IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14612

DETECTION OF PHISHING WEBSITE USING MACHINE LEARNING AND FEATURES EXTRACTION

T. Amalraj Victoire [1], Anusha D [2]

Professor, Department of Master Computer Applications,

Sri Manakula Vinayagar Engineering College, Pondicherry-605 107[1]

PG Student, Department of Master Computer Applications,

Sri Manakula Vinayagar Engineering College, Pondicherry-605 107[2]

Abstract: Phishing attacks represent a rapidly expanding threat in cyberspace that causes significant financial losses to web users and companies on an annual basis. Any sensitive data acquired from the customers by using different social engineering tactics is considered unauthorized, and any kind of website, pop-ups, instant message, emails, and other communication tools can be employed for recognizing phishing. In this paper, we propose a mechanism that could help recognize phishing or real URLs. The dataset consists of clean, spam, malicious, phishing, and defacement websites. Also, phishing URLs obtained from an open-source website known as "Phish Tank," which provides phishing URLs in different formats such as JSON, CSV, and others, have been included. Six models for recognizing phishing URLs based on the machine learning and deep neural network algorithms have been tested. With a set of about 10,000 random URLs, including up to 23,328 phishing URLs and 4894 valid URLs, divided into training and testing datasets the main aim of our study consists in creating software applications for detecting phishing URLs online. The dataset of Uniform Resource Locator has been tested and trained through feature selections like HTTPS and JavaScript-based features, domain-based features, address bar-based features to distinguish between genuine and phishing URLs. This study has offered an approach towards the classification of URLs into legitimate and phishing URLs.

Keywords: Phishing, Detection, Machine Learning, Neural Network, Authentication, Identification.

INTRODUCTION

We currently depend on digital channels for the collection and distribution of information; especially, social media is extensively used for these purposes.

Therefore, one of the most common types of attacks associated with social engineering is social engineering attacks, which are a kind of attacks aimed at stealing users' information; such information include credit card details and login credentials. An attack takes place if the attacker deceives his victim into opening an instant text message or chat, or email, which appears to be from someone he trusts. When clicking on the link, he will think he has been gifted and accidentally click on a harmful link, leading to malware installation, ransomware attack, or even exposure of his personal details.

The concerns regarding cyber security have greatly increased during the last few decades due to the rapid adoption of new technological developments. Individuals must know how the hackers work and how not to become their target. Phishing techniques represent one of the major threats today.

Fraudsters' techniques become more sophisticated with technological advancement. In addition to phishing, there are other means to steal sensitive information from the customers.

Phishing Method

Phishing websites are imitations of existing websites and generally possess form fields (such as textboxes). The intruder gets the information of the target when he/she uploads it. An intruder can steal credentials of one person from the other in the following way:

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14612

i. Building a phishing website: At first, the intruder recognizes the target as a legitimate company. After that, to get complete information about the company, the intruder visits the website of the firm. Using this information, the intruder develops the phishing website.

