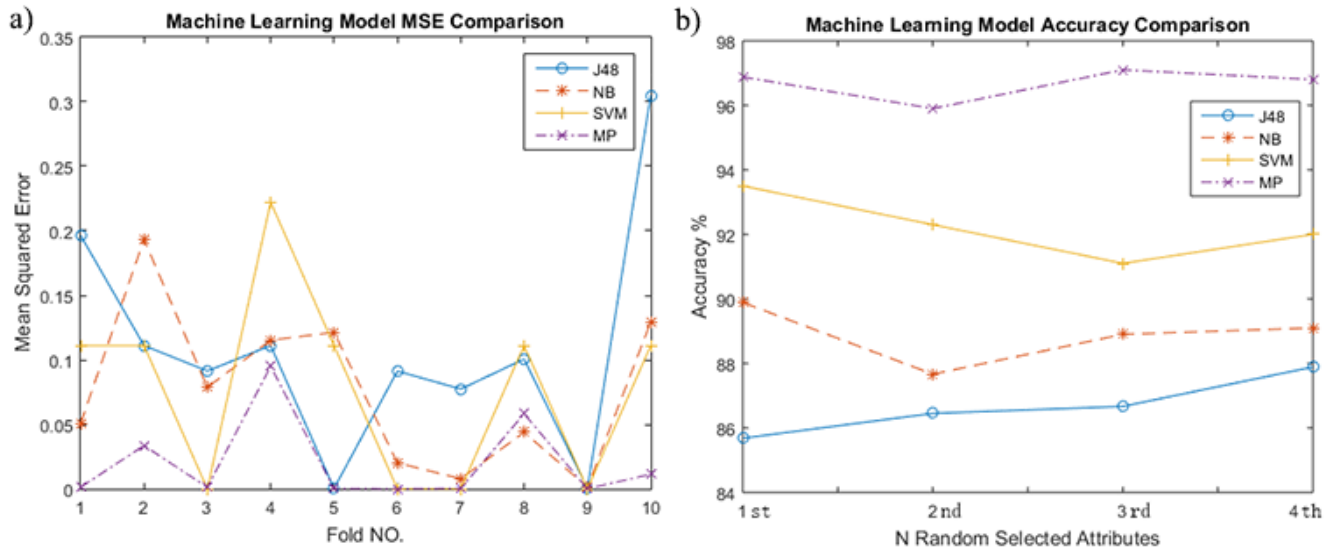

Figure 7. (a) Evaluation of 10-fold cross-validation Mean Squared Error (MSE) of each fold for different machine learning models. (b) Accuracy of N random selected attributes.

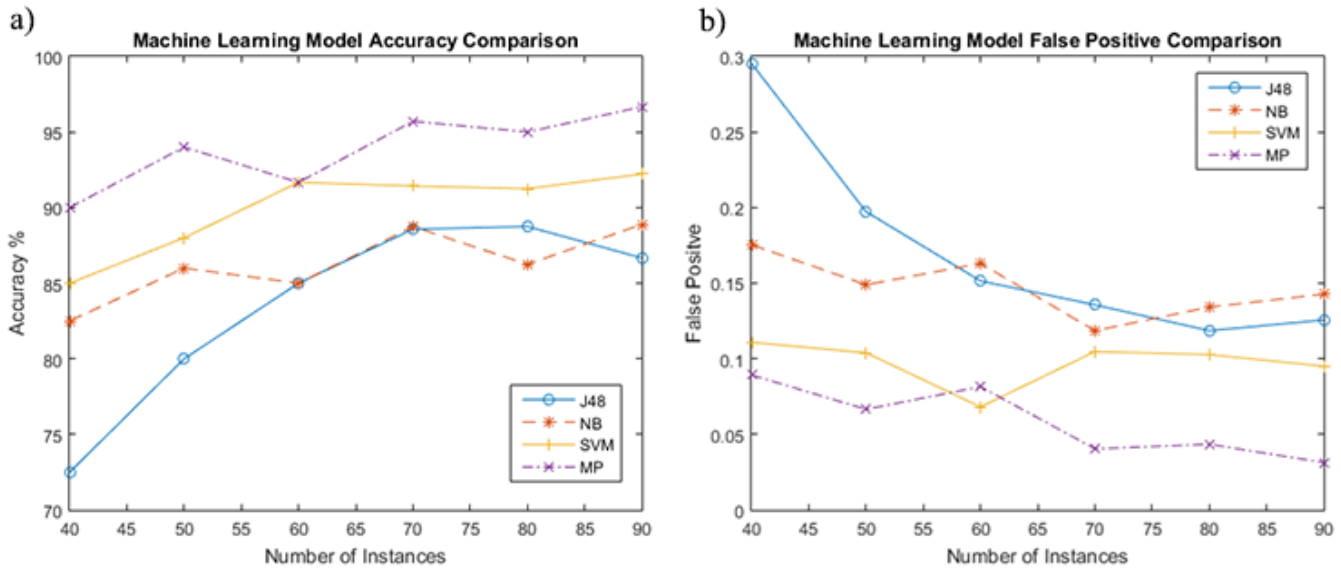

Figure 8. (a) Accuracy of each machine learning model with different number of instances. (b) Evaluation of false positive rate for different machine learning models.

