IMAGE SCALING ATTACK SIMULATION: A MEASURE OF STEALTH AND DETECTABILITY

An Undergraduate Research Scholars Thesis

by

DEVON KELLY IV

Submitted to the Engineering Honors in Computer Science and Engineering at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

ENGINEERING HONORS GRADUATE

Approved by Faculty Research Advisors:	Dr. John Andrew Hamilton Dr. Christiana Chamon Garcia

December 2023

Major:	Computer Science

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

I, Devon Kelly IV, certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Faculty Research Advisors prior to the collection of any data used in this final thesis submission.

This project required approval from the Texas A&M University Research Compliance & Biosafety office.

TAMU IRB #: 2023-1081M Approval Date: 11/15/2023 Expiration Date: 11/15/2026

TABLE OF CONTENTS

Page

ABSTRACT 1

ACKNOWLEDGMENTS 3

NOMENCLATURE 4

1. INTRODUCTION 5

    1.1 Social Engineering 5
    1.2 Transition 12

2. METHODS 13

    2.1 Inclusion and Exclusion Criteria 13
    2.2 Survey Setup and Format 13

3. RESULTS 19

    3.1 Responses Count and Minimal Information Section 19
    3.2 Debrief and Following Quiz 21

4. CONCLUSION 24

REFERENCES 26

ABSTRACT

Image Scaling Attack Simulation: a Measure of Stealth and Detectability

Devon Kelly IV
Department of Computer Science and Engineering
Texas A&M University

Faculty Research Advisor: Dr. John Andrew Hamilton
Texas A&M Cybersecurity Center
Texas A&M University

Faculty Research Advisor: Dr. Christiana Chamon Garcia
Department of Computer Science and Engineering
Texas A&M University

Cybersecurity practices require constant effort to be maintained, and one major weakness within the machine learning field is a lack of awareness regarding potential attacks not only in the usage of machine learning models, but in the development process of models as well. It is possible to poison datasets for the benefit of attackers, and for the poor performance of models using data. Previous studies have already determined that preprocessing attacks, such as image scaling attacks, can be difficult to detect both visually and algorithmically. However, there is a lack of emphasis in these studies regarding the real world performance of these attacks and the detectability of the presence of one of these attacks. The purpose of this work is to analyze the relationship between awareness of image scaling attacks with respect to demographic background and experience. We conduct a survey where we gather the subjects' demographics, analyze the subjects' experience in cybersecurity, record their responses to a poorly performing convolutional neural network model that has been unknowingly hindered by an image scaling attack of a used dataset, and note their reactions after we reveal to them that the images used within the broken

models have been attacked. The subjects in our pilot analysis consist of students taking computer science courses and professors in computer science within Texas A&M University. We find in this study that the overall detection rate of the attack is low enough to be viable in a workplace or academic setting, and that after discovery subjects cannot conclusively determine benign images from attacked images.

ACKNOWLEDGMENTS

Contributors

I would like to thank my faculty advisors, Dr. John Andrew Hamilton and Dr. Christiana Chamon Garcia, and my classmate, Sarah Flanery, for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my family for their encouragement and to my partner for her patience and love.

The data analyzed/used for Image Scaling Attack Simulation: a Measure of Stealth and Detectability were provided by anonymous faculty, staff, and student volunteers within the Computer Science field. The analyses depicted in Image Scaling Attack Simulation: a Measure of Stealth and Detectability were conducted in part by the Department of Computer Science and Engineering and these data are unpublished as of the time of publishing of this thesis.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

No funding was provided during the work conducted for the thesis and was completed by the student independently.

NOMENCLATURE

ML	Machine Learning
FaceID	Apple’s facial recognition authentication tool
CNN	Convolutional Neural Network

1. INTRODUCTION

1.1 Social Engineering

Human cognition consists of obliviousness and a tendency to trust one another, rendering humans susceptible to deception, manipulation, and exploitation. Manipulation in today's society most commonly shows itself in the digital world, as humans have not merely adopted social media as a tool for communication and connection, but an integral part of daily life [1–28]. In response to the COVID-19 pandemic, remote communication has become an essential part of everyday life, and social media platforms have increased in popularity. In the professional realm, remote work became normalized with virtual meetings, collaboration tools, and telecommuting technologies [8–10]. Simultaneously, the entertainment industry has also surged in popularity, as streaming services, online gaming, and virtual events such as "Zoom Happy Hour" have found their way into the daily routines of individuals. The routine exchange of images is a feature of contemporary communication, and as pictures are shared and received on a daily basis, their transmission also serves as a vector for social engineering attacks [1–28].

1.1.1 Vulnerabilities Caused by Social Engineering

The crux of social engineering attacks is manipulation, a psychological vulnerability rather than an algorithmic flaw. Through deception, the victim feels rushed to make a decision from a seemingly reliable source, showing that even systems that contain layers of security can be infiltrated through social engineering [1–45]. Despite the frequency of security breaches, not all social engineering attacks are immediately identified, as social engineers disguise their ulterior motives behind phishing emails, phone calls, or even face-to-face encounters [15, 16]. As such, the nature of these attacks raises a demand for constant vigilance and increased layers of protection on cyber-physical systems, including photo editing software.

1.1.1.1 Phishing

In phishing attacks, fraudsters may superimpose legitimate company logos and branding onto emails or websites that harbor malicious intent, luring unsuspecting victims into divulging sensitive information. Likewise, social media manipulation campaigns often rely on the superposition of images to create false narratives. Social engineering attacks that involve scaling superimposed images add an additional layer of deception, where cybercriminals employ image manipulation techniques to carefully adjust the size and proportions of overlaid elements, making them appear consistent with the overall visual context [29–45]. For example, attackers may alter the scale of a superimposed phishing link or a malicious message within an image, making it blend with the surrounding content in an email or a webpage. This attention to detail aims to exploit the human tendency to overlook discrepancies in visual cues.

1.1.1.2 AI Prompt Engineering

Machine learning itself is susceptible to various attacks as well. These often operate in similar ways to social engineering techniques. In large language model based chatbots and other types of prompt related AI, it is possible to trick the model into learning incorrect details [46]. One of the most notorious instances of this type of attack goes back to Microsoft's Tay.AI chatbot on the platform formerly known as Twitter. Through continuous queries involving derogatory material, the chatbot was convinced that the vulgar language and input was valid and should be repeated. In the span of a few days Tay.AI went from behaving in what Microsoft believed to be similar to that of a teenage girl to repeating offensive and vulgar tweets. This happened because the inputs were not sanitized nor monitored to the degree it should have been in retrospect. Despite only being online for a short period of time, Tay.AI had to be shut down since the innocent chatbot had now become a major public relations issue for the company [47, 48]. Even to this day with ChatGPT and other large language models, people can trick these chatbots into repeating vulgar or otherwise prohibited phrases. ChatGPT in particular, although hasn't been as corrupted as Tay.AI, is susceptible to "prompt engineering" in which users can bypass the filters put in place by OpenAI by asking the bot, in very specific terms, to play the role of a character who operates outside of the OpenAI terms of service and morals. Usually the chatbot is not supposed to give instructions on how to perform illegal activities or create malicious software, but through this type of manipulation it is possible to convince the bot to create viruses or explain how to make controlled substances, along with other illegal material that is outside the OpenAI terms of service [49]. However, when it comes to these types of attacks they all stem from a similar source: the inputs to the model themselves are the vulnerable site. Without proper filtering and sanitizing of inputs and trusting the inputs of every user, these open access models have made themselves vulnerable to these types of manipulations. Even outside the context of publicly accessible models, this vulnerability still exists.

