

# The Ineffectiveness of Anti-Phishing Tools on Perturbed Logos

Faith Chernowski, Reagan Sanz, Radhika Garg  
 Department of Computer Science  
 University of Tennessee  
 Knoxville, TN, USA  
 Email: {fchernow, rsanz, [rgarg4@vols.utk.edu](mailto:rgarg4@vols.utk.edu)}

**Abstract**— Phishing Attacks continue to be one of the most common cyberattacks online. Anti-Phishing tools that utilize AI image classifiers are often used to combat the increased quality and quantity of phishing websites by identifying logos that correspond to popular brands and companies. However, when small levels of perturbation are added to the phishing websites' logos, the image classifiers may fail to classify the stolen logos. The minor level of noise added to these logos can often be imperceptible to website users. This can result in dangerous websites avoiding detection by anti-phishing tools and continuing to acquire private information from victims. This is especially true for people unfamiliar with image perturbation methods, those who are older, and those who use classes for near-range sight.

**Index Terms**— AI, Phishing, Perturbed Noise, PGD, Human Factors

## 1 INTRODUCTION

Phishing attacks are among the most prevalent and damaging cyber threats today, targeting individuals and organizations alike. Leveraging social engineering, phishing schemes manipulate users into revealing sensitive information such as login credentials, financial data, or personal details. Among the tactics employed, visual deception has emerged as a particularly effective strategy. Attackers often utilize logos and branding elements to fabricate fake websites, emails, and advertisements that appear convincingly legitimate.

As the sophistication of phishing attacks increases, anti-phishing tools have evolved to incorporate artificial intelligence (AI) and machine learning models. These tools often rely on visual similarity detection algorithms, using convolutional neural networks (CNNs) to identify logos and branding features that signal malicious intent. However, the adversarial capabilities of these systems remain underexplored, particularly their susceptibility to perturbed images.

Perturbation techniques involve introducing imperceptible noise to images, altering their pixel values while preserving their visual structure. This subtle manipulation can render logos unrecognizable to AI-based classifiers, even as they remain

visually identical to human observers. For phishing campaigns, this represents a potent avenue for evading detection. If attackers can systematically bypass anti-phishing tools using perturbed logos, the efficacy of these systems could be severely compromised.

The objective of this research is to assess the vulnerability of AI-based anti-phishing tools to perturbed logos and to evaluate the imperceptibility of noise to human observers. By addressing the following research questions, this study aims to illuminate the limitations of current detection systems and propose pathways for improvement:

## 2 RESEARCH QUESTIONS AND HYPOTHESES

### 2.1. Research Question 1 (RQ1)

**RQ1:** Can the addition of imperceptible noise/perturbation to logos allow these images to evade detection by anti-phishing tools?

**Hypothesis:** Adding noise to these images will prevent the model from correctly identifying the logos

### 2.2. Research Question 2 (RQ2)

**RQ2:** Can these noisy logos avoid detection by users

**Hypothesis:** Adding noise to these images will prevent the model from correctly identifying the logos

### 2.3. Reasoning and Objective

This study is significant for both technical and practical reasons. On the technical front, it identifies critical gaps in AI-based security measures, emphasizing the need for adversarial robustness in visual similarity detection. Practically, it underscores the importance of educating users about the visual cues of phishing and designing multi-layered defenses that extend beyond AI. As phishing attacks continue to evolve, understanding their mechanics and developing robust countermeasures remain essential to safeguarding digital spaces.

## 3 BACKGROUND

### 3.1 Related Work

Existing studies have demonstrated that adversarial samples can avoid detection effectiveness of image classification models. Ji et al. showed that visual perturbations can evade many models, while Hao et al. found that slight distortions significantly lower detection accuracy. This study builds upon these findings by systematically evaluating current tools against perturbed logos to identify potential improvements.

Code snippets from “Unlearnable Examples: Making Personal Data Unexploitable” [9] were utilized in the generation of noise to images and to implement the Projected Gradient Decent method of applying perturbation.

### 3.2 Image Perturbation

Image perturbation is a technique that subtly modifies an image by introducing adversarial noise. Unlike traditional image alterations, such as resizing or cropping, perturbation manipulates pixel values without altering the image's overall structure or content. This ensures that the modified image remains visually indistinguishable from its original version to human observers while appearing entirely different to

machine learning models.

Adversarial noise is generated using algorithms such as Projected Gradient Descent (PGD) and Fast Gradient Sign Method (FGSM). These algorithms iteratively apply small perturbations to an image, guided by a loss function that maximizes the likelihood of misclassification. The resulting adversarial examples expose vulnerabilities in AI systems, particularly those relying on convolutional neural networks (CNNs) for image classification.