for Multilayer Perceptron. SVM has the second best accuracy, the average accuracy for SVM is 89.93%. In addition, when using all 90 instances, it has the best accuracy, which is 92.22%. The average accuracies for Naive Bayes and J48 are 86.23% and 83.58%, respectively.

Figure 8 (b) shows the false positive rate for all four models. We also compute false positive rates because false positive is also an important measurement that we want to keep it as low as possible. Multilayer Perceptron has the lowest false positive rate at instance number of 90, which is 0.0314. J48 has the highest false positive rate, which is 0.2953 when the number of instance is 40. We can see as the number of instances increases, the trend of false positive rate for all four machine learning models are decreasing. The average false positive rate for J48, Naive Bayes, SVM and Multilayer Perceptron is 0.1707, 0.1472, 0.0976 and 0.0588, respectively. Among them, J48 has the highest false positive rate and Multilayer Perceptron has the lowest false positive rate.

Aside from just compare the accuracy, we also use other metrics to compare them, such as false positive rate, F-Measure, MCC and area under ROC. The comparison result is shown in Table 9. The

Table 9. Evaluation of Machine Learning Models

Evaluation Method	*J48*	*Naive Bayes*	*SVM*	*MP*
Accuracy	86.67	88.89	92.22	96.67
FP	0.1257	0.1428	0.0951	0.0314
TP	0.8667	0.8889	0.9222	0.9667
Precision	0.8717	0.8936	0.9232	0.967
Recall	0.8667	0.8889	0.9222	0.9667
F-Measure	0.8675	0.887	0.9216	0.9667
MCC	0.732	0.7725	0.8394	0.9317
Area under ROC	0.936	0.9444	0.9136	0.998
MSE	0.1085	0.0762	0.0778	0.0201
Time(s)	0.089	0.08	0.208	4.552

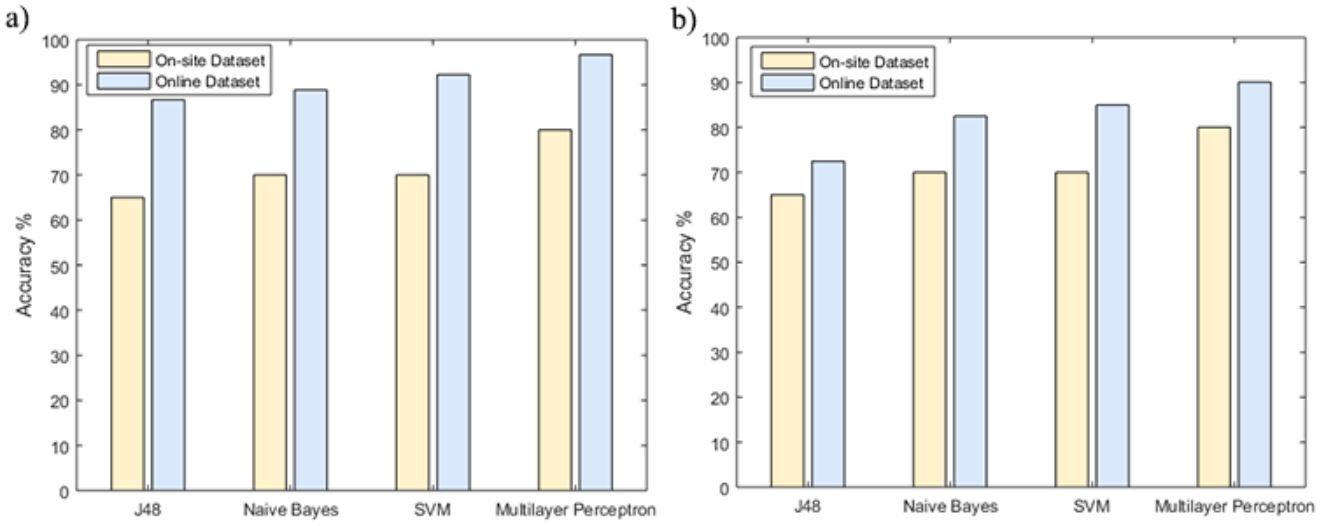

Figure 9. (a) Accuracy comparison of four machine learning models using on-site dataset and online dataset with 90 instances. (b) Accuracy comparison of four machine learning models using on-site dataset and online dataset with 40 instances.