1.1.1.3 Superposition

Superposition attacks in the context of social engineering involve the simultaneous execution of multiple deceptive strategies to exploit human cognitive vulnerabilities. By superimposing various social engineering tactics, attackers aim to overwhelm an individual's cognitive defenses, making it more challenging for the target to discern the malicious intent behind the orchestrated campaign. These attacks often exploit trust, authority, or urgency, capitalizing on human instincts and emotions to increase the likelihood of success. As defenders improve their awareness and countermeasures against individual social engineering techniques, the emergence of superposition attacks highlights the need for comprehensive security training and robust cybersecurity protocols to mitigate the evolving and complex nature of social engineering threats.

Superposition attacks in the context of machine learning (ML) refer to a sophisticated class of adversarial techniques aimed at undermining the robustness and reliability of ML models [50]. Unlike traditional adversarial attacks that perturb inputs with imperceptible changes to deceive models, superposition attacks involve overlaying multiple perturbations simultaneously. This method leverages the principles of superposition from combining diverse adversarial signals to create a more potent and challenging attack vector. By simultaneously injecting multiple subtle manipulations into a given input dataset, superposition attacks can exploit vulnerabilities that may not be apparent when considering individual perturbations in isolation. More specifically, by adding similar noise patterns in an image it would be possible to teach a given convolutional neural network model to respond with a certain classification when the noise pattern is present [51]. This type of attack would be generated by methods such as having an active malware injecting the pattern throughout the desired classification's training dataset images, or by directly uploading a dataset with the noise present. Especially if each classification in a given model was given its own noise pattern that is injected, it would be possible to completely control the output of a given convolutional neural networking model by simply adding the noise pattern into the image uploaded to the model for classification. Applications for this attack in particular are devastating. For machine learning powered authentication tools such as TouchID or FaceID, it would be possible to trick similar models into generating an unlock result with a manipulated noise image. This would potentially give an attacker who created the poisoned data a backdoor into their victim's machine learning powered authentication tools. Defending against superposition attacks requires advanced strategies, such as incorporating ensemble models, deploying anomaly detection techniques, or enhancing model interpretability to detect and mitigate these complex and layered adversarial manipulations. As machine learning systems become increasingly integrated into critical applications, understanding and addressing the threat of superposition attacks is crucial for ensuring the security and reliability of these systems.

1.1.2 Image Scaling

Beyond the superposition techniques, cybercriminals also employ image scaling attacks, which consist of using the known dimensions and proportions of an image after preprocessing has occurred within a convolutional neural network model and injecting a secondary image within the pixels that will be selected by the scaling algorithm, often with the goal of distorting or concealing malicious content or otherwise skewing the training results of a model. In the context of machine learning (ML), image scaling takes advantage of the static ways different ML libraries downscale images to fit the model, independent of the input image, whether the libraries use bilinear interpolation or other algorithms such as [30, 31, 37]. Another application includes not only creating incorrect output in a given model, but teaching a model to provide a fixed response when it sees a certain pattern, e.g. injecting a noise image to unlock FaceID or other facial recognition. Consider a scenario where an attacker, posing as a legitimate organization or government entity, sends an email containing a seemingly official document or ID card with altered dimensions. Recipients, trusting the source, may overlook these nuanced changes and unwittingly open the attachment, exposing their systems to malware or falling victim to a phishing scheme. Another example involves the manipulation of product images on e-commerce platforms, where cybercriminals resize and enhance images of counterfeit goods to make them appear genuine, and unsuspecting buyers may make purchases based on these deceptive visuals. Several defense mechanisms have been implemented against such attacks. However, given that such attacks are carried out by humans, and the victims of these attacks are humans, it is unclear whether the ignorance and lack of awareness of humans are contributing factors to the success of such attacks. In the present thesis, we conduct a survey where we analyze the subjects' experience in cybersecurity, record their responses to compromised images as a result of image scaling attacks, and make note of their reactions after we reveal to them that these images have been attacked. The purpose of this survey is to provide a better understanding on how people will likely respond if their dataset has been hijacked through the use of image scaling attacks. This division of sections serves to simulate a realistic scenario where the computer scientist debugging a ResNet model may not be aware of a security breach or a tampered dataset.

Another nuance within the superimposition techniques are the variations between the colors in between inserted pixels and the scaling ratio therein make a difference. If the original image has a small resolution, then the attack pixels will take up much of the pixel space of the image, creating a more detectable attack to the naked human eye. However, with the image quality used in the current year along with the relatively high quality images used in datasets, the scaling ratio becomes closer to and sometimes will exceed a 100:1 scaling ratio. This would mean that only 1 of 100 pixels in a given image will leave behind artifacts of an attack. When viewing these higher quality attacked images, the viewing tool may scale down the image as well, further obfuscating the attack, since if a different scaling algorithm is used for a different target size, it is likely that the attack pixels will not be used in the shrunk image being used for viewing. Not only that but if the injected images are chosen right, then the resulting image will blend well with the original victim image. This is because if a like colored image is injected into a given victim image, the artifact pixels which display evidence of an attack will be hidden due to the similar colors within the injected pixels versus the original pixels. In other words, if the colors between the anticipated victim image and the injected image are similar, they blend well together due to the small difference in color values. Through the manipulation of these images in adaptive ways such as attacking specific high quality image, increasing the scaling ratio by doing so with a high quality image, and by choosing a target image that blends well with the original, human detection of this attack can be significantly mitigated since the artifacts left behind by the attack will be small and similar to the image itself.

1.1.2.1 Current Defenses

Several defense mechanisms have been proposed against image scaling attacks. However, such defenses have yet to be implemented. Whether due to a lack of human awareness of the preprocessing vulnerabilities, or due to a lower perception of threat from these attacks, most environments have not yet employed respective defense schemes. Some of the currently proposed defenses against image scaling attacks include a frequency analysis. Since the defender also knows the model information, it is possible to know ahead of time how pixel selection works with the scaling algorithm chosen, and the expected locations of anomalous pixel peaks can be approximated ahead of time. This allows the defender to lay out a trap using this information by storing the expected attack locations ahead of time, then this defense strategy measures the various peaks of pixel colors and compares the distance of the peaks to that of where the peaks are expected to be in an attack. If the average distances found through the process are low, then an attack is likely. This defense seemed to work against attacks that were injected across the whole pixel space of an image, named global attacks in the paper, and local attacks, which only inject a certain subset of the pixel space [52]. Another defense strategy proposed by the same paper suggests performing the downscale operation, following up with the inverse upscale operation. The resulting image of this can be compared to the original to see if there are major statistical differences between the two images. This strategy proved successful for the aforementioned global attacks, for the whole image changed as a result of the attack because of the significant differences. However, against a local attack such as the left quadrant of an image being coated with a pattern, this defense did not perform as well. This could have been due to much of the resulting image being left intact after the injection attack. The third type of defense proposed in the paper was to use a filter to try to clean a potentially attacked image. This is done by choosing between different types of potential filtering algorithms, then applying them around targeted areas within the image itself. In theory, this would remove a lot of the attacked pixels from the pixel space altogether. With this type of defense, it would prevent and mitigate the effects of a potential attack. The performance of this strategy was largely effective on the global attacks, where the entire pixel space was attacked. However, performance dampened during the local, limited pixel space attacks. To prevent the adaptive nature of attackers from bypassing one of these chosen defenses alone, it was suggested that these defenses be used in ensemble with each other. The success of such a strategy was effective in both scenarios, i.e. the global attacks and the local attacks [52].