In this study, perturbations were applied to logos from popular brands, including Facebook, Instagram, and YouTube. By manipulating epsilon values—a parameter controlling the magnitude of noise—we tested the resilience of anti-phishing tools and evaluated their ability to identify perturbed logos. The results reveal significant weaknesses in visual similarity detection models, underscoring the importance of robust adversarial training for AI systems.

### 3.3 Phishing

Phishing is a form of cybercrime where attackers impersonate legitimate entities to deceive users into revealing sensitive information. It is one of the most common and effective methods of online fraud, accounting for billions of dollars in losses annually. Phishing attacks exploit trust, familiarity, and urgency, employing techniques such as fraudulent emails, websites, and advertisements.

A particularly concerning subset of phishing is visual phishing, which leverages logos, brand names, and other visual elements to trick users into believing a fraudulent website or email is genuine. For example, a phishing website mimicking a bank's login page might use the bank's logo, color scheme, and layout to appear authentic.




Recent advancements in AI-based anti-phishing tools have improved the detection of visual phishing. These tools utilize CNNs and other machine learning models to analyze logos and branding elements, identifying discrepancies that signal malicious intent. However, adversarial attacks, such as perturbed logos, exploit the limitations of these models. By introducing imperceptible noise, attackers can bypass visual similarity detection, rendering these tools ineffective.

## 4 METHODS

### 4.1 Data Collection: Generating Noise

Perturbed logos are generated using PyTorch, focusing on common brands like Facebook, Instagram, and YouTube. Across the 3 brands, 6 logos are generated per brand, of epsilon values ranging from 0.005 to 0.05. The size of the steps remains at a constant ratio of epsilon/2, while the number of steps is 120 across all tests.

Noise is generated using Projected Gradient Decent (PGD) on a custom renet18 model. This is an image classification model that has been previously trained on a dataset of 278 popular logos. This custom model represents possible Convolutional Neural Network (CNN) models that are used by phishing tools to identify stolen logos on possible phishing websites.

Epsilon:	0.005	0.01	0.02
Step Size:	0.0025	0.005	0.01
Num Steps:	120	120	120
Logo:			
Guess/Confidence	"Facebook" (99.62%)	"Facebook" (81.53%)	"Delta Airlines" (99.88%)
Tricked the Model?	NO	NO	YES







Epsilon:	0.03	0.04	0.05
Step Size:	0.015	0.02	0.025
Num Steps:	120	120	120
Logo:			
Guess/Confidence	"Delta Airlines" (98.93%)	"Delta Airlines" (95.41%)	"Delta Airlines" (98.45%)
Tricked the Model?	YES	YES	YES

Table 1- Noise Generated Facebook Logos

After the noisy images are generated, the images are then sent as inputs to the original trained model that identifies logos- where the model will output its guess as to what the logo is and its confidence in this guess.

As seen in Table 1, the model correctly identifies the logo as "Facebook" at epsilon values of 0.005 and 0.01 at a confidence rate of 99.62% and 81.53%, respectively. However, beyond that point (at epsilon  $\geq 0.02$ ), it is no longer able to detect the logo as "Facebook" and misidentifies it as being Delta Airline's logo with high confidence rates ( $\geq 95\%$ ).

Epsilon:	0.005	0.01	0.02
Step Size:	0.0025	0.005	0.01
Num Steps:	120	120	120
Logo:			
Guess/Confidence	"Instagram" (95.86%)	"Instagram" (62.33%)	"Square" (98.45%)
Tricked the Model?	NO	NO	YES







Epsilon:	0.03	0.04	0.05
Step Size:	0.015	0.02	0.025
Num Steps:	120	120	120
Logo:			
Guess/Confidence	"Square" (98.45%)	"Square" (99.46%)	"Square" (99.68%)
Tricked the Model?	YES	YES	YES

Table 2- Noise Generated Instagram Logos

Instagram and YouTube's logos had similar results (Tables 2 and 3)- will all three of them avoiding detection at epsilon  $\geq 0.02$  (with YouTube being even more successful at avoiding detection. As it is unable to be identified by the image classifier at an epsilon of 0.005). The epsilon value of 0.02 is significant, as this represents the point in which all 3 logos were able to avoid detection by the model- indicating that this level of noise would most likely avoid detection by anti-phishing tools that utilize CNNs in image classification.

Epsilon:	0.005	0.01	0.02
Step Size:	0.0025	0.005	0.01
Num Steps:	120	120	120
Logo:			
Guess/ Confidence	"Square" (73.82%)	"Square" (93.35%)	"Square" (98.45%)
Tricked the Model?	YES	YES	YES




Epsilon:	0.03	0.04	0.05
Step Size:	0.015	0.02	0.025
Num Steps:	120	120	120
Logo:			
Guess/ Confidence	"Square" (99.54%)	"Square" (99.70%)	"Square" (99.67%)
Tricked the Model?	YES	YES	YES

Table 3- Noise Generated YouTube Logos

#### 4.2 Data Collection: Noise Imperceptibility

To test the imperceptibility of the noise added to the samples, a form was created and sent out to 46 participants. They were asked 22 questions on whether logos appeared changed or unchanged. An example of the original, clean logos was given at the beginning of the survey (for those unfamiliar with the brands' logos).