False Positive (FP) rate and True Positives (TP) rate are common measures in machine learning, the TP rate is the higher the better and the FP rate is the lower the better. Precision is defined as the number of TP over the number of TP plus the number of FP. Recall is defined as the number of TP over the number of TP plus the number of False Negatives (FN). The F-Measure is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test. The Matthews correlation coefficient (MCC) is used in machine learning as a measure of the quality of binary (two-class) classifications, the value are more approximate to positive 1 represents a perfect prediction. The Receiver Operating Characteristic (ROC) curve is created by plotting the true positive rate against the false positive rate at various threshold settings. In this experiment, we calculate the area under the ROC curve. We can see that Multilayer Perceptron has the highest value of MCC, area under ROC curve, F-Measure, etc. This demonstrates that Multilayer Perceptron has the best performance among all four classifiers.

We also use the on-site dataset to do the user performance prediction, we followed the same procedure as we discussed above. However, in the on-site dataset, we only have 40 participants, which means we can only use 40 instances. Figure 9 (a) shows the accuracy comparison of four machine learning models using on-site dataset and online dataset. The online dataset contains 90 instances. We can see the accuracy of the online study is much better than the on-site study. For J48, the on-site study accuracy is 65% and accuracy of the online study is 86.67%. With Naive Bayes, we have the accuracy of 70% in the on-site study and 88.89% in the online study. The accuracies by using SVM for the on-site study and the online study are 70% and 92.22% respectively. Multilayer Perceptron has the highest accuracy in both the on-site study and

the online study. The accuracies are 80% in the on-site study and 96.67% in the online study.

Figure 9 (b) shows the accuracy comparison of four machine learning models using on-site dataset and online dataset with 40 instances. We can see the online study has much better accuracy than the on-site study. For J48, the on-site study accuracy is 65% and accuracy of the online study is 72.5%. With Naive Bayes, we have the accuracy of 70% in the on-site study and 82.5% in the online study. The accuracies by using SVM for the on-site study and the online study are 70% and 85% respectively. Multilayer Perceptron has the highest accuracy in both on-site study and online study. The accuracies are 80% in the on-site study and 90% in the online study. With the same number of instances, the online study still has better prediction performance. The reason is that we the attributes in our online study are more significant than the attributes used in the on-site study.

6. Discussions

In this research, we have collected data from both on-site study and online study. In the on-site study, we applied statistical methods to analyze the data. The on-site study aims at answering the questions regarding intervention, phishing types, and monetary incentive factors. Through statistical methods, we have first analyzed the data collected from the on-site study, where we introduced intervention in the second round. Our analysis demonstrates that the participants with intervention and a monetary incentive perform better than the ones in other cases. Our data analysis also showed the performance of participants in the second round had been improved due to the use of the intervention. However, we noticed that in the third round, some participants' performance was be even worse compared with the one in the first round. We suspect that the worse performance could be thank to the participants' fatigue in the third round. To address this phenomenon, we plan to conduct further experiments in our future research. Because of the limitation of budget and resources, we were only able to recruit 40 participants in the on-site study. To increase the scalability and diversity of participants, we designed the online study and collected data from more participants using Amazon mechanical turk. We applied four machine learning models, J48, Naive Bayes, SVM and Multilayer Perceptron, to predict a participant's performance that was classified as *Good* or *Poor*. Our data analysis results showed that Multilayer Perceptron performed the best where its accuracy was 96.67%. However, there was a weakness. That is, in this study, we did an attribute selection or reduction through fitting all attributes in a linear regression that might cause the problem of multicollinearity because some of the attributes were somewhat correlated. This is due to the small dataset in both our studies, resulting in a limitation in the current research. To address this issue, we plan to recruit more participants in the future research. Furthermore, as we see in section 5, both intervention and the monetary incentive could improve the user performance when dealing with phishing emails. Therefore, these factors could be applied in user education. We could design an education game that can be used to predict the user's performance based on user behaviors by applying our machine learning framework. Then, we could design specific schemes by helping them to be aware of phishing attacks so that they could achieve better performance. We could motivate them by giving them a hint (intervention) or an award (monetary incentive). User education is out of the scope of this research. We leave it in the other paper.