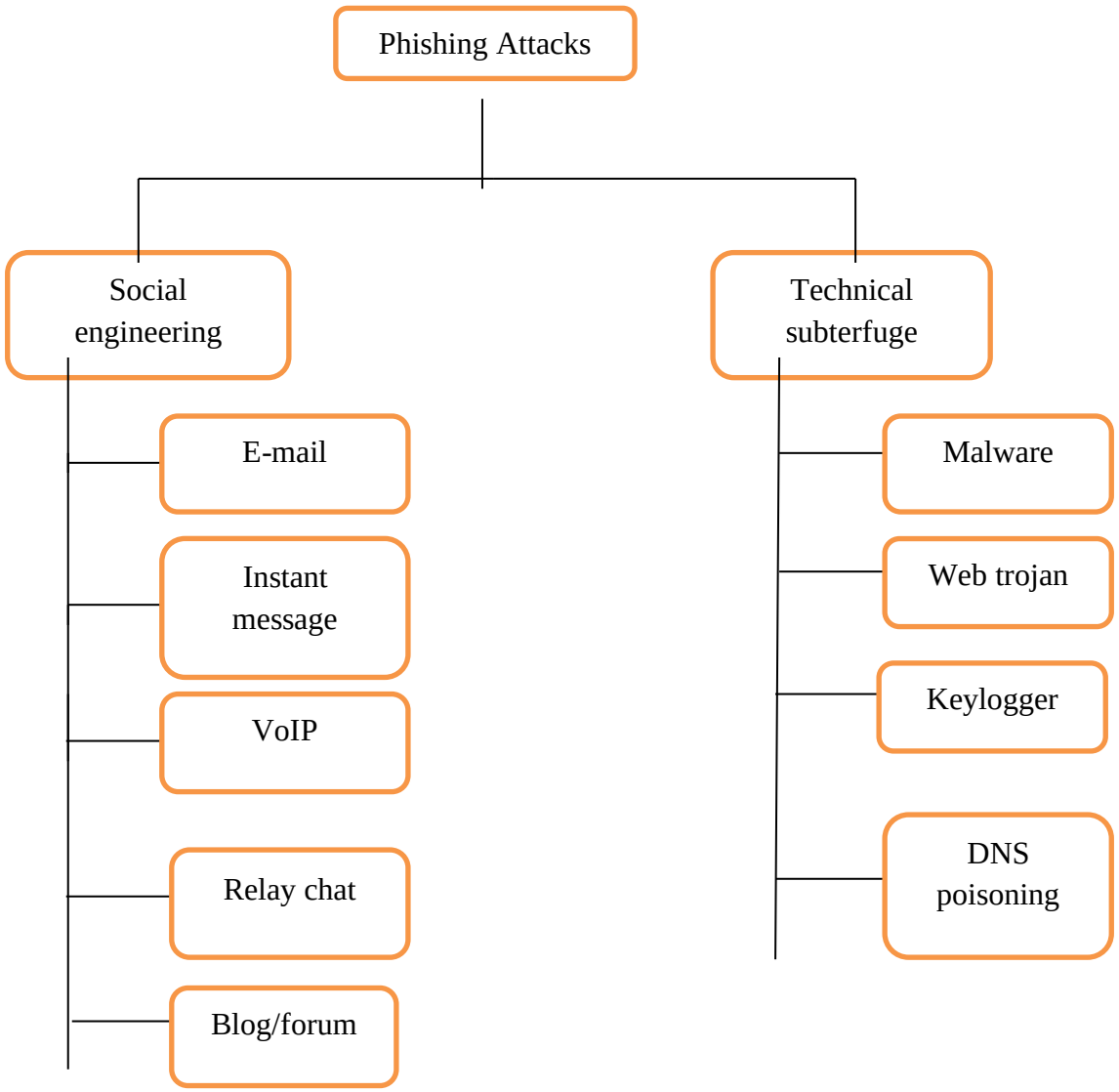
ii. Delivering links: The hacker makes a fake mail and delivers it to several people. The hacker gave the URL of the fraudulent website in the fake email. A spear phishing attack is conducted targeting a certain group of people with the email. In addition, hackers can share the URL of the fraudulent sites through forum postings, blog postings, and social media networking sites.

iii. Stealing credentials: A fake web page loads in the computer once the recipient clicks on the URL link. An unsuspecting guest's information is stolen using a fake registration form on the fraudulent site. The information entered by the recipient is also known to the hacker.

iv. Identity Theft: The details are misused by the criminal. For example, a fraudster might employ an individual's credit card details to complete a purchase.

Phishing Attack Taxonomy

The attacker will use social engineering and electronic trickery methods to conduct a phishing attack. This attack is executed through the forwarding of an impersonation email. These attacks are frequently carried out using requests from the hacker to respond using credit card issuers, banks, online vendors, and many more.



IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14612

The harmful malware is then loaded onto the victim's machine after they open the link in the false email, and information is acquired and communicated back to the perpetrator. Fake links or malicious software was embedded into the false emails by the criminals. They might also have access to the computer system and obtain the information at their convenience.