The questions contained a mixture of control samples (clean and very obviously modified images) and noisy logos with epsilon values ranging from 0.01 to 0.05.



Figure 1- Clean Logos

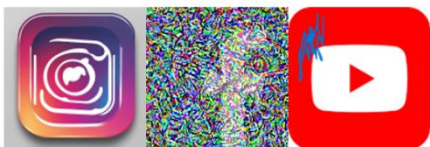


Figure 2- Control Group (Modified Logos)

The addition of these control groups was to ensure that all samples were valid. Only two of the 46 survey responses needed to be removed due to an issue regarding Facebook's logo. Both participants marked the clean Facebook logo (and all consecutive "noisy" Facebook logos) as being "incorrect". Upon reading their response to this decision, they stated the logo differed from the current app logo. (Even though an example of the clean Facebook logo was provided at the beginning, they still identified all the Facebook logos- including clean ones- as incorrect). The Facebook logo used in our test is different from the current logo used on mobile devices, and this led to 2 invalid data points in regard to Facebook.

Lastly, participants were asked for various demographic information. These questions were as follows:

- Name?
- Age?
- Do you wear glasses/contacts?
- Are you colorblind?
- How often do you use Facebook/Instagram/YouTube?
- Do you have experience in Computer Science and/or AI?
- Did you complete this form on a Phone, Tablet, or Computer?
- For Logos you didn't trust, what were the reasons?

This survey was sent via text message, discord, email, and in-person to various friends, family, classmates, and strangers. The survey, however, was not sent to anyone within the class this project takes place in- to avoid bias by those who are familiar with the goal and intention of the project.

Originally, this research intended to observe the perceptibility of noisy logos on those who are colorblind. However, due to the limitations of our sample size, only 2 participants surveyed were colorblind. For this reason, analysis of this demographic could not be concluded.

### 4.3 Data Analysis

Quantitative analysis involves comparing detection accuracy for clean and perturbed logos, using statistical methods to determine significance. Qualitative analysis identifies specific logos that bypass detection, exploring reasons for failure.

## 5 PRELIMINARY RESULTS

### 5.1 Fast Gradient Sign Method

Originally, the Fast Gradient Sign Method (FGSM) technique of generating noise was used to create adversarial examples. However, this method required a significantly high epsilon to avoid detection by the model (epsilon = 0.3-0.4). This noise was very noticeable and could easily be identified as a modified logo. This shows the importance of fine-tuning adversarial perturbations as crucial to achieving imperceptibility while evading detection.

### 5.2 Lesson Learned from Pilot Study

Some methods of generating noise are far more effective at achieving both imperceptibility and avoiding detection by the model. Originally, the Fast Gradient Sign Method (FGSM) technique of generating noise was used to create adversarial examples. However, this method required a significant high epsilon to avoid detection by the model (epsilon = 0.3+). This high level of noise was incredibly noticeable and could easily be identified as being a modified logo. This shows the importance of fine-tuning adversarial perturbations as crucial to achieving imperceptibility while evading detection.

## 6 RESULTS

### 6.1 Image Detection by Image Classifier

The study evaluated the ability of AI-based anti-phishing tools to detect logos perturbed with adversarial noise. Detection performance varied significantly based on the epsilon value applied during perturbation. Clean logos consistently achieved high detection confidence across all tools, but the introduction of noise drastically reduced classifier accuracy:

#### Facebook Logos

- At epsilon = 0.005, the classifier confidently identified the logo with a 99.62% accuracy.
- As epsilon increased to 0.01, confidence dropped to 81.53%.
- Beyond epsilon = 0.02, the classifier misidentified the Facebook logo as belonging to other brands, such as Delta Airlines, with confidence rates exceeding 95%.

#### Instagram Logos:

- At epsilon = 0.005, detection accuracy remained above 98%
- At epsilon = 0.02, detection confidence fell below 20%.
- Similar to Facebook, Instagram logos failed detection entirely at epsilon  $\geq$  0.03.

#### YouTube Logos:

- Surprisingly, YouTube logos were highly susceptible to perturbations, failing detection even at epsilon = 0.005
- At epsilon = 0.02, the classifier could no longer identify the logo, rendering it ineffective for detection

The results underscore a threshold effect at epsilon = 0.02, where all logos consistently evaded detection. This suggests that attackers can exploit this threshold to bypass anti-phishing tools without significantly compromising visual quality.