7. Conclusions and Future Work

In this paper, we studied the user behavior related with phishing emails. We did comprehensive and quantitative investigation of how users react in email checking and reading that have become an integral part of our daily life. We have designed two studies, on-site study and online study. We have applied statistical methods to analyze our on-site dataset and explore the answers to the questions on how intervention, phishing types, and a monetary incentive affect user behaviors when phishing attacks are encountered. Our analysis have showed that participants with intervention and a monetary incentive perform better than the ones in other cases. Phishing type 1, suspicious senders' email addresses, tends to be more harmful to users compared to other two phishing types. We have further developed machine learning techniques with the 10-fold cross-validation to analyze the data collected in the online study. We have analyzed the best attributes and found the preferred critical point used in our machine learning framework. By choosing 16 attributes and critical point $c = 30$, we have achieved the user performance prediction accuracies of 86.67%, 88.89%, 92.22%, and 96.67% for J48, Naive Bayes, SVM, and Multilayer Perceptron, respectively.

Based on the findings from our study, we would suggest users pay more attention to the sender's email addresses, links, and contents in the email in order to avoid being a victim from phishing email attacks. In the future, we plan to conduct more experiments and recruit more participants to perform the experiment. In daily-life scenarios, we tend to deal with many other things while checking our emails; thus, we plan to investigate a multitasking experiment platform to understand how multitasking will affect the behavior of a user accordingly besides a couple of future work discussed in section 6.

Acknowledgement. We acknowledge National Science Foundation (NSF) to partially sponsor the work under grants #1620868, #1620871, #1620862, and #1651280. We also thank the Florida Center for Cybersecurity for a seed grant. Moreover, we thank the JHU team that provided their study design document with some data used in this research. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of NSF.

References

[1] Li, Y., Xiong, K. and Li, X. (2019) An analysis of user behaviors in phishing email using machine learning techniques. In *Proceedings of The 16th International Conference on Security and Cryptography (SECRYPT) and the 16th International Joint Conference on e-Business and Telecommunications* (ICETE).

[2] (Accessed May 13, 2018) *Gartner survey shows phishing attacks escalated in 2007; More Than 3 Billion Lost to These Attacks*. https://www.gartner.com/newsroom/id/565125.

[3] Almomani, A., Gupta, B., Atawneh, S., Meulenberg, A. and Almomani, E. (2013) A survey of phishing email filtering techniques. *IEEE communications surveys & tutorials* **15**(4): 2070–2090.

[4] Chin, T.J., Xiong, K. and Hu, C. (2018) Phishlimiter: A phishing detection and mitigation approach using software-defined networking. In *IEEE Access*.

[5] Puthal, D., Nepal, S., Ranjan, R. and Chen, J. (2017) A dynamic prime number based efficient security mechanism for big sensing data streams. *Journal of Computer and System Sciences* **83**(1): 22–42.

[6] Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E. (2007) Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (ACM): 905–914.

[7] Downs, J.S., Holbrook, M. and Cranor, L.F. (2007) Behavioral response to phishing risk. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit* (ACM).

[8] Arachchilage, N.A. and Cole, M. (2011) Design a mobile game for home computer users to prevent from "phishing attacks". In *International Conference on Information Society (i-Society)* (IEEE): 485–489.

[9] Kirlappos, I. and Sasse, M.A. (2012) Security education against phishing: A modest proposal for a major rethink. *IEEE Security & Privacy, 2012* **10**(2).

[10] Yang, W., Chen, J., Xiong, A., Proctor, R.W. and Li, N. (2015) Effectiveness of a phishing warning in field settings. In *Proceedings of the Symposium and Bootcamp on the Science of Security* (ACM): 14.

[11] Brase, G.L. (2009) How different types of participant payments alter task performance. *Judgment and Decision Making* **4**(5): 419.

[12] Drake, C.E., Oliver, J.J. and Koontz, E.J. (Accessed Jan 20, 2020) *Anatomy of a Phishing Email*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.9431&rep=rep1&type=pdf.

[13] Hong, J. (2012) The state of phishing attacks. *Communications of the ACM* **55**(1): 74–81.

[14] Wang, J., Herath, T., Chen, R., Vishwanath, A. and Rao, H.R. (2012) Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE transactions on professional communication* **55**(4): 345–362.

[15] Jang-Jaccard, J. and Nepal, S. (2014) A survey of emerging threats in cybersecurity. *Journal of Computer and System Sciences* **80**(5): 973–993.

[16] Vishwanath, A., Herath, T., Chen, R., Wang, J. and Rao, H.R. (2011) Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems, Elsevier, 2011*, **51**(3).

[17] Dhamija, R., Tygar, J.D. and Hearst, M. (2006) Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (ACM).