LITERATURE REVIEW

Over 38% of the world's population, or 2.97 billion individuals, accessed the Internet in 2014, according to Internet World Stats 9. Hackers can deceive gullible individuals into becoming victims of phishing scams because the Internet is insecure. Phishing emails are employed online in order to scam financial institutions and individuals. A record number of phishing attacks was reported in 2012, when there was a 160% rise in attacks over the year 2011. The nearly 450,000 phishing attacks reported in 2013 amounted to over 5.9 billion USD in losses.

In 2013, the total number of attacks increased by 1% compared to 2012. In the first quarter of 2014, a total of 125,215 phishing attacks were reported, which was 10.7% more than that reported in the last quarter of 2013. More than 55% of the phishing websites utilize the name of the target website to deceive users, and 99.4% utilize port 80.

Phishing attacks are conducted through an attacker who implements social engineering techniques and cyber fraudulent practices. Hackers use the social engineering technique to conduct the attack through forwarding a false email. The criminal normally asks the victim to provide the names of banks, credit card firms, online shops, and other institutions. Malicious software gets installed into the machine through clicking on fake emails that transmit private information to the attackers. In the fake emails, the criminals have embedded malicious links or malware which convey to the attackers what keys were pressed by the user.

Proposed various techniques, as well as components used in phishing and techniques to identify phishing websites. One of the best things about the paper is that it has covered several methods to identify phishing attacks. It also gives a technique to identify fake websites.

The paper proposes an approach to grouping phishing attacks into categories. The approach consists of categorization of web pages and feature extraction from them. Thirty variables have been identified, and they were acquired from the UCI machine learning repository dataset once phishing feature extraction concepts were specified. Support Vector Machine (SVM), Naive Bayes (NB), and Extreme Learning Machine (ELM) were applied for classification of the data according to the above attributes.

METHODOLOGY

The proposed approach for detecting phishing attacks utilizes machine learning algorithms and deep neural networks. The complete framework comprises two main components: an online application and machine learning models including Auto Encoder Neural Network, Multilayer XGBooster, Random Forest, Support Vector Machine, and Decision Tree. These models have been chosen based on multiple comparisons made between various machine learning approaches. A feature of websites based on real and phish datasets is employed to train and test each algorithm. Thus, the most effective model is chosen and embedded into the online application that helps users identify whether the URL link is legitimate or not.

MODEL DEVELOPMENT

Both supervised and unsupervised learning techniques were employed during the development of the machine learning models for phishing identification techniques. The source for the data required to create the datasets for training the machine learning models comes from numerous open-source platforms. Datasets comprising both legitimate and phishing URLs are included in the data collection process. This website offers an hourly updated list of phishing URLs in various formats, including CSV and JSON. For training purposes, a total of more than 24,442 phishing URLs were randomly chosen from this database. Moreover, a dataset containing URLs that do not contain malicious, spam, phishing, or altered content can be acquired from the same free resources. No matter what variety exists, the authentic URL database will be taken into consideration for this study. For training purposes, over 5000 authentic URLs have been randomly selected from the above-mentioned database.

In order to determine the trends and knowledge contained in the data set, it went through exploratory data analysis, which entailed analyzing, investigating, and summarizing the data using visualization techniques. Some of the techniques used are heat maps, scatter plots, and pair plots. The use of phishing and legitimate data sets was used to

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14612

extract attributes related to website content, such as the address bar-based attribute (with eight attributes), domain-based attribute (with three attributes), and HTML & JavaScript-based attribute (with four attributes). In total, there were fifteen attributes extracted for the detection of phishing attacks.

Among them is the provision of information to machine learning algorithms so that they can detect and understand the positive aspects of the dataset. The classification problem is the issue that is being considered in this research and has been formulated through supervised learning. The dataset has been trained to identify phishing through these algorithms: Decision Tree, Random Forest, Support Vector Machines, XGBooster, Multilayer Perceptron, and Auto-Encoder Neural Network. This has been done through the dataset. A training set and testing set have been developed through the dataset. Half of the data is used for the training phase in order for machine learning algorithms to understand the data and the capacity of distinguishing between genuine and fraudulent URLs. Fifty percent of the training sample is used for assessment in order to ascertain the performance of the training datasets after half of the data set has been successfully trained.