### 6.2 Imperceptibility of Noise

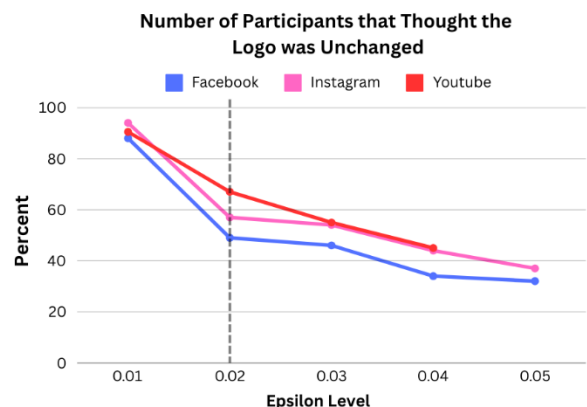


Figure 3: Comparison between YouTube, Facebook, and Instagram in Imperceptibility.



The human perception study revealed the extent to which noise added to logos was imperceptible to participants. Key findings include:

Detection Rates:

- At  $\epsilon = 0.02$ , only 42% of participants identified changes in Facebook logos.
- For Instagram logos, 54% failed to notice any alterations.
- YouTube logos showed the highest imperceptibility, with 67% of participants unable to detect changes.

Demographic Trends:

- Older participants and those with vision impairments were significantly less likely to detect noise.
- Participants with experience in computer science or AI were more adept at identifying perturbations, particularly at lower  $\epsilon$  values.

Participant Feedback:

Several participants noted difficulty distinguishing noisy logos from their original versions, citing familiarity with branding as a key factor in their decisions.

Even some of the participants who were able to identify the added noise noted the difficulty in observing the perturbation. They expressed concern that the real-world use of these noisy logos would probably go unnoticed. It required a significant amount of focus to notice the modified logos.

Some participants expressed skepticism about certain logos even when clean, highlighting potential limitations in user trust.

These results highlight the dual challenge faced by anti-phishing systems: adversarial noise undermines AI-based detection while also exploiting human cognitive biases and perceptual limitations.

## 7 DISCUSSION

### 7.1 Age on Image Perception

The findings of this study reveal a significant impact of age on the ability to detect perturbed logos. Older participants consistently demonstrated lower detection rates, particularly for logos

with  $\epsilon$  values of 0.02 and above. Several factors likely contribute to this trend. Age-related declines in visual acuity, such as presbyopia and reduced sensitivity to contrast, may impair the ability to discern subtle alterations in logos. Additionally, older individuals might be less familiar with specific logos, especially those associated with brands like Instagram or YouTube, which tend to be more popular among younger demographics. This lack of familiarity could hinder their ability to identify small discrepancies between clean and noisy logos. Cognitive load also plays a role, as older participants may find it more challenging to analyze complex visual stimuli while distinguishing between similar images.

Does this Instagram Logo Look Correct? ( $\epsilon = 0.02$ , All)



Figure 4: Number of Participants who thought the 0.02 Epsilon Instagram logo was unchanged. 56.5% of all participants stated the logo was correct

Does this Instagram Logo Look Correct? ( $\epsilon = 0.02$ , Age > 45)

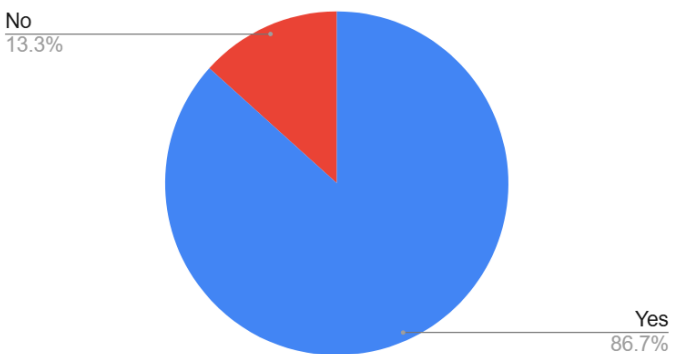


Figure 5: Number of Participants over the age of 45 who thought the 0.02 Epsilon Instagram logo was unchanged. 86.7% of those over 45 years old stated the logo was correct

These findings have important implications for phishing defense strategies. Attackers could exploit this demographic vulnerability, targeting older individuals with perturbed logos that are more likely to bypass human scrutiny. Educating older users about the visual tactics employed in phishing attacks is essential. Awareness campaigns should emphasize the importance of scrutinizing logos and branding elements, particularly in contexts where sensitive information is requested. Tailored approaches, such as simplified training modules and visual aids, could help mitigate the risks faced by this vulnerable demographic.

## 7.2 Computer Science Experience on Image Perception

Participants with a background in computer science or artificial intelligence exhibited an advantage in detecting perturbed logos. This group was consistently more adept at identifying changes across all epsilon levels, with particularly high detection rates at lower levels of perturbation. Their familiarity with concepts like adversarial noise and image manipulation likely heightened their sensitivity to subtle inconsistencies in logo design.