[18] Vishwanath, A., Harrison, B. and Ng, Y.J. (2016) Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research, 2016.* .

[19] Alnajim, A. and Munro, M. (2009) An anti-phishing approach that uses training intervention for phishing websites detection. In *Sixth International Conference on Information Technology: New Generations* (IEEE): 405–410.

[20] Burns, M.B., Durcikova, A. and Jenkins, J.L. (2013) What kind of interventions can help users from falling for phishing attempts: a research proposal for examining stage-appropriate interventions. In *46th Hawaii International Conference on System Sciences (HICSS)* (IEEE): 4023–4032.

[21] Liang, H. and Xue, Y. (2010) Understanding security behaviors in personal computer usage: A threat avoidance perspective. *Journal of the Association for Information Systems, 2010*, **11**(7).

[22] Wu, M., Miller, R.C. and Garfinkel, S.L. (2006) Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (ACM).

[23] Downs, J.S., Holbrook, M.B. and Cranor, L.F. (2006) Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security* (ACM).

[24] Stuart, L.M., Park, G., Talor, J.M. and Raskin, V. (2014) On identifying phishing emails: Uncertainty in machine and human judgment. In *IEEE Conference on Norbert Wiener in the 21st Century (21CW)* (IEEE): 1–8.

[25] Ma, L., Torney, R., Watters, P. and Brown, S. (2009) Automatically generating classifier for phishing email prediction. In *10th International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN)* (IEEE): 779–783.

[26] Smadi, S., Aslam, N., Zhang, L., Alasem, R. and Hossain, M. (2015) Detection of phishing emails using data mining algorithms. In *9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)* (IEEE): 1–8.

[27] Shirazi, H., Haefner, K. and Ray, I. (2017) Fresh-phish: A framework for auto-detection of phishing websites. In *IEEE International Conference on Information Reuse and Integration (IRI)* (IEEE): 137–143.

[28] Zeng, Y.G. (2017) Identifying email threats using predictive analysis. In *International Conference on Cyber Security And Protection Of Digital Services (Cyber Security)* (IEEE): 1–2.

[29] Şentürk, Ş., Yerli, E. and Soğukpinar, İ. (2017) Email phishing detection and prevention by using data mining techniques. In *International Conference on Computer Science and Engineering (UBMK)* (IEEE): 707–712.

[30] Park, G., Stuart, L.M., Taylor, J.M. and Raskin, V. (2014) Comparing machine and human ability to detect phishing emails. In *IEEE International Conference on Systems, Man and Cybernetics (SMC)* (IEEE): 2322–2327.

[31] Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M.A. and Pham, T. (2009) School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of Symposium On Usable Privacy and Security (SOUP)* (ACM).

[32] Kawakami, M., Yasuda, H. and Sasaki, R. (2010) Development of an E-learning content-making system for information security (ELSEC) and its application to anti-phishing education. In *International Conference on E-Education, E-Business, E-Management, and E-Learning* (IEEE): 7–11.

[33] Tseng, S.S., Chen, K.Y., Lee, T.J. and Weng, J.F. (2011) Automatic content generation for anti-phishing education game. In *Proceedings of International Conference on Electrical and Computer Engineering (ICECE)* (IEEE).

[34] Stembert, N., Padmos, A., Bargh, M.S., Choenni, S. and Jansen, F. (2015) A study of preventing email (spear) phishing by enabling human intelligence. In *Intelligence and Security Informatics Conference* (IEEE).

[35] Underhay, L., Pretorius, A. and Ojo, S. (2016) Game-based enabled E-learning model for E-Safety education. In *IST-Africa Week Conference* (IEEE).

[36] Muthal, S., Li, S., Huang, Y., Li, X., Dahbura, A., Bos, N. and Molinaro, K. (2017) A phishing study of user behavior with incentive and informed intervention. In *Proceedings of the National Cyber Summit*.

[37] (Accessed February 10, 2018) *Phish Bowl Database*. https://it.cornell.edu/phish-bowl.

[38] (Accessed February 5, 2018) *Amazon Mechanical Turk - Welcome - MTurk*. https://www.mturk.com/mturk/welcome.

[39] Ping Liang and Kaiqi Xiong (1999) On the analysis of neural networks with asymmetric connection weights or noninvertible transfer functions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **29**(5): 632–636.

[40] Sontag, E.D. (1998) VC dimension of neural networks. In *NATO ASI Series F Computer and Systems Sciences*.

[41] Chin, T., Xiong, K., Hu, C. and Li, Y. (2018) A machine learning framework for studying domain generation algorithm (DGA)-based malware. In *International Conference on Security and Privacy in Communication Systems* (Springer): 433–448.