SYSTEM ARCHITECTURE

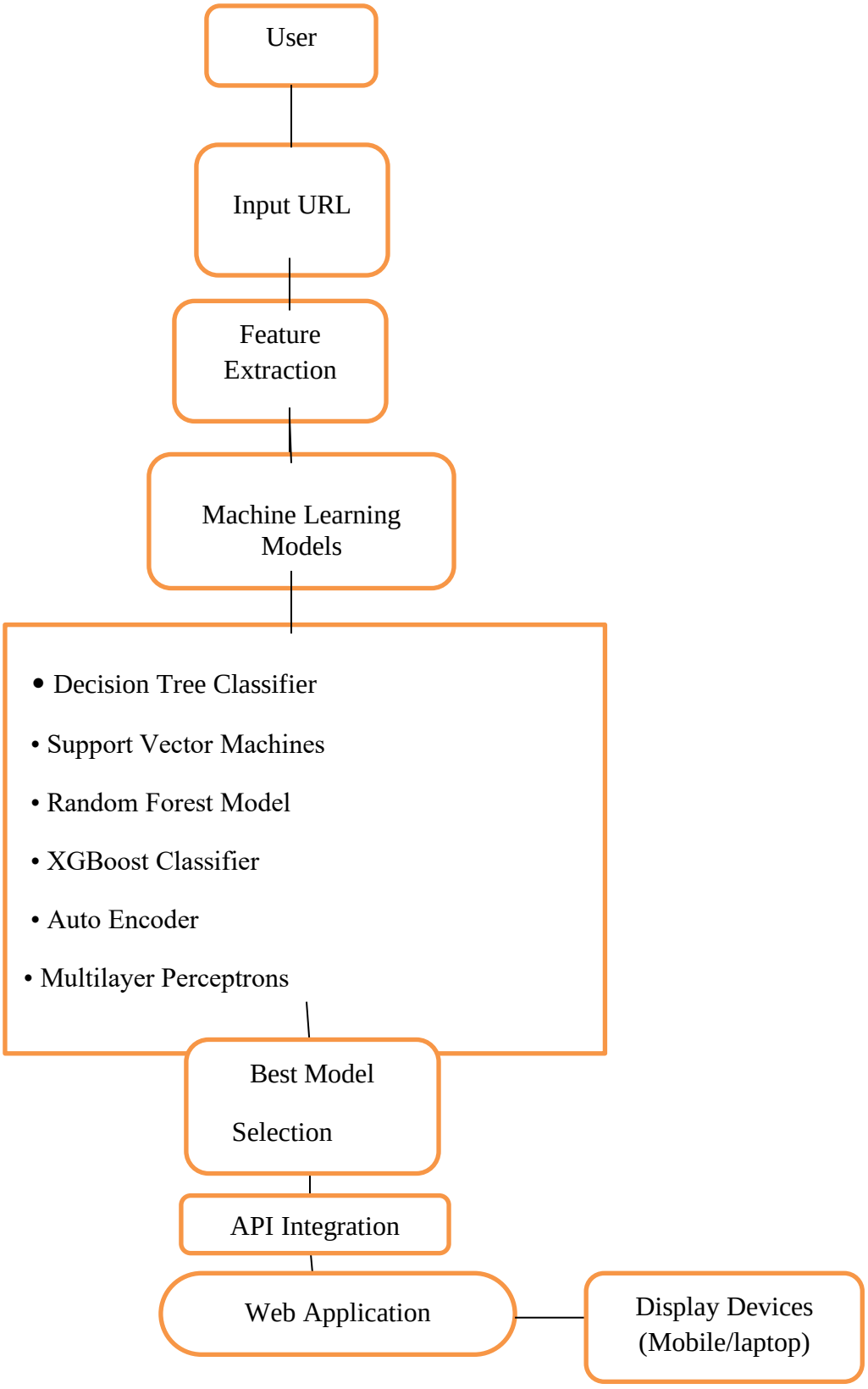
The proposed phishing detection technique architecture is illustrated. Prior to selecting the best algorithm with the highest accuracy, a URL entered by a person will be examined through several machine learning and deep neural networks. The best algorithm will then be included in the web application after being developed as an API (Application Programming Interface). Therefore, the person using the web application will use it through different display devices, such as computers, laptops, tablets, and smartphones.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14612