Additionally, participants with technical expertise approached the task with a more critical mindset, questioning the authenticity of logos even when perturbations were minimal. This skepticism, combined with their training in pattern recognition and visual systems, enhanced their ability to detect manipulations that might go unnoticed by less experienced individuals.

In addition, participants with experience in AI were more likely to outright state that the logos had “image perturbation added” when asked why they didn’t trust the logos. Given that image perturbation is part of the curriculum here at the University of Tennessee, this aids in these individuals’ abilities to identify the added noise.

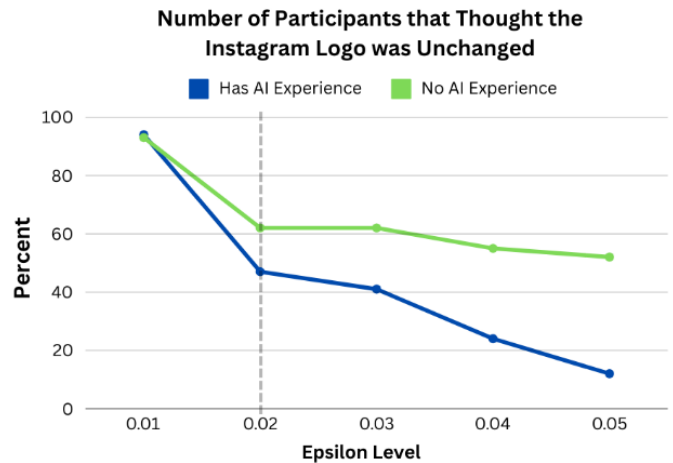


Figure 6- Comparison of participants with AI experience vs. without AI experience who did not notice the added noise in the Instagram logos.

The results underscore the value of technical education in combating phishing. By equipping users with the knowledge and tools to critically analyze visual elements, organizations can reduce the success rates of adversarial attacks. Training programs targeting non-technical users could incorporate basic principles of image analysis and adversarial techniques, fostering greater awareness and vigilance. Such initiatives would not only strengthen individual defenses but also complement broader organizational efforts to mitigate phishing risks.

## 8 IMPLICATIONS OF WORK

### 8.1 Impact on AI Phishing Checker Tools

This study highlights critical vulnerabilities in current AI-based anti-phishing tools, particularly in their susceptibility to adversarial noise. Tools that rely on convolutional neural networks (CNNs) for visual similarity detection struggled to recognize perturbed logos once noise levels reached an epsilon value of 0.02 or higher. This consistent failure indicates a need for adversarial training techniques during model development. Adversarial training involves exposing AI models to perturbed logos during training to improve their ability to generalize and detect manipulated images effectively. In addition to improving robustness through adversarial training, integrating

multi-modal detection strategies can significantly enhance these tools. For instance, combining visual analysis with contextual and behavioral data, such as URL legitimacy or user interaction patterns, could provide more comprehensive protection against phishing attempts. These strategies are essential to overcoming the limitations of current AI systems and addressing the rapidly evolving threat landscape.

8.2 Implications for Human Factors

The human perception findings from this study reveal significant vulnerabilities that phishing attackers can exploit. Older participants, in particular, demonstrated reduced ability to detect perturbed logos. Age-related factors such as diminished visual acuity, slower cognitive processing, and limited familiarity with modern branding likely contributed to this trend. These vulnerabilities underscore the need for tailored educational initiatives to mitigate phishing risks for different demographic groups. Older users may benefit from simplified training materials that focus on recognizing key visual discrepancies in logos and understanding phishing warning signs.

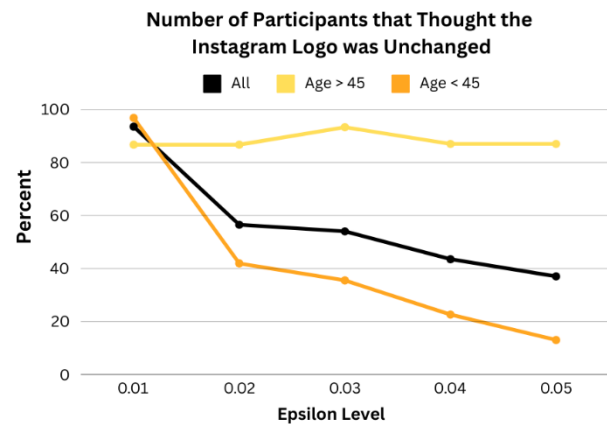


Figure 7: Percentage of Participants who did not notice the added noise, by age range, across the Instagram Logos. Those over 45 years old consistently believed the images were unchanged (at a rate around ~87%), even as epsilon increased.

Conversely, interactive and gamified training modules may better engage younger users and the general public. For technically inclined individuals, advanced training on the mechanics of adversarial noise and its implications for

cybersecurity could improve their detection rates and ability to inform others. Such human-centered approaches are vital for complementing technological defenses and creating a multi-layered defense strategy against phishing.