[42] Li, Y., Xiong, K., Chin, T. and Hu, C. (2019) A machine learning framework for domain generation algorithm-based malware detection. *IEEE Access* **7**: 32765–32782.

APPENDICES

A. ON-SITE STUDY SURVEY QUESTIONS:

Age	Age
Gender	Gender
Native_Spk	Are you a native English speaker?
Edu1	Are you currently, or have you previously been enrolled in a computer science/engineering or cybersecurity related degree program?
Edu2	Have you taken any cybersecurity courses?
Habit1	How often do you check your social media accounts?
Habit2	How often do you check your emails?
SelfEff1	Please rate your agreement with the following: I am very confident with my computer skills.
SelfEff2	Please rate your agreement with the following: I consider myself to be a cybersecurity expert.
SelfEff3	I feel confident in my ability to determine which emails are legitimate.
SelfEff4	I believe I was successful in the email sorting task.
Susp1	I was not generally suspicious of the emails.
Susp2	I generally noticed nothing unfamiliar about the emails.
Susp3	Overall, I thought that clicking on links/attachments would not make me vulnerable.
HP1	I skimmed through the emails.
HP2	I briefly looked at the sender/source of the emails.
HP3	I ignored the message content of the emails.
SP1	I thought about the action I took based on what I saw in the email.
SP2	I found myself making connections between the emails' requests and what I have heard about emails requesting such information.
SP3	I spent some time thinking about the request before I made my decision.
RB1	The risk of a security compromise is a lot less on a public computer than your personal computer.
RB2	The risk of a security compromise is a lot more when you click on a link in an email than when you respond to it.
HabStgth1	Checking messages and social media are something I start doing before I realize I'm doing it.

HabStgth2	Checking messages and social media are something I have been doing for a long time.
HabStgth3	Checking messages and social media are something I do automatically.
HabStgth4	Checking messages and social media belong in my daily routine.
HabStgth5	I feel my social media use and the amount of time spent checking messages has gotten out of control.
HabStgth6	I have tried unsuccessfully to cut down the amount of time I spend checking my messages and social media.
HabStgth7	I feel anxious when I am offline without access to messages and social media for an extended period of time.
Study_Know	How did you come to know about our study?

B. ONLINE STUDY SURVEY QUESTIONS:

Pre-Survey Questions:

Age	Age
Gender	Gender
Native_Spk	Are you a native English speaker?
Student	Are you currently a student?
Degree_program	What degree program are you currently enrolled in?
Highest_education	What is your highest level of completed education?
Completed_degrees	What is(are) your completed degree(s) in?
Cyber_experience	Have you completed any network engineering and/or cybersecurity courses or certifications?
Cyber_courses	Please list the cybersecurity courses/certifications you have taken.
Comp_type	What type of computer are you using to complete this experiment?
Input_type	Are you controlling the cursor with an external mouse or a trackpad?
Computer_agreement_1	I am very confident with my computer skills.
Computer_agreement_2	I consider myself to be a cybersecurity expert.
Beliefs_agreement_1	The risk of getting a computer virus is a lot less on a public computer than your personal computer.

Beliefs_agreement_2	The risk of getting a computer virus is a lot less on a mobile device than on a computer.
Beliefs_agreement_3	The risk of getting a computer virus is a lot more when you click on a link in an email than when you open an attachment.
Beliefs_agreement_4	Only Windows machines can get a computer virus.
Beliefs_agreement_5	If you have antivirus or anti-malware software, your computer is completely safe.
Email_habits_agreeme_1	Checking email is something I do frequently.
Email_habits_agreeme_2	Checking email is something I do without having to consciously remember.
Email_habits_agreeme_3	Checking email is something I have no need to think about doing.
Email_habits_agreeme_4	Checking email is something I start doing before I realize I'm doing it.
Email_habits_agreeme_5	Checking email is something I would find hard not to do.
Email_habits_agreeme_6	Checking email is something I have been doing for a long time.
Email_habits_agreeme_7	Checking email is something I do automatically.
Email_habits_agreeme_8	Checking email belongs in my daily routine.
SM_habits_agreement_1	Checking social media is something I do frequently.
SM_habits_agreement_2	Checking social media is something I do without having to consciously remember.
SM_habits_agreement_3	Checking social media is something I have no need to think about doing.
SM_habits_agreement_4	Checking social media is something I start doing before I realize I'm doing it.
SM_habits_agreement_5	Checking social media is something I would find hard not to do.
SM_habits_agreement_6	Checking social media is something I have been doing for a long time.
SM_habits_agreement_7	Checking social media is something I do automatically.
SM_habits_agreement_8	Checking social media belongs in my daily routine.
email_reg_agreement_1	I feel my email use has gotten out of control.
email_reg_agreement_2	I have tried unsuccessfully to cut down the amount of time I spend checking my email.
email_reg_agreement_3	I feel anxious when I am offline without access to email for an extended period of time.