The reason an ensemble was recommended earlier was because if the given attacker becomes aware of the mitigation strategy, then the attacker can simply use knowledge regarding the defense to subvert it through a more adaptive attack. In the instance of frequency analysis alone, a strategy would be to try to use more camouflaged images such that many of the injected pixels do not register as a peak, increasing the average distance and potentially throwing off the resulting conclusion regarding the status of the images. It becomes less obvious to determine the strategy within an ensemble defense, not to mention the difficulty required to bypass such a defense also becomes much higher.

When discussing details regarding the effectiveness of the defenses come the tradeoff. That being with a more complex preprocessing protection measure comes an increase in runtime. In particular, Quiring et al.'s earlier paper compares the runtime performance of lazier scaling algorithms such as nearest neighbor scaling versus filter based scaling versus the VGG19 model, showing that runtimes shift by factors of 10 as they become more complex, specifically nearest scaling ran below $10\hat{3}$ microseconds consistently but all other complex ones had runtimes ranging from $10\hat{4}$ and $10\hat{5}$ microseconds [30].

1.2 Transition

The rest of this thesis is organized as follows. Section 2 describes the methodology used to carry out the survey, Section 3 displays the results of the survey, and Section 4 concludes this thesis with discussion on the survey results.

2. METHODS

The subjects are asked to fill out a Qualtrics survey consisting of the following sections: Demographics and Background, Minimal Information Survey, Debrief, and follow-up survey. The multiphase approach was used to separate portions of the survey where details of what occurs during the simulation are left to speculation initially versus afterwards where questions will be asked after all details have been revealed to the subjects. This way it is possible to extract information from the subjects regarding their responsiveness to image scaling attacks prior to discovery and post detection performance in identifying the attack.

2.1 Inclusion and Exclusion Criteria

The target audience of the present study is participants that have experience in the software development and data science fields. The inclusion criteria consists of faculty, staff, and students who are at least 18 years old and pursuing a degree in computer science or a field related to machine learning. We would like to note that no personally identifying information will be collected, i.e. the survey is completely anonymous. The survey is advertised through emails sent to the relevant departmental Listservs and in-person within classes that allow advertisement during the designated lecture and lab periods. In a future study this would be broadened to include willing companies, their employees, as well as other universities and graduate students from them as well.

2.2 Survey Setup and Format

The subjects are asked to fill out a Qualtrics survey consisting of the following sections: Demographics and Background, Minimal Information Survey, Debrief, and follow-up survey. The multiphase approach was used to separate portions of the survey where details of what occurs during the simulation are left to speculation initially versus afterwards where questions will be asked after all details have been revealed to the subjects. This way it is possible to extract information from the subjects regarding their responsiveness to image scaling attacks prior to discovery and post detection performance in identifying the attack.

2.2.1 Demographics and Background

The Demographics and Background section gathers the age, gender, race, and experience in computer science, machine learning, and cybersecurity whether from work, academia, military, law enforcement, and/or hobbyism. These data points allow for the breakdown of data along these metrics with a sizable amount of data points. The demographics this study specifically was seeking was a large variation in years of experience and within academic credentials. The goal of asking these questions was to attempt to create a breakdown of the data, given that there are enough data points to do so reliably, thereby potentially lending analysis as to how different amounts of experience and academic credentials may affect the future results in other sections of the survey.

2.2.2 Minimal Information Survey

The Minimal Information Survey begins with a short explanation of the simulated scenario: the subject is attempting to figure out why a ResNet model is performing poorly on a cats and dogs dataset [53]. The survey inquires the subjects on what part of the model would be causing problems: the model itself, the hyperparameters used, the training images from the dataset, or no issue at all, and why the subject made their selection. To assist in the selection process, subjects are given the detailed model layout (Tensorflow's implementation of ResNet50, a GlobalAveragePooling2D layer, a 128 node dense layer, and a 1 node dense layer), knowledge that the starting weights are the same as the subject's simulated coworker, and a specific selection of the training images to supplement their decision-making process. All four of the images provided to subjects in this selection have been injected with image scaling attacks with around a 100:1 scaling ratio to match the standard quality of dataset images and the resulting ratio created after downscaling. Not only that, but the images used within this section of the survey have been injected with contrasting color schemes such that the attack would represent the worst case of a random injection method: matching two oppositely color schemed images from opposite datasets to be injected into one another. This is because in the case where an attacker is using an automated attack approach relying on random matching of opposing datasets as proposed in this simulation, it would be better to measure the worst case where the produced images do not blend well with the background image. The goal of the ranking question and the following explanation is to gather data specifically regarding the thought process of subjects on a quantitative level. Through these questions the goal is to extract if the attack is indirectly noticed versus directly found or missed entirely.

We would like to note that in this phase, the subjects will not be told that several images have been tampered with. Only after completion of the initial questions will they be informed that the images had been tampered with via image scaling attacks. An example is shown in Figure 1, where a cat photo has been injected into a dog photo, yet the resulting photo appears visually to be the dog, as it has not been scaled down. Similar images to that of the attacked image from Figure 1 have been chosen, such that the misclassified images represent the most clear instances of image scaling to see if subjects respond that the images in question have been attacked or modified by a third party.

