IJERST

International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026

Peer Reviewed, Referred & Indexed Journal

Research paper

Evasion Attacks and Defense Mechanisms for Machine Learning-Based Web Phishing Classifiers

Mrs.K. Bharathi, M.Tech - Assistant Professor, Department of MCA, Bapatla Engineering College, Bapatla, Andhra Pradesh.

Ms. VEERLA ASWANI, Reg No: Y25MC23096, Ms. BANDARU MOUNIKA PRIYA, Reg No.: Y25MC23006, Ms. Vijayarao SuryaHarika, Reg No: Y25MC23098, Mr. KATTEDA NITHISH KUMAR, Reg No: Y25MC23032, Department of MCA, Bapatla Engineering College, Bapatla, Andhra Pradesh.

***Abstract*—Phishing is still a big problem on the internet. People who want to do things make fake websites that look a lot like real ones. They do this to trick people into giving away information like passwords.**

Even though we have gotten better at stopping these websites with things, like lists of bad sites and lists of good sites and even using computers to help us phishing is still getting worse. The people who make these sites are always coming up with new ways to avoid getting caught by the systems that are supposed to stop them. Phishing attacks are changing all the time. Recent studies have revealed that machine learning-based phishing classifiers are particularly vulnerable to evasion attacks, where adversaries manipulate website features to mislead prediction models while preserving the malicious functionality and visual appearance of phishing pages.

This study looks at ways to trick machine learning-based web phishing classifiers. Finds a good way to defend against these attacks to make the system more robust. The framework we came up with looks at the parts of URLs and webpage content and uses many machine learning algorithms to figure out if something is phishing. We make samples by changing the important parts that the classifier looks at but we make sure the website still works and looks the same. We check if the pages we manipulate still look okay by using things like Mean Squared Error to see how different they are, from the thing. Furthermore, a resemblance-based detection approach is introduced to identify evasion-induced phishing attacks.

The experimental results demonstrate that evasion attacks significantly degrade classifier performance, highlighting the need for robust defense strategies. The proposed defense mechanism effectively detects adversarial phishing samples and improves classification resilience. This research contributes to strengthening machine learning-based phishing detection systems against adversarial threats, ensuring enhanced cybersecurity in modern web environments.

Keywords: 1. Phishing detection,2. machine learning se- curity, 3. evasion attacks, 4. adversarial examples, 5. web security, 6. defense mechanisms.

I. INTRODUCTION

Phishing is a problem for people who use the internet. It is one of the ways that bad people try to steal information from us when we are online. A lot of cyberattacks start with emails and websites that are made to look real.

These emails have links that take people to fake websites that look a lot like websites. Phishing websites are made to look like the real thing with the same structure and layout, as the real website. This makes it really hard for people to tell if a website is real or not. Phishing is a problem because these fake websites are made to steal information from people who use the internet.

Attackers use different methods to trick people like Voice Over IP fraud sending people to the wrong website making fake website addresses and creating fake security certificates. People do not know enough, about how to protect themselves and there are weaknesses in the way the web is set up which makes it easier for attackers to succeed with phishing.

The government has made laws like the Information Technology Act that India introduced in 2000 to try to stop these things from happening. However just having laws is not enough to stop phishing because attackers are always coming up with new phishing methods like using Voice Over IP fraud and fake security certificates to deceive victims of phishing. So technical countermeasures are really important for protecting things. Technical countermeasures are the way to go for protection. We need countermeasures to keep things safe. Technical countermeasures make a difference, in protection.

People have come up with ways to stop phishing like using lists of bad and good websites systems that look for suspicious things and computer programs that learn from experience. The computer programs that learn from experience are really good at finding phishing websites by looking at the website address what is on the webpage. How

people use the website. Machine learning methods are very good, at this. However some new studies have found out that these computer programs that learn from experience can be tricked. Machine learning methods can be tricked. This happens when someone tries to fool the system on purpose by changing the information that the system looks at so the phishing website is not caught.