SM_reg_agreement_1	I feel my social media use has gotten out of control.
SM_reg_agreement_2	I have tried unsuccessfully to cut down the amount of time I spend checking my social media.
SM_reg_agreement_3	I feel anxious when I am offline without access to social media for an extended period of time.

Post-Survey Questions:

Sort_correct_1	How many emails do you think you sorted correctly?
Sort_confident_1	How confident are you in your assessment of the number of correctly sorted emails?
Sort_agreement_1	I felt hurried and rushed when sorting emails.
Sort_agreement_2	Completing the email sorting task was mentally demanding.
Sort_agreement_3	It took a lot of effort to sort emails.
Sort_agreement_4	I felt irritated and stressed while sorting emails.
Sort_agreement_5	I spent more time thinking about each email while doing this task than I usually would.
Strat_gen	What is your general strategy for determining if an email was legitimate?
Strat_cues1_1	Importance of Sender Display Name
Strat_cues1_2	Importance of Sender Email Address
Strat_cues1_4	Importance of Hyperlinked URL
Strat_cues1_5	Importance of HTTPS in URL
Strat_cues2_1	Importance of Amount of Logos/Branding
Strat_cues2_2	Importance of Overall Design/Formatting
Strat_cues2_5	Importance of In-email Security Scanning Notices/Indicators
Strat_cues3_1	Importance of Spelling and Grammar Errors
Strat_cues3_2	Importance of Lack of Personalization
Strat_cues3_3	Importance of Type of Information Requested
Strat_cues3_5	Importance of Use of Time Pressure (ex. "you have 24hrs to respond")
Strat_cues3_6	Importance of Use of Threats (ex. threatening legal action)
Strat_cues3_7	Importance of Too Good to be True Offers (ex. you won $4,000,000)

Incent_strat	Did the possibility for a financial incentive change your strategy for identifying suspicious emails?
Incent_strat_change	How did your strategy change?
Incent_agreement_1	I spent more time thinking about each email while doing this task than I usually would.
BFI_BFI_1	I am someone who: Is talkative
BFI_BFI_2	I am someone who: Tends to find fault with others
BFI_BFI_3	I am someone who: Does a thorough job
BFI_BFI_4	I am someone who: Is depressed, blue
BFI_BFI_5	I am someone who: Is original, comes up with new ideas
BFI_BFI_6	I am someone who: Is reserved
BFI_BFI_7	I am someone who: Is helpful and unselfish with others
BFI_BFI_8	I am someone who: Can be somewhat careless
BFI_BFI_9	I am someone who: Is relaxed, handles stress well
BFI_BFI_10	I am someone who: Is curious about many different things
BFI_BFI_11	I am someone who: Is full of energy
BFI_BFI_12	I am someone who: Starts quarrels with others
BFI_BFI_13	I am someone who: Is a reliable worker
BFI_BFI_14	I am someone who: Can be tense
BFI_BFI_15	I am someone who: Is ingenious, a deep thinker
BFI_BFI_16	I am someone who: Generates a lot of enthusiasm
BFI_BFI_17	I am someone who: Has a forgiving nature
BFI_BFI_18	I am someone who: Tends to be disorganized
BFI_BFI_19	I am someone who: Worries a lot
BFI_BFI_20	I am someone who: Has an active imagination
BFI_BFI_21	I am someone who: Tends to be quiet
BFI_BFI_22	I am someone who: Is generally trusting
BFI_BFI_23	I am someone who: Tends to be lazy
BFI_BFI_24	I am someone who: Is emotionally stable, not easily upset

BFI_BFI_25	I am someone who: Is inventive
BFI_BFI_26	I am someone who: Has an assertive personality
BFI_BFI_27	I am someone who: Can be cold and aloof
BFI_BFI_28	I am someone who: Perseveres until the task is finished
BFI_BFI_29	I am someone who: Can be moody
BFI_BFI_30	I am someone who: Values artistic, aesthetic experiences
BFI_BFI_31	I am someone who: Is sometimes shy, inhibited
BFI_BFI_32	I am someone who: Is considerate and kind to almost everyone
BFI_BFI_33	I am someone who: Does things efficiently
BFI_BFI_34	I am someone who: Remains calm in tense situations
BFI_BFI_35	I am someone who: Prefers work that is routine
BFI_BFI_36	I am someone who: Is outgoing, sociable
BFI_BFI_37	I am someone who: Is sometimes rude to others
BFI_BFI_38	I am someone who: Makes plans and follows through with them
BFI_BFI_39	I am someone who: Gets nervous easily
BFI_BFI_40	I am someone who: Likes to reflect, play with ideas
BFI_BFI_41	I am someone who: Has few artistic interests
BFI_BFI_42	I am someone who: Likes to cooperate with others
BFI_BFI_43	I am someone who: Is easily distracted
BFI_BFI_44	I am someone who: Is sophisticated in art, music, or literature