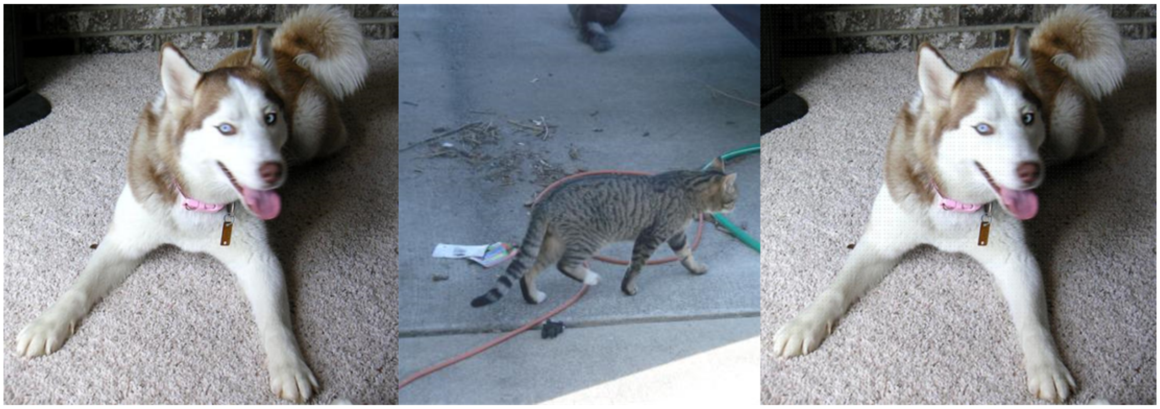
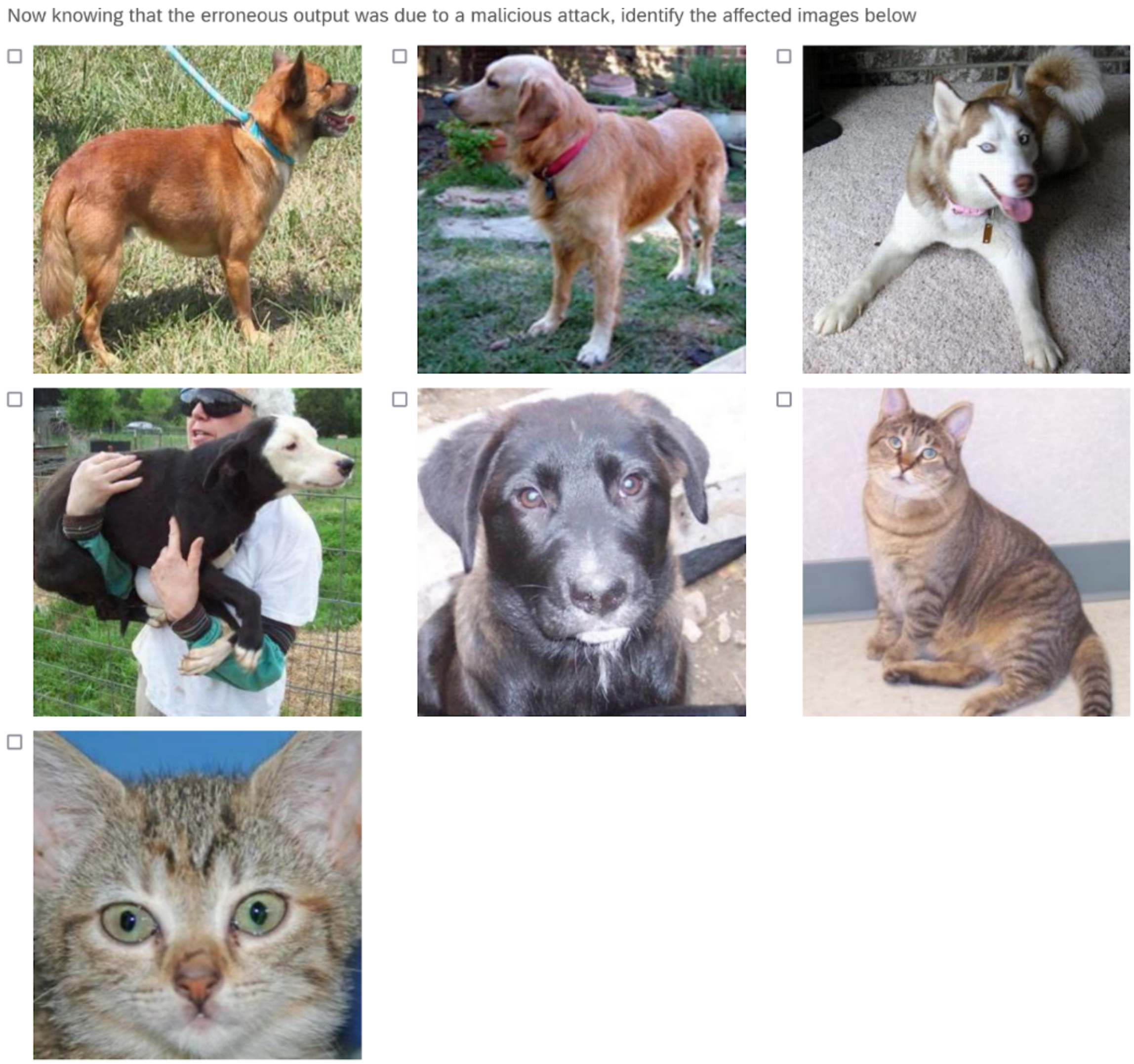


Figure 1: Example of an image infected by an image scaling attack. The original image (left) is a photograph of a dog. The middle image is injected into the attacked image (right). When scaled down, the attacked image becomes an image of a cat (middle).

2.2.3 Debrief and Post Quiz

The debrief section of the survey observes the awareness of the subject to image scaling attacks along with other attacks that target the preprocessing phase of machine learning models to manipulate results. In this phase, subjects are given the explanation that some of the images used in the inception of the model have been compromised by a third party, causing the problems mentioned in the Minimal Information phase–more specifically, subjects will be informed about image scaling attacks and how they operate specifically. They are told about the small artifacts within compromised images and are told those dots are directly caused by the attack, as when pieced together those dots form the target image just as displayed in Figure 1. They are told that this type of attack is done to confuse and manipulate the results of various convolutional neural networking models, and the potential implications that this vulnerability holds. Subjects are then further informed that some images have been affected by this attack (i.e. pictures of dogs appear to the model as a cat and vice versa), and that the reason for the poor performance of the model can be explained by the failed recognition of these images after their subsequent downscaling and transformation into the opposite classification (dog to cat or vice versa). The subjects will then be inquired on their prior awareness of image scaling attacks. Subjects will also be asked whether an attack was considered as a possibility during the Minimal Information phase. This information can be used to compare to how and why the results of the previous questions in the minimal information phase came to be by putting the responses into a better context.

After the debrief, subjects will take a multiple-choice quiz on more images from the dataset, which are different from those which have been seen prior to this section in the explanations and in the introductory dataset. The quiz will consist of three infected images and four benign images, as shown in Figure 2, and the result will determine whether they can identify attacked images after being aware of image scaling attacks and their presence, along with the details regarding ways to determine if an image has been infected. The post-debrief seeks to analyze how now-aware subjects are able to determine the integrity of potentially attacked images after gaining awareness regarding the type of attack that has been performed along with the details required to detect indicators of the attack themselves for the purpose of determining how effective the attack is post-discovery of a breach. Although this has been the emphasis of previous papers, comparing the results of those who have just learned about the attack versus those who had just learned about this vulnerability would be the novelty of including it in this survey in particular.


Now knowing that the erroneous output was due to a malicious attack, identify the affected images below

□ □ □

□ □ □

□

Figure 2: Selected images for the quiz, in this non-randomized order the first 3 images have been injected via image scaling attacks. The rest of the 7 images are benign and unaffected. When presented to users, the order of the selection is randomized. Users are tasked with selecting the attacked images and leaving the benign images unchecked. A zooming tool is also provided in the survey.

By collecting responses from subjects with varying levels of academic and industry experience, we analyze the true detectability of image scaling attacks. We also analyze each groups' ability to identify such attacks, thus indicating the overall population's awareness of such attacks during the training step of machine learning.