8.3 Broader Organizational Impacts

Organizations must adopt a holistic approach to combating phishing that combines technological solutions with employee education and awareness. AI-based anti-phishing tools, while essential, cannot be relied upon as standalone defenses due to their limitations in detecting perturbed logos. To address these gaps, organizations should implement layered security measures that include browser extensions capable of analyzing user interactions with logos and flagging suspicious behavior. Contextual analysis tools that assess website elements and behavioral patterns can also provide additional safeguards. Employee training is another critical component. Tailored workshops and phishing simulations can prepare employees to recognize and respond to phishing attempts effectively. IT staff, in particular, require specialized training on adversarial techniques to enhance their ability to identify sophisticated attacks. By integrating these strategies, organizations can foster a culture of cybersecurity awareness and resilience, reducing their overall exposure to phishing threats.

8.4 Future Threat Study & Recommendations

The future of phishing defense faces significant challenges as adversarial techniques continue to evolve. Generative adversarial networks (GANs) and other advanced perturbation methods have the potential to create logos that evade detection by both AI systems and human users. These techniques could produce perturbations that are visually indistinguishable from clean logos, making detection nearly impossible without specialized tools. Additionally, attackers may combine adversarial noise with social engineering tactics, exploiting psychological triggers such as fear, urgency, or authority to prompt users into action. These hybrid attacks will require new approaches to



defense, including enhanced AI systems capable of adversarial robust detection and ongoing user education to address the human factors involved. Collaborative efforts between researchers, industry professionals, and government agencies will be crucial in developing innovative solutions to counteract these threats and protect digital ecosystems. To address the vulnerabilities identified in this study, several actionable recommendations emerge. Developers of anti-phishing tools should focus on implementing adversarial training to improve the robustness of AI models. Regular updates to detection algorithms are necessary to ensure that these systems keep pace with evolving adversarial techniques. Integrating multi-modal detection strategies, such as combining visual analysis with contextual and behavioral cues, can further enhance their effectiveness. For organizations, regular phishing simulations and tailored employee training programs are essential to building resilience against phishing attacks. These training programs should emphasize visual literacy, teaching employees to critically evaluate branding elements and recognize phishing warning signs. For researchers, future studies should explore the potential of advanced adversarial techniques, such as GAN-based perturbations, and investigate demographic-specific vulnerabilities to inform targeted defense strategies. By adopting these measures, the cybersecurity community can strengthen its defenses against the growing threat of phishing.

## 9 CONCLUSION

This study provides critical insights into the vulnerabilities of current anti-phishing tools and highlights the dual challenge posed by adversarial attacks on both AI systems and human users. By introducing subtle noise to logos, attackers can effectively bypass visual similarity detection models, with detection rates dropping sharply at an epsilon threshold of 0.02. Human perception studies further reveal demographic-specific vulnerabilities, with older users and those without technical expertise being significantly less likely to identify perturbed logos. These findings underscore the

limitations of existing defenses and the pressing need for more robust solutions.

Addressing these vulnerabilities requires a multi-faceted approach. AI-based detection systems must incorporate adversarial training to enhance their resilience to manipulated images. Multi-modal detection strategies that integrate visual, contextual, and behavioral data can provide a more comprehensive defense against sophisticated phishing attempts. On the human side, education and training are critical to raising awareness and equipping users with the skills needed to recognize phishing attacks. Tailored programs, particularly for at-risk demographics such as older users, can significantly mitigate phishing risks.

The implications extend beyond individual users and tools, highlighting broader organizational and societal challenges. Organizations must adopt layered security measures, combining technological innovations with human oversight and robust training programs. Furthermore, as adversarial techniques evolve, collaborative research and proactive investments in cybersecurity will be essential to staying ahead of emerging threats.

In conclusion, this study emphasizes the importance of an integrated defense strategy that combines advancements in AI technology with human-centered approaches. By addressing the gaps identified in this research, the cybersecurity community can take meaningful steps toward creating safer digital environments and reducing the impact of phishing attacks. Continued research, education, and innovation are essential to ensuring resilience against the ever-evolving tactics of adversarial threats.

## 10 REFERENCES

- [1] C. Kane, "The most common visual elements exploited in phishing attacks," VISUA, Jun. 22, 2021. Available: <https://visua.com/graphical-phishing-attacks>. [Accessed: Nov. 10, 2024].
- [2] P. Panda, A. K. Mishra, and D. Puthal, "A Novel Logo Identification Technique for Logo-Based Phishing Detection in Cyber-Physical Systems," *Future Internet*, vol. 14, no. 8, p. 241, Aug. 2022. doi: <https://doi.org/10.3390/fi14080241>.