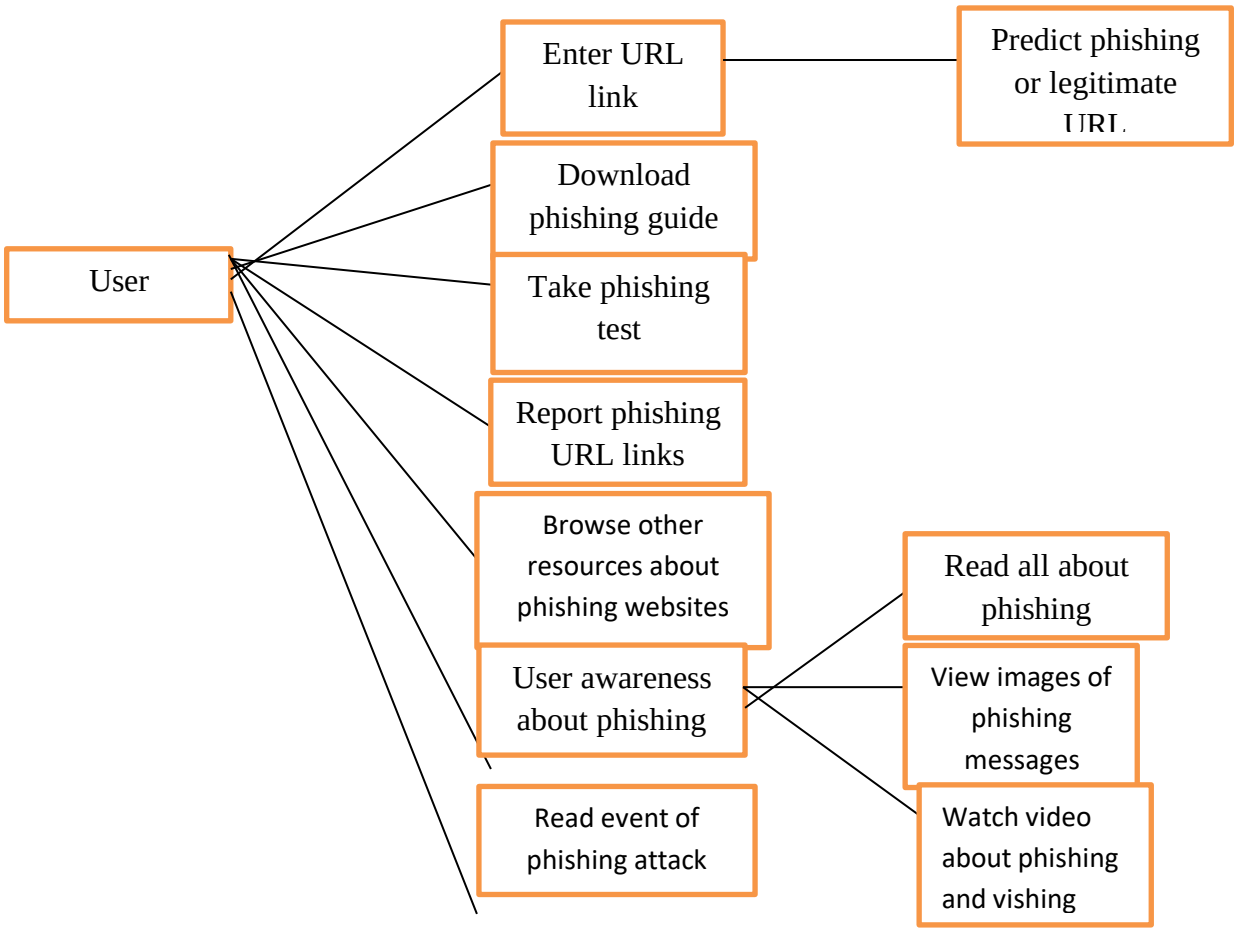
IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14612