3. RESULTS

3.1 Responses Count and Minimal Information Section

At this current time, 129 responses have been gathered, but only 78 of those meet the qualifying requirements outlined in Section 2. Figure 3 illustrates a quantitative comparison of what users believed to be the issue with the image in the Minimal Information survey (see Section 2). Table 1 displays the statistics of the same data points, showing the mean ranking of each selection along with standard deviation, which displays the overall presence of each factor across all rankings. This was done with the purpose of reducing the likelihood of an error influenced by outlying results such as subjects choosing either most likely or least likely without accounting for the likelihood of the intermediate options. From the data provided by the subjects, most of them did not have much industry experience, with the average years of experience being .38. Despite that, most of the respondents had above 3 years of experience academically in the computer science field, with an average of 3.29. This is also reflected in the academic credentials question, where 92.13% of respondents are currently in college for their Bachelors degree, with 7.09% of respondents having their undergraduate degree, and 1 respondent having their doctorate degree.

The metrics in Figure 3 and Table 1 indicate that most subjects found an issue with the training images, but despite that, Figure 4 shows that only 2 directly pointed out that the images had been tampered with, and 5 noted that the noise in the image may have thrown the model off. The remainder of the explanations for their selections included comments such as the dogs had too many cat features and the cats had too many dog features, that there was a sampling bias, and other reasons that did not indicate perception of an attack having been implemented. This would Indicate that less than 10% of subjects noticed any of the manipulation done to the image dataset with all photos shown to subjects being infected in an attempt to show an exaggerated attack with greater artifacts. Additionally, only 2 out of the 78 subjects noted that the attack was present, thus in the present experiment, the attack had a true discovery rate of 3%.

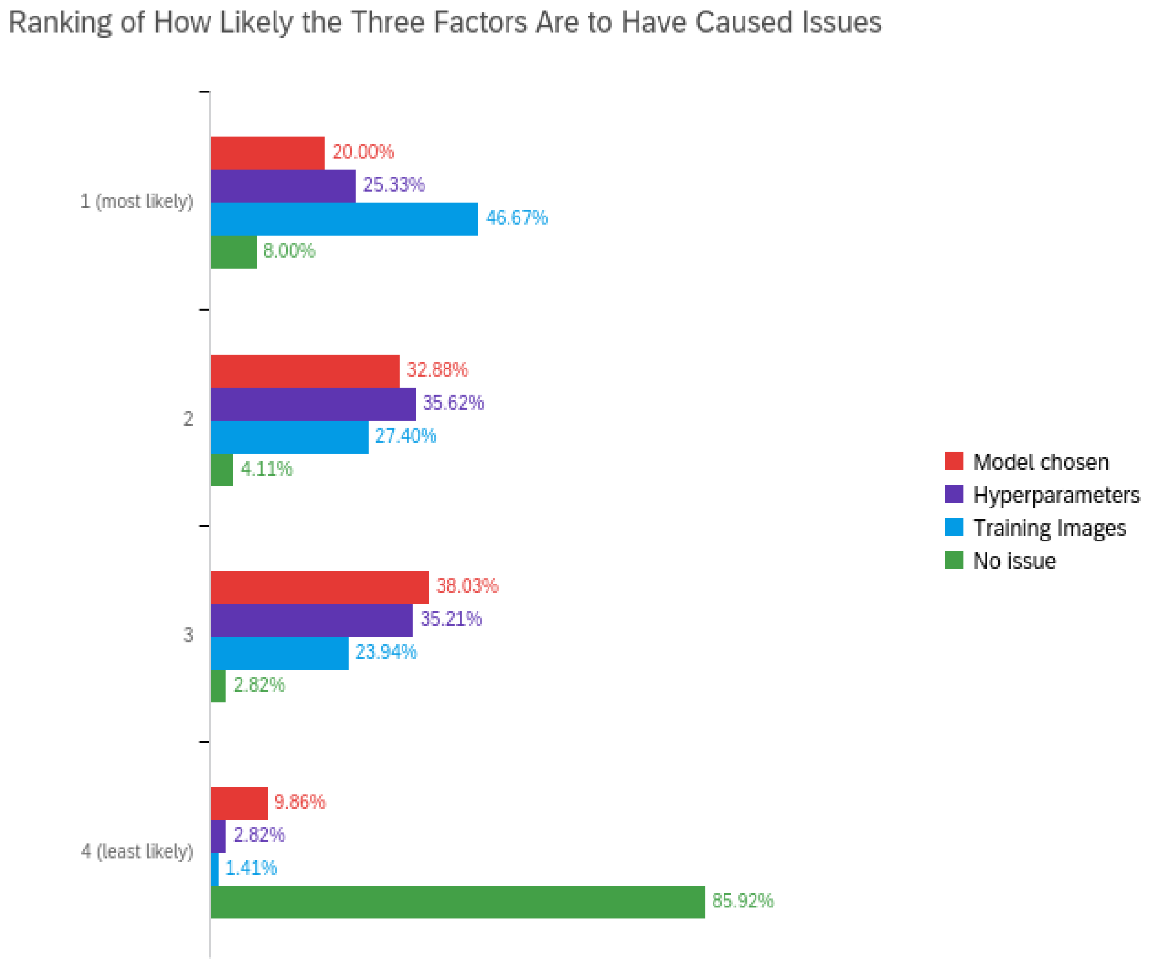


Figure 3: Percentage of users responses to the ranking question regarding asking users to evaluate likely causes of the issue in the simulation during the minimal information section.

Field	Minimum	Maximum	Mean	Std Deviation	Variance
Training Images	1.00	4.00	1.78	0.85	0.72
Hyperparameters	0.00	4.00	2.11	0.87	0.76
Model Chosen	1.00	4.00	2.36	0.91	0.83
No Issue	0.00	4.00	3.59	0.99	0.98

Table 1: Statistical information of the data in Figure 2, specifically focusing on the mean ranking of each potential issue from least likely (bottom) to most likely (top), as indicated by the black arrow.

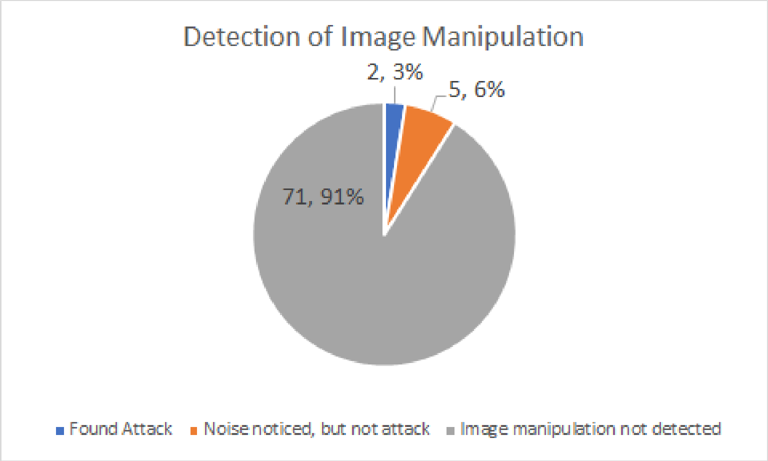


Figure 4: Percentage of users who detected problems with the dataset or immediately detected the attack. This data was taken from the portion of the survey asking for the explanation as to why their selection was chosen.