- [3] K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, "Utilisation of website logo for phishing detection," *Computers & Security*, vol. 54, pp. 16–26, Oct. 2015. doi: <https://doi.org/10.1016/j.cose.2015.07.006>.
- [4] CoreTech, "In a new twist on phishing campaigns, cybercriminals are luring victims to click on images rather than downloading malicious files or clicking suspicious links. Image-based phishing attacks, also known as 'image phishing' or 'visual phishing,' are cyber attacks that use images or graphics to deceive," *Linkedin.com*, Jul. 26, 2023. Available: <https://www.linkedin.com/pulse/look-dont-touch-hackers-sending-targets-image-based-phishing/>. [Accessed: Nov. 10, 2024].
- [5] "Phishing with Images — Office of Innovative Technologies," *Utk.edu*, 2024. Available: <https://oit.utk.edu/security/learning-library/article-archive/phishing-with-images/>. [Accessed: Nov. 10, 2024].
- [6] A. S. Bozkir and M. Aydos, "LogoSENSE: A Companion HOG based Logo Detection Scheme for Phishing Web Page and E-mail Brand Recognition," *Computers & Security*, p. 101855, May 2020. doi: <https://doi.org/10.1016/j.cose.2020.101855>.
- [7] C. Kane, "How is Logo Detection Used in Anti-Phishing," *VISUA*, Apr. 19, 2022. Available: <https://visua.com/logo-detection-in-anti-phishing>. [Accessed: Nov. 11, 2024].
- [8] "Overview of phishing techniques: Brand impersonation — Infosec," *Infosecinstitute.com*, 2020. Available: <https://www.infosecinstitute.com/resources/phishing/overview-of-phishing-techniques-brand-impersonation/>. [Accessed: Nov. 11, 2024].
- [9] H. Huang, X. Ma, S. M. Erfani, J. Bailey, and Y. Wang, "Unlearnable Examples: Making Personal Data Unexploitable," *arXiv:2101.04898 [cs, stat]*, Feb. 2021. Available: <https://arxiv.org/abs/2101.04898>.
- [10] lindsey98, "GitHub - lindsey98/Phishpedia: Official Implementation of 'Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages' USenix'21," *GitHub*, 2021. Available: <https://github.com/lindsey98/Phishpedia>. [Accessed: Nov. 11, 2024].
- [11] lindsey98, "GitHub - lindsey98/PhishIntention: Phishing detection through webpage intention," *GitHub*, 2021. Available: <https://github.com/lindsey98/PhishIntention>.
- [12] Fujiaoji, "GitHub - Fujiaoji/LogoClassification," *GitHub*, 2024. Available: <https://github.com/Fujiaoji/LogoClassification/tree/main>. [Accessed: Nov. 11, 2024].
- [13] F. Ji et al., "Evaluating the Effectiveness and Robustness of Visual Similarity-based Phishing Detection Models," *arXiv.org*, 2024. <https://arxiv.org/abs/2405.19598> (accessed Dec. 01, 2024).
- [14] A. S. Bozkir and M. Aydos, "LogoSENSE: A Companion HOG based Logo Detection Scheme for Phishing Web Page and E-mail Brand Recognition," *Computers & Security*, p. 101855, May 2020, doi: <https://doi.org/10.1016/j.cose.2020.101855>.
- [15] J. Lee, P. Lim, B. Hooi, and D. D. Mon, "Multimodal Large Language Models for Phishing Webpage Detection and Identification," *arXiv.org*, 2024. <https://arxiv.org/abs/2408.05941> (accessed Nov. 28, 2024).
- [16] H. Wang and B. Hooi, "Automated Phishing Detection Using URLs and Webpages," *arXiv.org*, 2024. <https://arxiv.org/abs/2408.01667>.
- [17] W. Yao, Y. Ding, and X. Li, "Deep Learning for Phishing Detection," 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/Sustain-Com), Dec. 2018, doi: <https://doi.org/10.1109/bdcloud.2018.00099>.
- [18] K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, "Utilisation of website logo for phishing detection," *Computers & Security*, vol. 54, pp. 16–26, Oct. 2015, doi: <https://doi.org/10.1016/j.cose.2015.07.006>.
- [19] C. C. L. Tan, K. L. Chiew, K. S. C. Yong, Y. Sebastian, J. C. M. Than, and W. K. Tiong, "Hybrid phishing detection using joint visual and textual identity," *Expert Systems with Applications*, vol.

220, p. 119723, Jun. 2023, doi: <https://doi.org/10.1016/j.eswa.2023.119723>.

[20] C. Kane, "How is Logo Detection Used in Anti-Phishing," VISUA, Apr. 19, 2022. <https://visua.com/logo-detection-in-anti-phishing>

[21] A. Gendre, "Spear Phishing Impacts from the HR Perspective," Vadesecure.com, Jan. 2016, doi: <https://doi.org/109028204803/1729501169013>.