USE-CASE DIAGRAM



Phishing prediction system

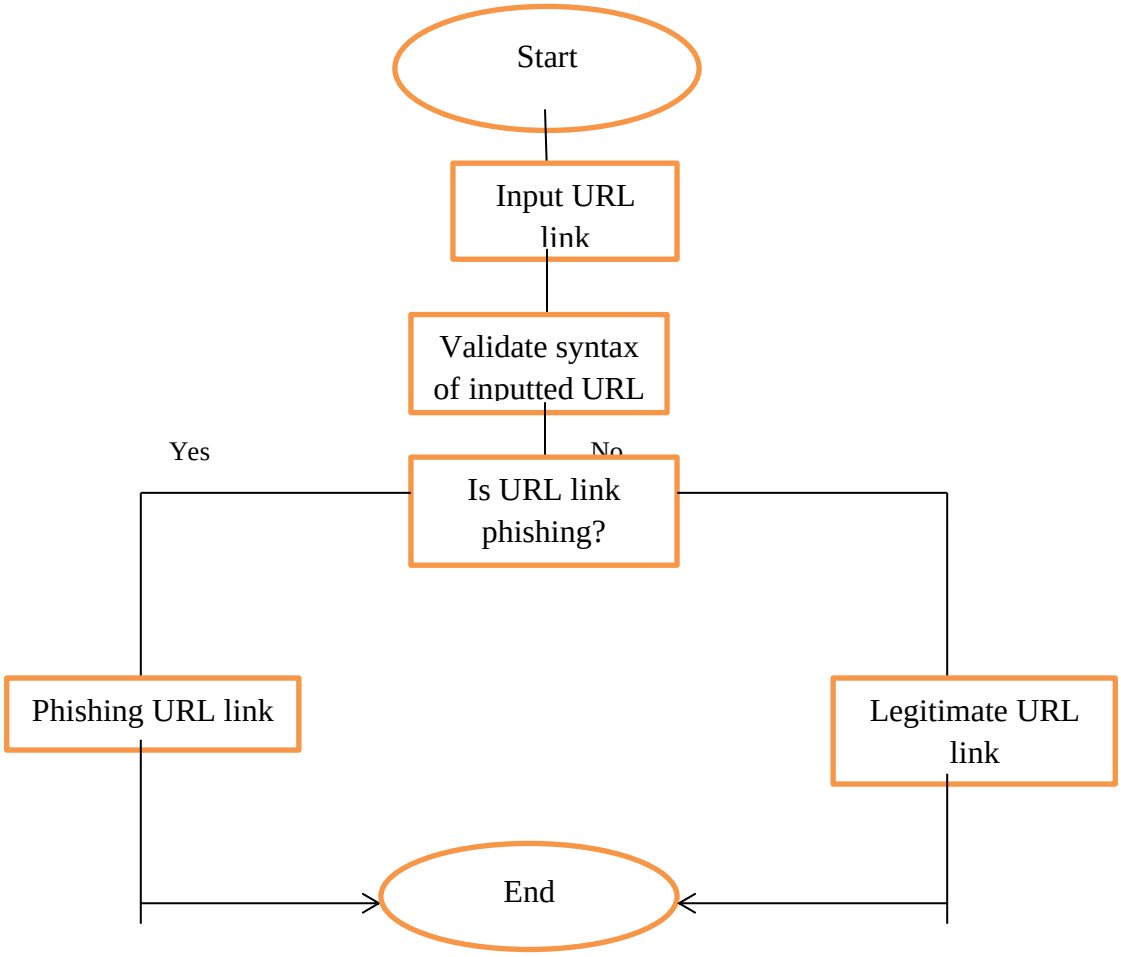
IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14612