3.2 Debrief and Following Quiz

After explaining to the subjects what image scaling attacks are and how this form of attack had been applied to the dataset as per the Debrief section (see Section 2), subjects were quizzed on which three of the seven images had been affected by image scaling. As shown in Figure 5, which displays the results of the quiz question, about 57.29% of subjects correctly identified the three attacked images. However, about 37.11% of subjects incorrectly marked the remaining 4 benign images as attacked. These metrics were calculated by averaging the results of the first three images and the last four respectively from Figure 5. This high amount of false positives and false negatives indicate that even after detection, many images still blend in, evading detection. Not only that, but the false positive marking of benign images after detection could be caused by a general paranoia of the noise being present. In a scenario where the attack images were chosen in a way that better blends the target image in visually, it would be likely that the numbers would stratify more with more false negatives and false positives respectively. This is because subjects are able to determine the infected nature of the images through the artifacts left behind by the attack itself. However, if a more well-blended output image was created through smart pairing of original images and target images, then the artifact traces would be less clear to both the human eye and any algorithm attempting to detect the attack's presence.

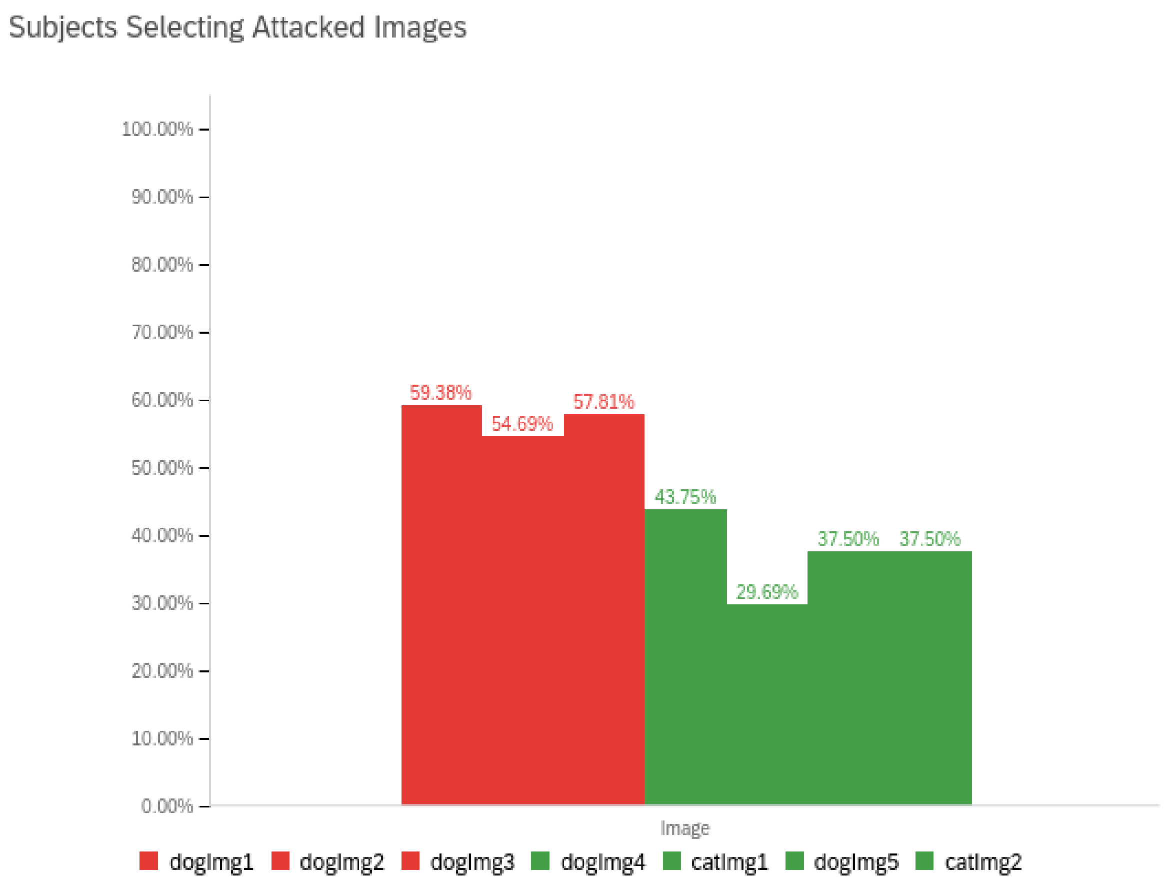


Figure 5: Percentage of respondents that have identified an attacked/benign image. The red bars represent the correct identification of the images that were attacked, and green bars represent the correct identification of the benign images. Let it be noted that the order of options were randomized for subjects while taking the survey.

When asked regarding the thought process that subjects had during their deliberation throughout the first section, 48.72% of subjects claim to have considered an attack being present in the minimal information section. Subjects were also asked as to their prior familiarity with this type of attack, to which only 21.79% of subjects were familiar with Image Scaling (or similar preprocessing) attacks. Interestingly, Figure 6 shows that subjects who had only just become aware of this type of attack had a higher rate of detecting the attacked images. This came with the trade off of subjects who just learned about the attack had a higher false positive rate than that of the group of subjects with prior knowledge of the attack.

