International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Email Phishing Detection System using Url Phishing Technique and Machine Learning

Amaka Eugenia Ngozi[1], Ezea Jonathan Ikechukwu[2], Okpalla Chidimma Lilian[3] Theodora Onwuama [4], Ibeneme-Sabinus Ifeoma Livina[5] and Atomatofa Emmanuel Oghenero[6]

[1,5,6] Department of Cybersecurity, School of Information and Communication Technology, Federal University of Technology, Owerri, Imo State, Nigeria.

[2] Department of Information Technology, First Bank Nigeria Ltd, 35 Marina Lagos, Nigeria.

[3] Department of Computer Science, School of Information and Communication Technology, Federal University of Technology, Owerri, Imo State, Nigeria.

[4] Toronto Business College, 4000 Victoria Park Avenue, Toronto ON M2H 3P4. Canada

Abstract - The damage caused by phishing attacks is heart-breaking. Hence, APWG, decade report ascertained accelerated speed of phishing attacks to 26.19% in 2023. However, the new system developed browser plugin service at real-time phishing detection to educate the internet users on the danger posed on phishing exploits. Thus, 40,000 datasets were collected from PhishStat and the data collected was splitted into train-test split of 80% training and 20% testing respectively. Also, Random Forest was adopted for the training since it's capable of detecting non-linear patterns in data and handling imbalance features. Similarly, JavaScript was used for the implementation and results obtained obtained obtained ascertained 82.20% accuracy from post-test conducted. Therefore, getting internet users pre-informed stands out to be the best strategy to curb the phishing exploits.

Keywords: Email, Phishing, URL, ML, Websites

I. INTRODUCTION

Phishing is a malicious attempt of an impersonator to deceitfully obtain the targets' sensitive information [27,29] Though, phishing has been an old cybercrime and maintains its' criminal activities because it has an effective dynamic and ever-evolving practices [24]. More technically in detecting the phishing threats is the advanced development techniques using Jframe to design illegitimate and web-windows in a similar-look like legitimate browser in a webpage [5]. Consequently, attackers keep innovating their techniques to have better chances of executing attacks. ([12].

However, it is heart-breaking to see the damage caused by the phishing exploits as reported by Anti-Phishing Working Group (APWG, 2014 – 2023), [11,16,17,18]. Hence, the decade report ascertained gradual phishing exploits of 3.23% of phishing websites detected in 2014 and maintained an accelerated speed of 26.19% of phishing websites detected in 2023. Though, this was because of a tactical phishing approach, which craftily deceived the targets into clicking on malicious URLs or requesting the targets to open an infected attachment [29]. Although, the influential factors for victims falling into phishing susceptibilities are poor security tips awareness to avert uncertainties, and individual's greed to accept unnecessary and unverified internet offers. Therefore, a browser plugin service for Real-Time Detection of Website Phishing Attacks was developed. This address pressing need for a simple yet innovative solution for mitigating phishing attacks by educating internet users with detailed knowledge on preventive measures.

II. ATTACK PATTERN

Figure 1 depicts the process of attacking the victim. It takes consistent observation to study target victim and obtain sensitive information.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