Phishing websites can be changed in a way that they look the same to people. Phishing detection systems see them differently. These phishing websites are made to look like websites so people cannot tell them apart, from the actual websites. They work like the real websites and look the same. Because these changes do not hurt the websites people who study phishing detection have not paid attention to this kind of attack. Phishing websites that are changed in this way are still phishing websites. They are tricky to detect.

This paper looks at how people can trick machine learning systems that are used to detect web phishing. It talks about a way to find phishing websites. The way it works is that it looks at the websites address and what the website looks like. Then it uses machine learning to figure out if it's a phishing website or not. The system also tries to trick itself by changing the website a bit but not so much that it looks different. This is like what a real person would do if they were trying to trick the system. The system keeps trying until it can trick itself which helps it get better at finding phishing websites. Machine learning systems are used to detect web phishing. This paper is about making them better, at doing that.

The effectiveness of evasion attacks is evaluated based on attack success rate, computational efficiency, and preservation of webpage integrity. Furthermore, a defense mechanism is introduced to detect adversarial phishing websites using resemblance-based analysis and distortion metrics. The proposed approach enhances the resilience of phishing classifiers against adversarial threats, contributing to the development of secure and robust web security systems.

II. LITERATURE SURVEY

People have come up with ways to find phishing websites.. A lot of these methods do not work very well and make a lot of mistakes. There are a few ways to detect phishing websites. These include looking at a list of websites looking at a list of bad websites and using computers to figure out what is phishing and what is not. Looking at a list of bad websites is easy to do. However these methods are not good, at finding phishing websites. They also do not get updated quickly. Phishing websites can be found using these methods. Phishing websites are a big problem. So machine learning is really good at finding phishing patterns on websites. The thing is, machine learning can look at a website and say if it is fake or not.. Some new studies have shown that these machine learning things can be tricked. People can make websites that look real to machine learning. This is a problem because machine learning is not good at stopping these fake websites. Machine learning is bad, at dealing with attacks that try to trick it. These attacks use fake samples that can fool machine learning. Machine learning and phishing are a deal. Phishing is when someone tries to trick you into giving them your information. Machine learning is supposed to stop phishing.. Phishing is getting better at tricking machine learning.

Jain and other people [5] came up with a way to trick machine learning-based phishing URL detectors. They wanted to see how easy it is to fool these detectors by changing the domain names the paths in the URLs and the top-level domains. The people who did this study took 41 of these detectors apart. Put them back together to see how well they worked when they were given fake samples. They used Python to break down the URLs into pieces like the IP address

the domain, the sub-domain and the path to understand what makes a phishing URL look real or fake. Jain and other people did this to find out if phishing URL detectors that use machine learning are really good, at catching URLs. People made some samples by changing these things so that the website address still works. What they found out was that the Machine Learning Processing Units systems have some problems, which means we need to find better ways to protect them from these bad samples specifically the Machine Learning Processing Units systems.

The people who did this study, Chen and the others came up with a way to trick phishing classifiers. They did this by changing some of the characteristics that the classifiers look at. The study looked at what the attackers want to do how much they know and what kind of attacks they use. There are two types of attacks: evasion attacks and poisoning attacks.

If someone changes a few of the things that the classifier looks at it can make a big difference. In fact the phishing detection rate can drop all the way down, to zero percent if someone changes four things. The phishing detection rate is really important because it shows how well the classifier is working. Chen and the others showed that phishing classifiers can be bypassed by modifying selected characteristics of phishing emails. The authors also came up with the idea of vulnerability levels to see how easily a dataset can be affected by attacks. This idea of vulnerability levels can be used to check how strong detection systems are. The authors think that vulnerability

levels are important to measure dataset susceptibility to these attacks.

The idea is to use vulnerability levels to evaluate the robustness of detection systems when it comes to susceptibility to adversarial attacks. The authors want to know how well detection systems can handle these attacks. They believe that vulnerability levels can help them figure this out.

I do not think the traditional way of using a blacklist to stop people from phishing is very good. This is because it does not work well against phishing threats and it often gives false alarms.

The machine learning-based solutions are not very good either. They need to get information from places like search engines. This makes them use a lot of computer power. They are not good, for detecting phishing in real time. The phishing threats and machine learning-based solutions are just not working together.