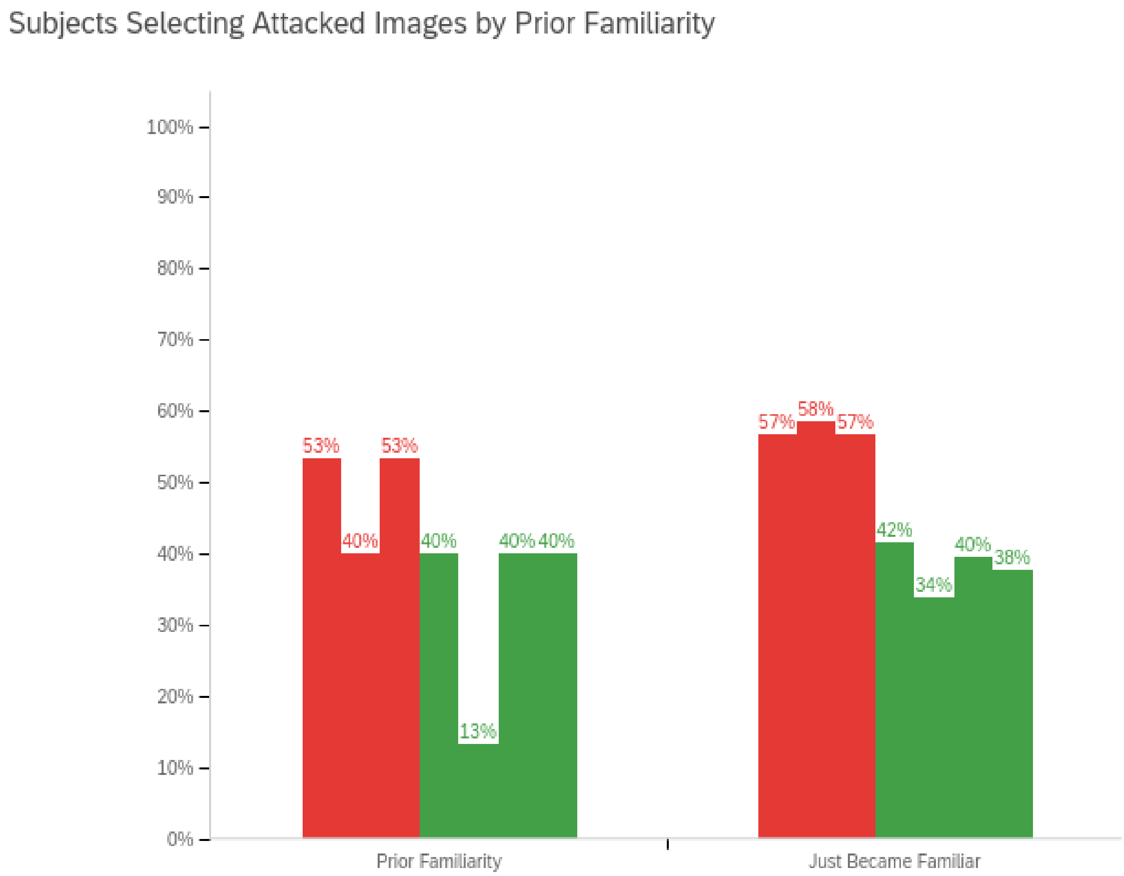


Figure 6: Percentage of respondents that have identified an attacked/benign image split across subjects who had prior familiarity with image scaling attacks versus those who had gained familiarity with this type of attack during the survey. The red bars represent the correct identification of the images that were attacked, and green bars represent the correct identification of the benign images. Let it be noted that the order of options were randomized for subjects while taking the survey.

4. CONCLUSION

We conducted a survey to simulate the debugging experience of a neural network model while under an Image Scaling attack. We performed the survey by guiding subjects through a scenario which described the model's setup, parameters, and input. The input in this case had been modified through Image Scaling attacks, specifically choosing images that would display the worst case result of a random selection of images from opposing datasets. When subjects were prompted regarding which part of the setup was the problem, most subjects identified that the images were the source of the problem, yet only 9% of them noticed the noise in the image from the image scaling attack, and only 2 directly stated that the images had been tampered with.

With the low detection rate of the attack and even lower detection of the noise artifacts in the image, this experiment suggests that the image scaling attack is effective in a workplace environment, especially with the low familiarity that the subjects had with the ability to attack the input of machine learning models.

We also showed that, after detection, subjects have difficulty differentiating attacked images from benign images, which rendered both high false positive and high false negative rates. To the extent that a little under half of the subjects failed to identify the attacked image, and a similar proportion of subjects marked the benign images as attacked. This indicates that the attack is not only effective from a detection standpoint, but that post-detection it is hard to determine the integrity of images.

If a potential attacker is to use an image scaling attack, whether through a malware that mixes dataset images together, by releasing poisoned datasets, or other means, then there would be significant damage to the workplace productivity for the models attempting to use the tampered images. It would also potentially cause many images to be falsely removed for being potentially tampered with post-discovery, or eliminating good datasets for being falsely assumed to be infected.

Future work would involve an expansion of the target audience to include a larger sample size for those who have experience in cybersecurity, as well as a greater diversity in industry experience and academic credentials. This would account for the cohesivity and reproducibility of survey results for more diversity in experience with respect to the population. With a larger dataset of subjects, it would be possible to add granularity to the survey metrics and analyze them more reliably. For instance, it would be more feasible to try to indicate if having certain certifications, years of experience, or other credentials would indicate likelihood of detection in lieu of aggregating the responses to binary "yes" and "no" categories. A future study could also implement using a full office environment and potentially a team trying to debug the given machine learning model, as this study could provide insight on how Image Scaling attacks perform in a realistic office environment across such groups of people. Another interesting approach a future study could use would be to implement a fully automated attack through a malware or other tool that actively searches for matching images to better camouflage the target images within the victim image, that way not only could human detection be potentially evaded but maybe even algorithmic detection methods. One final addition for a future work would be to experiment with the awareness of more local image scaling attacks, since if only a portion of the image space is injected into the overall impact of the artifacts are minimized. It has already been shown that the algorithmic detection of this is decreased when only a portion of the image is attacked, but if presented to a human for a determination, the difficulty would potentially be scaled as well since the human would theoretically not be able to figure out which parts of the image have been left alone versus potentially attacked.

REFERENCES

[1] Z. Wang, H. Zhu, and L. Su, “Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods,” *IEEE Access*, vol. 9, pp. 11895–11910, 2021.

[2] R. Montañez, E. Golob, and S. Xu, “Human cognition through the lens of social engineering cyberattacks,” *Frontiers in Psychology*, vol. 11, 2020.

[3] M. A. Siddiqi, W. Pak, and M. A. Siddiqi, “A study on the psychology of social engineering-based cyberattacks and existing countermeasures,” *Applied Sciences*, vol. 12, no. 12, p. 6042, 2022.

[4] K. Weber, A. E. Schütz, T. Fertig, and N. H. Müller, “Exploiting the human factor: Social engineering attacks on cryptocurrency users,” *Learning and Collaboration Technologies. Human and Technology Ecosystems*, pp. 650–668, 2020.

[5] S. Lineberry, “The human element: The weakest link in information security,” *Journal of Accountancy*, vol. 204, no. 5, p. 44, 2007.

[6] S. M. Albladi and G. R. S. Weir, “Predicting individuals’ vulnerability to social engineering in social networks,” *Cybersecurity*, vol. 3, no. 1, 2020.

[7] R. Heartfield and G. Loukas, “Detecting semantic social engineering attacks with the weakest link: Implementation and empirical evaluation of a human-as-a-security-sensor framework,” *Computers & Security*, vol. 76, pp. 101–127, 2018.

[8] S. Venkatesha, K. R. Reddy, and B. R. Chandavarkar, “Social engineering attacks during the covid-19 pandemic,” *SN Computer Science*, vol. 2, no. 2, 2021.

[9] C. M. Williams, R. Chaturvedi, and K. Chakravarthy, “Cybersecurity risks in a pandemic,” *Journal of Medical Internet Research*, vol. 22, no. 9, p. 23692, 2020.

[10] M. Hijji and G. Alam, “A multivocal literature review on growing social engineering based cyber-attacks/threats during the covid-19 pandemic: Challenges and prospective solutions,” *IEEE Access*, vol. 9, pp. 7152–7169, 2021.

[11] E. U. Osuagwu, G. A. Chukwudebe, T. Salihu, and V. N. Chukwudebe, “Mitigating social engineering for improved cybersecurity,” in *2015 International Conference on Cyberspace (CYBER-Abuja)*, pp. 91–100, 2015.