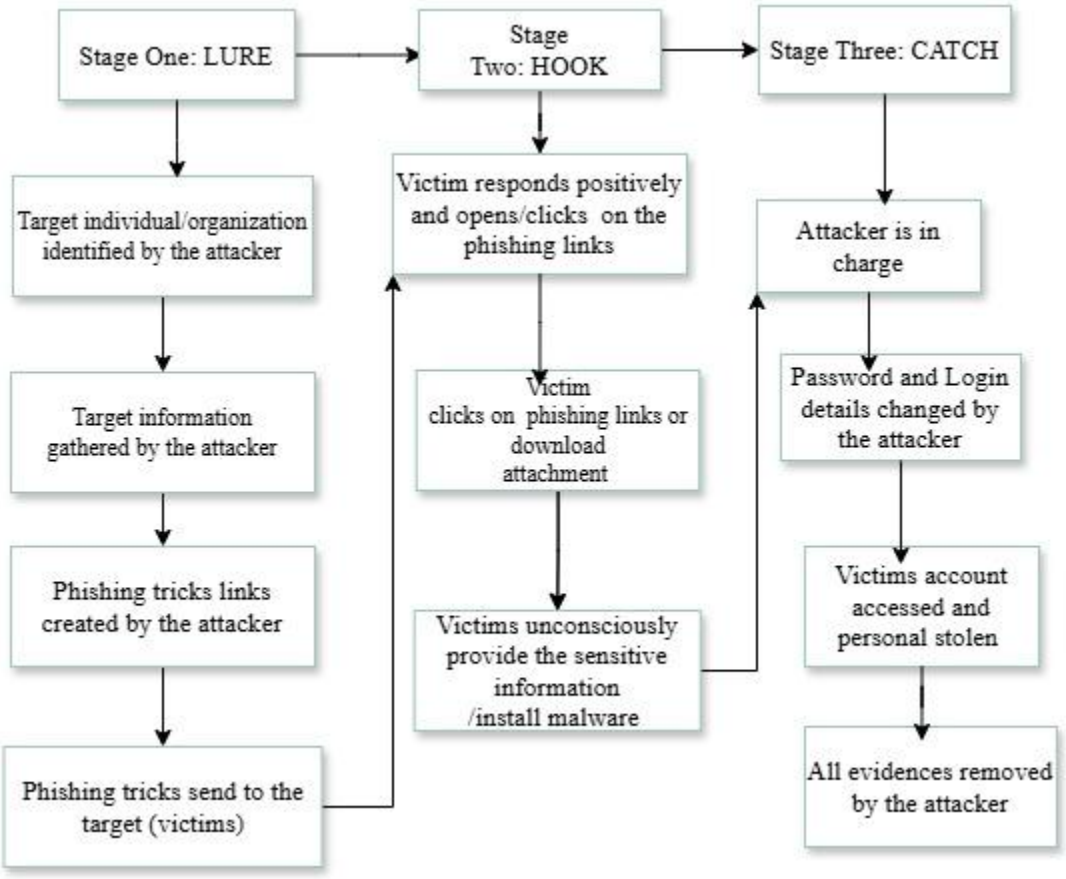

Figure 1: Lure-Hook-Catch Trend Pattern

Figure 1 captured the phishing trick phases, which were organized in 3 different patterns, such as lure, hook and catch, called lure-hook-catch trend pattern. Hence, the attacker in this pattern studies the individuals or organizations to identify the target and further obtains the target information to trick. However, the attacker creates phishing links such as websites, emails, or messages that have a similar legitimate-look, and sends them to the target. Consequently, as the victim is being lured by the tricks of the attacker, he discloses sensitive information. Thus, this leverages the attacker to take charge of the victims' account, changes the login information and knocks out the victim to exploit attacks.

III. LITERATURE REVIEW

A good number of previous research has been reviewed, and various phishing threats exploits, both financial and reputation damage caused have been identified. Also, the methods identified in a previously research include the methods adopted by the various researchers enhancing the performances of the existing phishing detection models and results obtained. Thus, the following are the reviewed literature.

Email remains an official platform for transaction notification and a preferred communication technique to seamlessly communicate with each other globally [13,19,26]. In research carried out by [10,16], an increase of 65% phishing exploits was recorded in 2016, with a huge loss, worth of $1,220,523. Though, the researchers adopted a supervised machine learning models with a classification of training the various dataset such as decision tree, Random Forest, Support Vector Machine (SVM), XGBoost and multilayer perceptions. The conclusion of the research showed that random forests have the best security accuracy among other listed models.

Additionally, [21,22] in research identified spear phishing, Smishing and whaling as the most recent phishing techniques, with highly sophisticated tools difficult to detect. The researchers also analyzed the trends of phishing attack and equally evaluated the effectiveness of the preventive models developed for prevention of phishing attacks. The results of the analysis conducted show that

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

user education is a paramount significance in phishing attack prevention. The research concluded by recommending an increase in user awareness and policy enhancement in organisations to prevent phishing attacks. Nevertheless, [1,6,24,29,23] identified the consequences of phishing attacks such as identity theft, loss of sensitive information among others. The researchers evaluated these attacks to ascertain the status and reviewed the existing techniques of phishing. However, the researchers proposed an Anatomy of phishing, involving the attack phases, attacker's types, targets, and techniques, among others. Hence, the research concluded that the designed Anatomy will greatly help the understanding of the readers, not necessarily involving awareness on the life processes of phishing attack and the various techniques employed to have a successful attack.