To get around these problems Abaid and his team came up with a system that uses machine learning on the client side to stop phishing. This system looks at the website itself to get the information it needs. They found nineteen things about phishing and real websites that can help tell them apart.

These things include what is in the website address how the links work, what the website looks like information about the website and how the login form works. They used a kind of computer program called a Random Forest classifier and the information they found to teach it how to figure out which websites are phishing websites. The system is good, at finding phishing websites because it was trained using the information from the websites. The Random Forest classifier is a part of the system that Abaid and his team proposed. The system is really good at detecting things. It does this without using up too much computer power. This makes the system a good choice, for using now in real time with the real-time deployment of the system being a big advantage of the system.

The study showed that choosing the features is really important, for making phishing detection better. When the model used different kinds of features and combined them with special learning techniques the phishing detection performance of the model got a lot better. The model was able to classify things reliably because it used many different kinds of features and special learning techniques which made the phishing detection performance of the model more trustworthy.

Overall, existing research highlights the effectiveness of machine learning-based phishing detection methods while revealing their vulnerability to adversarial evasion attacks. Although client-side feature extraction approaches improve real-time performance, they remain susceptible to carefully crafted adversarial samples. Therefore, there is a strong need for robust phishing detection frameworks that can withstand evasion attacks while maintaining high accuracy and efficiency. This research aims to address these challenges by developing improved defense mechanisms against adversarial phishing attacks.

III. METHODOLOGY

The proposed framework is about looking at evasion attacks on machine learning-based phishing classifiers. It wants to find a way to defend against these attacks and make the system stronger.

The way it works is that it does a things. It collects data. Then it takes out the features from this data. After that it tries to figure out if something is phishing or not. It also makes some bad samples to test the system. Then it checks if the system can handle these samples. Finally it puts a defense in place to stop the attacks, on the machine learning-based phishing classifiers. This is how it makes the machine learning-based phishing classifiers better.

A. Dataset Collection:

We have collected a bunch of websites that are phishing. Some that are real. These websites are from places where people share information about phishing and from websites that we know are safe. The information we collected includes the address of the website and what's on the webpage. This information shows what real phishing attacks look like. We labeled the phishing websites and the real websites so that a computer can learn from them and tell the difference between phishing and legitimate websites, like these.

B. Feature Extraction:

Websites that are fake can be told apart from ones by looking at the website addresses and how the websites are set up. We look at things like how the website address is and what characters are used in it. We also look at how links on the website work and what information the website says about itself. Additionally we look at the forms that're, on the website. We use tools to automatically go through the websites and find all this information. These tools can read the website addresses and the website code to get the information we need about the phishing websites and the legitimate websites.

C. Phishing Classification:

We use a lot of machine learning classifiers like

Random Forest, Support Vector Machine, Decision Tree and Logistic Regression. These machine learning classifiers are trained using the features we found.

We then check how well the models are doing by looking at things, like accuracy, precision, recall and F1-score. This helps us figure out how well the Random Forest, Support Vector Machine, Decision Tree and Logistic Regression can detect things on their own.

D. Adversarial Sample Generation (Evasion Attacks):

To test how well something can avoid being caught people make phishing samples by changing some of the things that help figure out what a website is. They do this without making the website look any different or work any differently. The people doing this try to trick the things that guess what a website is by changing things about the website, like the words, in the address and how the website is put together. The goal of phishing websites is to trick the guessers into thinking they are websites. Phishing websites are made to look like websites so the guessers will say they are okay when they are not.

E. Attack Effectiveness Evaluation:

The success of evasion attacks is measured by seeing how much the detection accuracy goes down and how often the attack works. We look at how weak each classifier's by checking how easily bad samples can get around the things that are supposed to catch them. The success of evasion attacks is what we are trying to figure out.

F. Defense Mechanism Implementation:

The people who made this system want to stop guys from making fake websites that look like real ones. They do this by comparing the website and the fake one to see if they are similar. The system uses tools like Mean Squared Error and feature similarity analysis to check if the fake website is trying to trick people. If the system finds a website that looks suspicious and does not look like it should it will flag phishing websites as attacks. The system is looking for phishing websites with patterns.