PHISHING DETECTION WEB INTERFACE SYSTEM



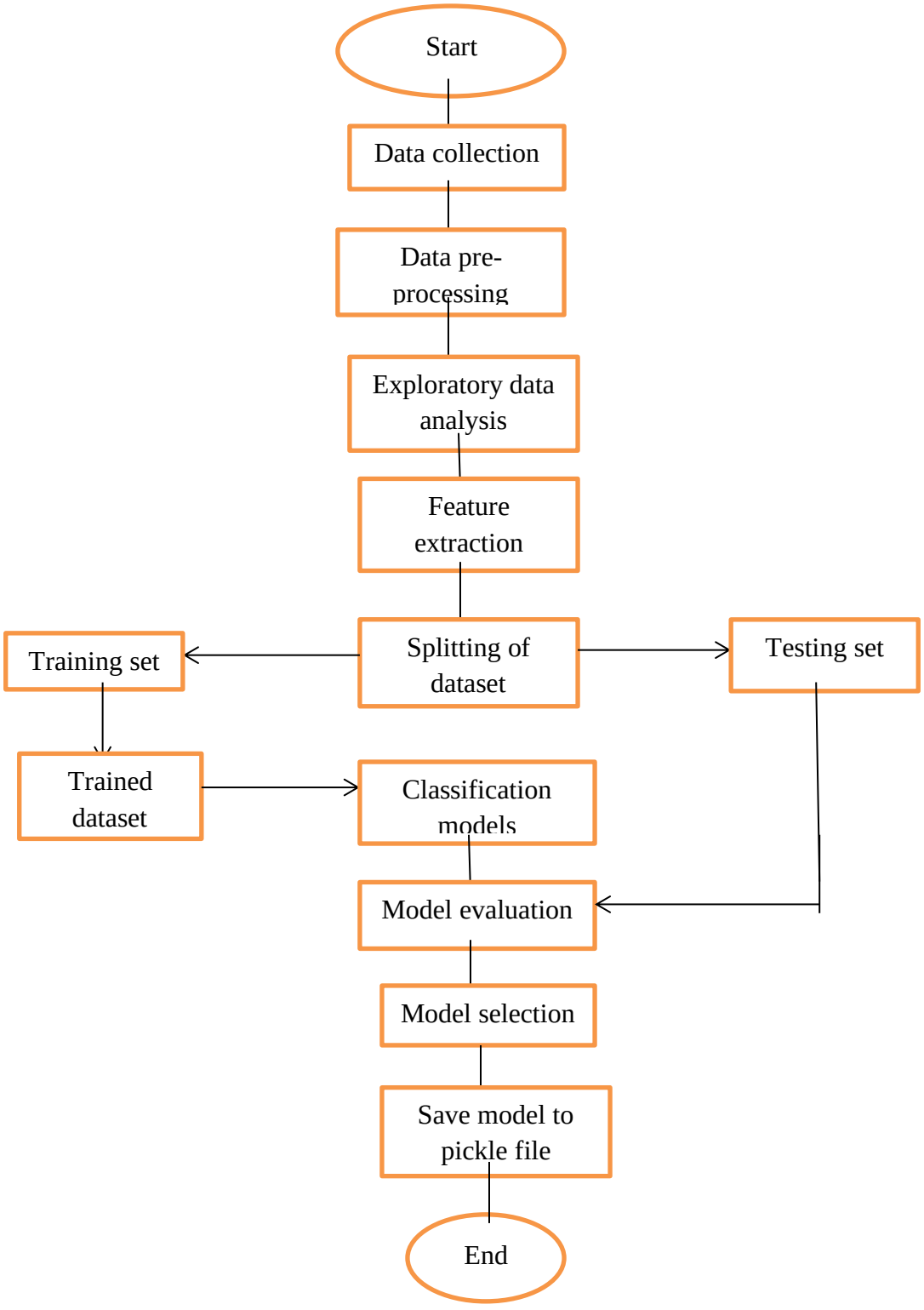
IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14612

MACHINE LEARNING TECHNIQUE FOR PHISHING DETECTION SYSTEMS



IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14612

PHISHING MITIGATION TECHNIQUES

Despite the fact that phishing attacks have become more advanced and harder to detect, there are still warning signs that can help you detect phishing attacks before they get out of hand. Here are some important signs used by security experts to detect fake links:

1. Check the URLs That Are Suspicious: Phishing URLs typically consist of some odd characters, are long, and may even seem to be confusing. Such types of URLs are employed by hackers in order to mask the real destination of the URL and trick their customers. Checking the URL becomes the first step towards defending oneself. Check whether the URL begins with "HTTPS," as the "s" indicates a secure connection through the SSL certificate. But do not rely only on SSL certificates, because nowadays, many cybercriminals distribute illegal content via HTTPS URLs that look legitimate. It would be best if you had doubts about URLs that seem like a bunch of symbols or appear too complicated.