Furthermore, [3,23,15] ,identified sophisticated phishing techniques, employed to deceive victims to reveal their sensitive information or download a phishing website. The research presented a machine learning technique to identify phishing websites, focusing on accuracy and efficiency detection. Also, the research integrated the "CfsSubsetEval attribute evaluator with K-means clustering algorithms" for enhancing the phishing detection. Likewise**,** [9,20] identified in research conducted, the damage caused by the effect of phishing attacks to both individuals and organisations. The research explored the various techniques to detect and prevent the re-occurrences of phishing attacks, aided by a comprehensive experiment, with efficient demonstration of both detection and mitigation analysis. However, the result of the findings showed that with high offer demonstrated by machine learning algorithm on detection accuracy, it still requires a continuous update to maintain an effective technique against phishing attacks. The research also recommended user education as a lasting preventive technique.

Moreso, [8,2,4,14,28], discovered in research conducted that the phishing attack techniques are becoming more sophisticated, making it difficult to detect new attacks. Though, as phishers keep changing their attack tactics to have successful exploits, the researchers also strategically explore various available options to avert their malicious plans. Hence, in a quest to combat the phishing attack threats, [15] conducted a physical training program for employees in a specific organization, and discovered the impact it created, and further recommended for continuous training reinforcement to enable a lasting solution. Thus, the method employed so far cannot sufficiently eradicate the phishing attacks [19]. Alarmingly, to ascertain the effectiveness of the educative mechanisms used for training of users on minimizing the associated risk of being affected in a cyber-attack are not tested, rather, the result is being assumed [7]