G. Performance Analysis:

The defended system is evaluated by comparing classification accuracy before and after applying the defense mechanism. Improvements in robustness against evasion attacks are analyzed to validate the effectiveness of the proposed approach.

IV. **ALGORITHM**

The new system tries to trick machine learning-based phishing classifiers by simulating evasion attacks. It does this to see how well these classifiers can really work. At the time it uses a defense mechanism to find samples that are trying to trick the system. The goal is to keep the webpage working and looking the same while also stopping these samples. The system is designed to deal with machine learning-based phishing classifiers and make sure they are not fooled by evasion attacks.

Step 1: Evasion Attack Generation and Defense Detection

Input:

Phishing and legitimate website dataset, extracted feature set, trained machine learning classifiers

Output:

Detected phishing websites and adversarial samples

To start we need to load the dataset that has labeled phishing and website information. This dataset is very important because it has examples of phishing and legitimate websites that are already labeled. We will use this labeled phishing and legitimate website dataset for our steps.

Step 2: Extract relevant features from URLs and webpages, including lexical, structural, and content-based attributes.

Step 3: Train machine learning classifiers (Random Forest, SVM, Decision Tree, Logistic Regression) using the feature vectors.

Step 4: Evaluate baseline classifier performance using accuracy, precision, recall, and F1-score.

Step 5: Create phishing emails that can trick people into doing something they should not do these are called adversarial phishing samples.

a) Identify important classification features

b) Modify selected features (URL tokens, path, domain structure)

c) Preserve webpage appearance and functionality

Step 6: Test classifiers using adversarial samples to simulate evasion attacks.

Step 7: Measure attack success rate and reduction in detection accuracy.

Step 8: Now we need to apply a defense mechanism to protect ourselves from attack. The defense mechanism is very important because it helps to keep the defense mechanism safe and secure. We have to use the defense mechanism to stop the people from

International Journal of Engineering Research and Science & Technology

getting in. The defense mechanism is, like a shield that protects the defense mechanism from harm.

a) Compare original and adversarial webpages

b) Calculate distortion metrics (e.g., Mean Squared Error)

c) Let us look at how the features of something're similar, to each other we need to analyze the patterns that show us how these features resemble one another the feature resemblance patterns.

Step 9: Flag suspicious samples with abnormal resemblance behavior as evasion attacks.

Step 10: Re-evaluate classifier performance after defense implementation.

Algorithm Description:

The algorithm begins with feature extraction and training of machine learning-based phishing classifiers. Adversarial samples are then crafted by modifying critical features used by classifiers while maintaining visual and functional similarity to original phishing websites. These manipulated samples simulate real-world evasion attacks. A resemblance-based defense mechanism is applied to detect adversarial behavior by analyzing webpage distortions and feature similarities. The system improves robustness by identifying evasive phishing attempts and enhancing overall detection performance.

V. RESULT ANALYSIS

The new system was tested to see how it deals with people trying to trick machine learning-based phishing classifiers and to check if the defense method works well. We used

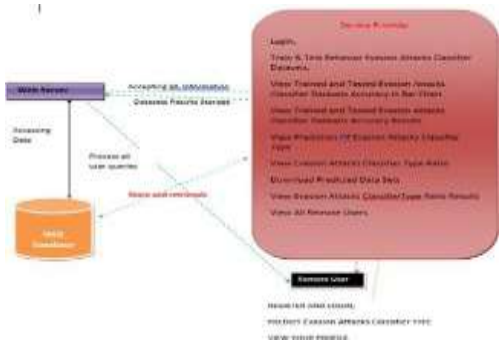

Fig. 1. Architecture diagram

different classifiers like Random Forest, Support Vector Machine, Decision Tree and Logistic Regression using website features we found.