C. Attributes in All Experiments:

Attributes Name:	**Description**	**Study**
Performance Score	Overall performances for the participants	Both
Condition	Condition 0 means this participant did not have monetary incentive, and condition 1 means the participant had monetary incentive	Both
Intervention	In round two, which type of phishing email we gave as an intervention based on the participants' performance of the first round	On-site
R1_P1_Time	Time used for sorting the phishing type 1 emails in round 1	On-site

R1_P1_Score	Score got for sorting the phishing type 1 emails in round 1	On-site
R1_P2_Time	Time used for sorting the phishing type 2 emails in round 1	On-site
R1_P2_Score	Score got for sorting the phishing type 2 emails in round 1	On-site
R1_P3_Time	Time used for sorting the phishing type 3 emails in round 1	On-site
R1_P3_Score	Score got for sorting the phishing type 3 emails in round 1	On-site
R1_Nr_Time	Time used for sorting the normal emails in round 1	On-site
R1_Nr_Score	Score got for sorting the normal emails in round 1	On-site
R1_Time	Time used for sorting all the emails in round 1	On-site
R1_Score	Score got for sorting all the emails in round 1	On-site
R2_P1_Time	Time used for sorting the phishing type 1 emails in round 2	On-site
R2_P1_Score	Score got for sorting the phishing type 1 emails in round 2	On-site
R2_P2_Time	Time used for sorting the phishing type 2 emails in round 2	On-site
R2_P2_Score	Score got for sorting the phishing type 2 emails in round 2	On-site
R2_P3_Time	Time used for sorting the phishing type 3 emails in round 2	On-site
R2_P3_Score	Score got for sorting the phishing type 3 emails in round 2	On-site
R2_Nr_Time	Time used for sorting the normal emails in round 2	On-site
R2_Nr_Score	Score got for sorting the normal emails in round 2	On-site
R2_Time	Time used for sorting all the emails in round 2	On-site
R2_Score	Score got for sorting all the emails in round 2	On-site
R3_P1_Time	Time used for sorting the phishing type 1 emails in round 3	On-site
R3_P1_Score	Score got for sorting the phishing type 1 emails in round 3	On-site
R3_P2_Time	Time used for sorting the phishing type 2 emails in round 3	On-site
R3_P2_Score	Score got for sorting the phishing type 2 emails in round 3	On-site
R3_P3_Time	Time used for sorting the phishing type 3 emails in round 3	On-site
R3_P3_Score	Score got for sorting the phishing type 3 emails in round 3	On-site
R3_Nr_Time	Time used for sorting the normal emails in round 3	On-site
R3_Nr_Score	Score got for sorting the normal emails in round 3	On-site
R3_Time	Time used for sorting all the emails in round 3	On-site
R3_Score	Score got for sorting all the emails in round 3	On-site

R1_Phis_Time	Time used for sorting all phishing emails in round 1	On-site
R1_Phis_Score	Score got for sorting all phishing emails in round 1	On-site
R2_Phis_Time	Time used for sorting all phishing emails in round 2	On-site
R2_Phis_Score	Score got for sorting all phishing emails in round 2	On-site
R3_Phis_Time	Time used for sorting all phishing emails in round 3	On-site
R3_Phis_Score	Score got for sorting all phishing emails in round 3	On-site
Num_sorted	The number of all emails that have been sorted in the task	Online
Phis_sorted	The number of phishing emails that have been sorted	Online
Phis_accuracy	The accuracy of phishing emails that have been sorted	Online
Nr_sorted	The number of normal emails that have been sorted	Online
Nr_accuracy	The accuracy of normal emails that have been sorted	Online
Pay	The amount money paid to participants	Online
Avg_rating	The average rating of confidence level	Online
Median_rating	The median rating of confidence level	Online
All_percent	The ratio of correctly sorted emails to all emails	Online
Phis_percent	The ratio of correctly sorted phishing emails to all emails	Online
Nr_percent	The ratio of correctly sorted normal emails to all emails	Online

Note: The attributes of survey questions can be found in Appendices A and B.