[22] H. Thakur and S. K. R. Cell, "Logo Image Based Approach for Phishing Detection," Online, 2016. Accessed: Dec. 01, 2024. [Online]. Available: <https://ijoes.vidyapublications.com/paper/Vol21/14-Vol21.pdf>

[23] J. Lee, Z. Xin, S. Melanie, K. Sabharwal, G. Apruzzese, and D. D. Mon, "Attacking logo-based phishing website detectors with adversarial perturbations," arXiv.org, 2023. <https://arxiv.org/abs/2308.09392> (accessed Dec. 01, 2024).

[24] B. Team, "Bypassing Object Detection: The Rise of Logo Manipulation in Phishing Attacks - BUFFERZONE," BUFFERZONE, Apr. 18, 2024. <https://bufferzonesecurity.com/bypassing-object-detection-the-rise-of-logo-manipulation-in-phishing-attacks/>

[25] J. Lee, Z. Xin, S. Melanie, K. Sabharwal, G. Apruzzese, and D. D. Mon, "Attacking logo-based phishing website detectors with adversarial perturbations," arXiv (Cornell University), Jan. 2023, doi: <https://doi.org/10.48550/arxiv.2308.09392>.

## 11 APPENDIX

### 11.1 Survey Questionnaire

Description: This is a collection of 3 popular logos: Facebook, Instagram, and YouTube. Some logos remain unchanged, while others have been edited or are incorrect. Please state whether these logos appeared unchanged or not.

Below, there are 3 examples of what the unchanged logos look like.

#### 1. Facebook (Example)



#### 2. Instagram (Example)



#### 3. YouTube (Example)



#### Research Questions:

##### 1. Does this FACEBOOK logo look correct?



☐ Yes

☐ No

##### 2. Does this YOUTUBE logo look correct?



☐ Yes

☐ No

...

##### 21. Does this INSTAGRAM logo look correct?



☐ Yes☐ No

22. Does this FACEBOOK logo look correct?

☐ Yes☐ No

Demographic Questions:

1. Name? \_\_\_\_\_

2. Age? \_\_\_\_\_

3. Do you Wear Glasses?

☐ Yes, For objects at a distance☐ Yes, for objects in close range☐ Yes, for both far-range and close-range☐ No, I do not wear any glasses

4. Do you have Colorblindness?

☐ Yes☐ No☐ Prefer not to say

5. How Often do you use Facebook?

☐ Daily/Weekl☐ Monthly☐ Yearly☐ Never Used

6. How Often do you use YouTube?

☐ Daily/Weekly☐ Monthly☐ Yearly☐ Never Used

7. How Often do you use Instagram?

☐ Daily/Weekly☐ Monthly☐ Yearly☐ Never Used

8. Do you have experience in Computer Science and/or AI?

☐ Yes, in Computer Science only☐ Yes, in AI only☐ Yes, in both☐ No, Neither







9. Did you Complete this form on a Computer, Tablet, or Phone?







☐ Computer/Laptop☐ Phone☐ Tablet







10. For the logos you didn't trust, what were some of the reasons? \_\_\_\_\_

The Full Questionnaire can be found at the following link: [Logo Research Report](#)

11.2 Graphs and Data

Epsilon: Step Size: Num Steps:	0.005 0.0025 120	0.01 0.005 120	0.02 0.01 120	0.03 0.015 120	0.04 0.02 120	0.05 0.025 120
Logo:						
Logo Guess/ Confidence:	"Square" (73.82%)	"Square" (93.35%)	"Square" (98.52%)	"Square" (99.54%)	"Square" (99.70%)	"Square" (99.67%)
Tricked the model?	YES	YES	YES	YES	YES	YES
Is the noise Perceptible?	NO	NO	NO	Probably not	Probably not	Probably not

Epsilon: Step Size: Num Steps:	0.005 0.0025 120	0.01 0.005 120	0.02 0.01 120	0.03 0.015 120	0.04 0.02 120	0.05 0.025 120
Logo:						
Guess/ Confidence:	"Facebook" (99.62%)	"Facebook" (81.53%)	"Delta Airlines" (99.88%)	"Delta Airlines" (98.93%)	"Delta Airlines" (95.41%)	"Delta Airlines" (98.45%)
Tricked the model?	NO	NO	YES	YES	YES	YES
Is the noise Perceptible?	NO	NO	Probably not	Probably Not	Maybe	Probably

Epsilon: Step Size: Num Steps:	0.005 0.0025 120	0.01 0.005 120	0.02 0.01 120	0.03 0.015 120	0.04 0.02 120	0.05 0.025 120
Logo:						
Guess/ Confidence:	"Instagram" (95.86%)	"Instagram" (62.33%)	"Square" (98.45%)	"Square" (99.46%)	"Square" (99.68%)	"Square" (99.75%)
Tricked the model?	NO	NO	YES	YES	YES	YES
Is the noise Perceptible?	NO	NO	Probably Not	Probably Not	Maybe	Probably