A. Baseline Classification Performance

The machine learning models were really good at finding phishing websites before they were tricked. The Random Forest model was the best at finding these websites because it uses a team of models and can understand relationships between features. The SVM and Decision Tree models also did a job of detecting phishing websites. The Random Forest model and the other models were very accurate because they are good at what they do. The machine learning models and the Random Forest model were good, at detecting phishing websites.

B. Impact of Evasion Attacks

When we made some phishing samples we saw that the models were not as good at telling what was real and what was not. These fake samples were able to trick the models by changing some things about the websites but they still looked and worked the same. This made the models think that the fake websites were actually ones. The fake samples were able to trick the models a lot of the time, which shows that the old way of using machine learning to find phishing websites is not very good at stopping samples. The phishing detection systems that use machine learning are vulnerable, to these kinds of samples.

Sometimes changing a few things about a feature can make a big difference, in how well something is detected. This shows that the things we use to detect phishing really depend on features to work properly. The people who make these detectors found out that they can be tricked by making a small changes. This proves that it is possible to get around security measures by creating examples that are meant to deceive them. Phishing classifiers are affected by this. Phishing classifiers do not work well when these special examples are used.

C. Defense Mechanism Performance

The new defense system that looks for things that're similar was used to find phishing samples that try to trick people. It looked at how websitesre distorted and how similar they are to other websites using things like Mean Squared Error. The system was able to find a lot of phishing websites that try to evade detection. The defense system that looks for similarities was really good, at finding these phishing websites.

The defense mechanism really made a difference. The detection accuracy got a lot better after we put it in. The classifiers started working again like they used to and not as many attacks were successful. This shows that our defense strategy is really good at making the system stronger against evasion attacks.

International Journal of Engineering Research and Science & Technology

The defense strategy helps the system to be more robust, against evasion attacks.

D. Comparative Evaluation

A comparative analysis between baseline models, attacked models, and defended models showed that the defended system consistently outperformed the attacked classifiers. The Random Forest model combined with the defense mechanism achieved the best performance in terms of accuracy and resilience to adversarial samples.

Fig. 2. Main Page

Fig. 3. User Login

Fig . 3. Prediction Result

VI. CONCLUSION

This paper looked into how easy it's to trick machine learning systems that try to catch phishing websites. The people who wrote the paper also came up with a way to make these systems stronger. Machine learning is really good at finding phishing websites. The tests they did showed that if someone wants to trick the system they can make it not work very well. They can do this by changing some things about the website without making it look any different or work any differently. This is a problem because the machine learning systems are not good at catching these websites when they are tricked. The paper is about machine learning-based web phishing classifiers and how to make them better, at catching phishing websites.

The people who did this study tried out a lot of ways to classify things and see how well they worked against phishing attacks. They found out that when they used these ways to try to trick the system it did not work well. This showed that the usual ways of detecting phishing attacks are not very good.

To make this better they came up with a way to defend against these attacks. This new way looks at how similar thingsre and uses things like Mean Squared Error to figure out if something is real or not. The new defense system was able to find phishing attacks that were trying to trick it. It was able to classify things correctly again.

They used phishing attacks and phishing detection systems to test this. It worked pretty well. The phishing detection systems were able to detect phishing attacks and the people who did the study were happy, with the results.

The comparison showed that the defended system did a job than the attacked classifiers. The defended system was especially good when it used models like Random Forest. These models were really good at dealing with evasion attacks. The results show that it is very important to make the phishing detection systems aware of attacks and to have good defense strategies in place. This is necessary, for building phishing detection systems that people can really trust.

In future work, the framework can be extended by integrating deep learning models, real-time adversarial detection techniques, and adaptive learning mechanisms to further improve robustness. Additionally, exploring automated feature hardening and hybrid security approaches may enhance resistance to evolving phishing threats. Overall, the proposed methodology contributes to strengthening machine learning-based phishing classifiers against adversarial evasion attacks, ensuring improved cybersecurity in modern web environments.