[12] W. Syafitri, Z. Shukur, U. A. Mokhtar, R. Sulaiman, and M. A. Ibrahim, “Social engineering attacks prevention: A systematic literature review,” *IEEE Access*, vol. 10, pp. 39325–39343, 2022.

[13] E. Ivanova, “Internet addiction and cyberchondria - their relationship with well-being,” *Journal of Education Culture and Society*, vol. 4, no. 1, pp. 57—70, 2020.

[14] K. Krombholz, H. Hobel, M. Huber, and E. Weippl, “Advanced social engineering attacks,” *Journal of Information Security and Applications*, vol. 22, pp. 113—122, 2015.

[15] D. B. Resnik and P. R. Finn, “Ethics and phishing experiments,” *Science and Engineering Ethics*, vol. 24, no. 4, pp. 1241—1252, 2017.

[16] P. Lawson, C. J. Pearson, A. Crowson, and C. B. Mayhorn, “Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy,” *Applied Ergonomics*, vol. 86, p. 103084, 2020.

[17] J. Chigada and R. Madzinga, “Cyberattacks and threats during covid-19: A systematic literature review,” *SA Journal of Information Management*, vol. 23, no. 1, 2021.

[18] A. Yasin, R. Fatima, L. Liu, J. Wang, R. Ali, and Z. Wei, “Understanding and deciphering of social engineering attack scenarios,” *Security and Privacy*, 2021.

[19] F. Mouton, L. Leenen, M. M. Malan, and H. S. Venter, “Towards an ontological model defining the social engineering domain,” *IFIP Advances in Information and Communication Technology*, vol. 431, pp. 266–279, 2014.

[20] C. Campbell, “Solutions for counteracting human deception in social engineering attacks,” *Information Technology & People*, vol. 32, 2018.

[21] I. Gulenko, “Social against social engineering,” *Information Management & Computer Security*, vol. 21, pp. 91–101, Jan 2013.

[22] R. Heartfield and G. Loukas, “A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks,” *ACM Comput. Surv.*, vol. 48, dec 2015.

[23] M. S. Jalali, M. Bruckes, D. Westmattelmann, and G. Schewe, “Why employees (still) click on phishing links: Investigation in hospitals,” *J Med Internet Res*, vol. 22, p. e16775, Jan 2020.

[24] A. Ferreira and G. Lenzini, “An analysis of social engineering principles in effective phishing,” in *2015 Workshop on Socio-Technical Aspects in Security and Trust*, pp. 9–16, 2015.

[25] A. Smith, M. Papadaki, and S. M. Furnell, “Improving awareness of social engineering attacks,” in *Information Assurance and Security Education and Training* (R. C. Dodge and L. Futcher, eds.), (Berlin, Heidelberg), pp. 249–256, Springer Berlin Heidelberg, 2013.

[26] D. Airehrour, N. Vasudevan Nair, and S. Madanian, “Social engineering attacks and countermeasures in the new zealand banking system: Advancing a user-reflective mitigation model,” *Information*, vol. 9, no. 5, 2018.

[27] H. Aldawood and G. Skinner, “Reviewing cyber security social engineering training and awareness programs—pitfalls and ongoing issues,” *Future Internet*, vol. 11, no. 3, 2019.

[28] H. Aldawood and G. Skinner, “Educating and raising awareness on cyber security social engineering: A literature review,” in *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pp. 62–68, 2018.

[29] E. Quiring and K. Rieck, “Backdooring and poisoning neural networks with image-scaling attacks,” in *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 41–47, 2020.

[30] E. Quiring, D. Klein, D. Arp, M. Johns, and K. Rieck, “Adversarial preprocessing: Understanding and preventing Image-Scaling attacks in machine learning,” in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1363–1380, USENIX Association, Aug. 2020.

[31] Y. Gao, I. Shumailov, and K. Fawaz, “Rethinking image-scaling attacks: The interplay between vulnerabilities in machine learning systems,” in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 7102–7121, PMLR, 17–23 Jul 2022.

[32] B. Kim, A. Abuadbba, Y. Gao, Y. Zheng, M. E. Ahmed, S. Nepal, and H. Kim, “Decamouflage: A framework to detect image-scaling attacks on cnn,” in *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 63–74, 2021.

[33] Q. Xiao, Y. Chen, C. Shen, Y. Chen, and K. Li, “Seeing is not believing: Camouflage attacks on image scaling algorithms,” in *28th USENIX Security Symposium (USENIX Security 19)*, (Santa Clara, CA), pp. 443–460, USENIX Association, Aug. 2019.

[34] A. Krizhevsky, “Learning multiple layers of features from tiny images.” University of Toronto, 2009.

[35] V. Chandrasekaran, C. Gao, B. Tang, K. Fawaz, S. Jha, and S. Banerjee, “Face-off: Adversarial face obfuscation,” in *Proceedings on Privacy Enchancing Technologies Symposium*, pp. 369–390, 2021.

[36] V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. Dickerson, G. Taylor, and T. Goldstein, “Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition,” 2021.

[37] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, “Evasion attacks against machine learning at test time,” in *Machine Learning and Knowledge Discovery in Databases* (H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, eds.), (Berlin, Heidelberg), pp. 387–402, Springer Berlin Heidelberg, 2013.

[38] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

[39] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, “Countering adversarial images using input transformations,” 2018.

[40] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, “Fawkes: Protecting privacy against unauthorized deep learning models,” in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1589–1604, USENIX Association, Aug. 2020.

[41] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, “Synthesizing robust adversarial examples,” in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 284–293, PMLR, 10–15 Jul 2018.

[42] J. Cohen, E. Rosenfeld, and Z. Kolter, “Certified adversarial robustness via randomized smoothing,” in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, PMLR, 09–15 Jun 2019.

[43] A. Rahmati, S.-M. Moosavi-Dezfooli, P. Frossard, and H. Dai, "Geoda: A geometric framework for black-box adversarial attacks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8443–8452, 2020.

[44] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 2484–2493, PMLR, 09–15 Jun 2019.

[45] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 3247–3258, PMLR, 13–18 Jul 2020.

[46] C. Weeks, A. Cheruvu, S. M. Abdullah, S. Kanchi, D. Yao, and B. Viswanath, "A first look at toxicity injection attacks on open-domain chatbots," in *Proceedings of the 39th Annual Computer Security Applications Conference*, ACSAC '23, (New York, NY, USA), p. 521–534, Association for Computing Machinery, 2023.

[47] C. Sinders, "Microsoft's tay is an example of bad design." Medium, 2016.

[48] H. Reese, "Why microsoft's 'tay' ai bot went wrong." TechRepublic, 2016.

[49] A. Ox, "Understanding dan prompts for chatgpt." Medium, 2023.

[50] S. Saxena, "A review of adversarial attacks on machine learning algorithms," 2023.

[51] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv:1412.6572*, 2015.

[52] E. Quiring, A. Müller, and K. Rieck, "On the detection of image-scaling attacks in machine learning," in *Proceedings of the 39th Annual Computer Security Applications Conference*, ACSAC '23, (New York, NY, USA), p. 506–520, Association for Computing Machinery, 2023.

[53] "Kaggle cats and dogs dataset." Microsoft Download Center.