2. Be Alert to Redirect Chains: As shown above, one of the foremost techniques used by fraudsters is redirects. Apart from being conscious about the complexity of the URL itself, observe where the link leads to. The technique makes it harder for users to determine the intent of fraud by making them confused and adding to the chain of distribution. One of the most common techniques used is sending an e-mail saying that some file needs to be downloaded. Instead of sending the file as an attachment, hackers redirect the link to ask for login credentials to access the file.

3. Watch Out for Strange Page Names and Missing Favicons: Checking the name and icon of websites is another way of detecting fraudulent URLs. A legitimate site must not contain any bizarre symbols or technical language, and its name must match the service that it provides. Any incompleteness or suspicion about the page name suggests that there might be some problem. The presence of a logo similar to the product or service on top of the name of the website is also essential for a legitimate site.

4. Avoid Being Scammed with Abused CAPTCHA and Cloudflare Verifications: One way that phishers will make use of CAPTCHA is through their abuse of "I am not a robot" verification, which is usually commonly used within phishing attacks. The phisher may make use of CAPTCHA by presenting fake, repeated tasks with CAPTCHA, despite its intention to identify users and prevent machine access. A similar scheme involves the abuse of websites like Cloudflare, where the phisher may obstruct delay to victims and hide the phishing scam using Cloudflare's verification process.

5. Verify Microsoft Domain Before Providing Passwords: Often, hackers create websites that look like trusted organizations like Microsoft to trick users into entering their credentials. Although Microsoft normally requires passwords only for some approved domains, it is recommended to be careful nonetheless. Keep in mind that your organization may even use the corporate domain to verify. Checking the website before providing the login credentials is therefore a prudent move.

6. **Analyze Connections to Recognizable Interface Elements:** You can also detect suspicious connections by closely analyzing the interface elements of programs. Keep in mind that the interface elements of programs that have an input box in order to enter a password on a browser web page is a highly suspicious sign. By mimicking interfaces of familiar software, such as those developed by Adobe or Microsoft, and using password input forms, phishers usually aim to gain the trust of clients. Individuals who are vulnerable relax and open themselves up because of this and end up falling prey to hacking tricks.

CONCLUSION

Phishing attacks are an increasing threat in cyberspace, causing the loss of billions of dollars for internet users each year. This type of cyber-attack uses various forms of social engineering to extract private information from consumers. Hence, any form of communication channel such as websites, popup messages, messaging apps, or electronic mails may be employed to detect the phishing activities. The various ways that have been applied by the investigators in an attempt to solve the issue of phishing attacks were analyzed in this study. The designed system applied diverse approaches of feature selection, deep neural network, and machine learning such as Decision Tree, Support Vector Machine, XGBooster, Multilayer Perceptions, Auto Encoder Neural Network, and Random Forest in determining patterns that made it easier to identify the URL links. Applying the feature extraction approach, the system was connected to the web application, where people could input website links and decide whether it is genuine or fake. Through the use of extracted features and algorithms applied to the data set, it was easier to detect malicious URLs,

DOI: 10.17148/IJIREEICE.2026.14612

hence improving the computational accuracy of the models. Additionally, it was extremely proficient in detecting whether the web address was genuine or not.

REFERENCES

[1]. Shekokar, N. M., Shah, C., Mahajan, M., & Rachh, S. (2015). An ideal approach for detection and prevention of phishing attacks. Procedia Computer Science, 49, 82-91.
[2]. Pujara, P., & Chaudhari, M. B. (2018). Phishing Website Detection using Machine Learning: A Review.
[3]. Lakshmi, V. S., & Vijaya, M. S. (2012). Efficient prediction of phishing websites using supervised learning algorithms. Procedia Engineering, 30, 798-805.
[4]. Kazemian, H. B., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. Expert Systems with Applications, 42(3), 1166-1177.
[5]. https://en.wikipedia.org/wiki/Phishing