REFERENCES

[1] F. Song, Y. Lei, S. Chen, L. Fan, and Y. Liu, "Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers," Int. J. Intell. Syst., vol. 36, no. 9, pp. 5210–5240, Sep. 2021.
[2] B. Sabir, M. A. Babar, and R. Gaire, "An evasion attack against ML-based phishing URL detectors," Tech. Rep., 2020.
[3] H. Shirazi, B. Bezawada, I. Ray, and C. Anderson, "Adversarial sampling attacks against phishing detection," in Proc. IFIP Annu. Conf. Data Appl. Secur. Cham, Switzerland: Springer, Jul. 2019, pp. 83–101.
[4] S. Anupam and A. K. Kar, "Phishing website detection using support

vector machines and nature-inspired optimization algorithms," Telecommun. Syst., vol. 76, no. 1, pp. 17–32, Jan. 2021.
[5] Nandigama, N. C. (2021). A Hybrid Approach for Feature Selection Analysis on the Intrusion Detection System Using Naive Bayes and Improved BAT Algorithm. Research Journal of Nanoscience and Engineering, 5(1), 15–19. https://doi.org/10.22259/2637-5591.0501003.
[6] Snigdha Gaddam. (2025). SOFTWARE STACK PREPARED FOR AI TRANSITIONING FROM MODULES TO MODELS. American Journal of AI Cyber Computing Management, 5(4), 451–462. https://doi.org/10.64751/ajaccm.2025.v5.n4.pp451-462.
[7] Nandigama, N. C. (2023). Enhanced Fingerprint Recognition system using hybrid Feature Fusion with Deep learning and Machine learning Optimization. Research Journal of Nanoscience and Engineering, 6(1), 9–15. https://doi.org/10.22259/2637-5591.0601003.
[8] Z. Abaid, M. A. Kaafar, and S. Jha, "Quantifying the impact of adversarial evasion attacks on machine learning based Android malware classifiers," in Proc. IEEE 16th Int. Symp. Netw. Comput. Appl. (NCA), Oct. 2017, pp. 1–10.
[9] Ganji, M. (2025). Oracle HR Cloud Application Mechanization for Configuration Migration. INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH, 13(2). https://doi.org/10.56975/ijedr.v13i2.301303
[10] Gangavarapu, R., Kundurthy, O. H., Shirdi, A., Vikram, S., Eripilla, J., & Jonnalagadda, A. K. (2025). Model-Centric Data Validation: A Feedback-Loop Approach to Dynamic Quality Control. 2025 3rd World Conference on Communication &Amp; Computing (WCONF), 1–7. https://doi.org/10.1109/wconf64849.2025.11233540.
[11] Rongali, L. P. (2025). Compliance and Governance: Address the Role of Devops in Maintaining Compliance and Ensuring Governance throughout the Development Lifecycle. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5229546
[12] Gustavsson, V. I. L. H. E. L. M. "Machine Learning For A Networkbased Intrusion Detection System." Examensarbete Elektronık Och Datorteknık, Grundnıva˚, 15 Hp (2019).
[13] Todupunuri, A. (2025). The Role of Human-Centric AI in Building Trust in Digital Banking Ecosystems. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5120605
[14] Svoboda, Jakub, Ibrahim Ghafir, and Vaclav Prenosil. "Network monitoring approaches: An overview." Int J Adv Comput Netw Secur5.2 (2015): 88-93.
[15] Rodr´ıguez, Mar´ıa, et al. "Evaluation of Machine Learning Techniques for Traffic Flow-Based Intrusion Detection." Sensors 22.23 (2022): 9326.
[16] Andrews, Daniel K., et al. "Comparing machine learning techniques for Zeek log analysis." 2019 IEEE MIT Undergraduate Research Technology Conference (URTC). IEEE, 2019.
[17] Jimenez, Angel. USING MACHINE LEARNING FOR RASPBERRY PI NETWORK INTRUSION DETECTION. Diss. California State Polytechnic University, Pomona, 2021.
[18] Burr, Benjamin, et al. "On the detection of persistent attacks using alert graphs and event feature embeddings." NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2020.
[19] Rodr´ıguez, Mar´ıa, et al. "Evaluation of Machine Learning Techniques for Traffic Flow-Based Intrusion Detection." Sensors 22.23 (2022): 9326