IV. RESEARCH METHODOLOGY

Figure 2 depicts the system's methodology

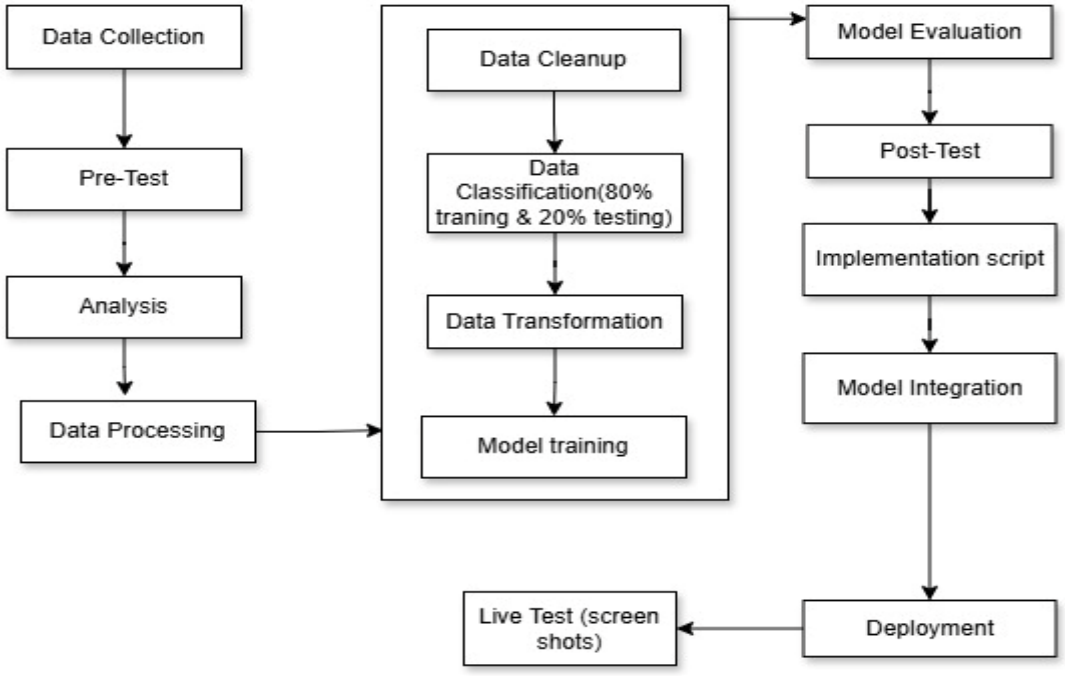

Figure 2: Research Methodology

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Data Collection: Datasets of 40,000 phishing websites were collected from PhishStat. The datasets contain URLs that are pre-labeled as either phishing (unsafe) or legitimate (safe), which is essential for training a detection model. Testing was conducted twice: Initial Testing: to evaluate the accuracy of existing phishing detection methods and Post-Training Testing: to test its performance on the testing set to validate accuracy and evaluate metrics, such as accuracy, precision, recall, and F1- score. The pre-test score obtained was 78%, which helped to establish baseline performance using default or no trained models. It also checks if the features extracted have any predictive power before model optimization, helps in comparing improvements after training and hyperparameter tuning. Thus, the data collected was splitted into two groups: 80% for training the system to detect phishing and 20% for testing. Also, data pre-processing was done to clean up any useless or empty parts since Machine learning models require numeric and structured input. Similarly, URLs, which are text-based, were transformed into numbers so that machine learning algorithms could understand them. Random Forest model was used for the training phase. This is a more sophisticated model using multiples decision trees to improve detection accuracy. Also, it handles imbalanced features well, detects non-linear patterns in data, is resistant to overfitting, especially with more trees, and performs feature selection implicitly. During training, these models learnt to recognize patterns in the data that distinguish phishing URLs from legitimate ones. The training was organized with Train-Test split of 80% training and 20% testing using Train-test_Split(). Furthermore, in actualizing Feature engineering, it applied the custom feature extraction function on both training and test data. while RandomForestClassifier of 42 as model choice was selected as a random state. The Hyperparameter Tuning used GridSearchCV with 3-fold cross-validation and the following parameter grid were obtained:
param_grid = {

'n_estimators': [100, 200],

'max_depth': [10, 20, None],

'min_samples_split': [2, 5],

'min_samples_leaf': [1, 2]

}

However, the new system is a browser plugin service that monitors URLs and links entered in the user's browser address bar and detects if the URL entered is a phishing attempt to mitigate it in real-time using traditional list-based technique and heuristic analysis. However, if it is a legitimate website, the user will be allowed to proceed to enter the website without any interruptions. Hence, the plugin will first check if the website URL entered is already blacklisted or whitelisted, if the status is ascertained, the system goes further to analyze the website's URL to classify it as either legitimate or phishing and sends a notification to the user. Thus, the alert notification is always active whenever the browser plugin service is accessed by the user, to popup security tips for awareness and safety on internet dangers.

V. IMPLEMENTATION SCRIPT

```
import pandas as pd

from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import train_test_split, GridSearchCV

from sklearn.metrics import accuracy_score

import re

import joblib

# Feature extraction from URL

def extract_features_from_url(url):

    features = {}

    features['length_url'] = len(url)

    features['nb_dots'] = url.count('.')
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

```
    features['nb_hyphens'] = url.count('-')
    features['nb_at'] = url.count('@')
    features['nb_qm'] = url.count('?')
    features['nb_and'] = url.count('&')
    features['nb_eq'] = url.count('=')
    features['nb_underscore'] = url.count('_')
    domain = re.findall(r'://([^/]+)', url)
    features['length_hostname'] = len(domain[0]) if domain else 0
    subdomains = domain[0].split('.')[:-2] if domain else []
    features['nb_subdomains'] = len(subdomains)
    features['contains_login'] = 1 if 'login' in url else 0
    features['contains_secure'] = 1 if 'secure' in url else 0
    return features
# Load data
data = pd.read_csv('phishing_urls.csv')
data['status'] = data['status'].map({'legitimate': 0, 'phishing': 1})
X = data['url']
y = data['status']
# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Extract features
X_train_features = [extract_features_from_url(url) for url in X_train]
X_test_features = [extract_features_from_url(url) for url in X_test]
X_train_df = pd.DataFrame(X_train_features)
X_test_df = pd.DataFrame(X_test_features)
# Model and Grid Search
model = RandomForestClassifier(random_state=42)
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [10, 20, None],
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

```
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2]
}
grid_search = GridSearchCV(model, param_grid, cv=3, verbose=2, n_jobs=-1)
grid_search.fit(X_train_df, y_train)
# Best model and evaluation
best_model = grid_search.best_estimator_
joblib.dump(best_model, 'model.pkl')
y_pred = best_model.predict(X_test_df)
accuracy = accuracy_score(y_test, y_pred)
print(f"Model Accuracy: {accuracy * 100:.2f}%")
Integration Model in Script (Process)
# Prediction on new URLs
new_urls = [
    "http://example.com/login.php",
    "http://malicious.com/steal_data",
    # ...
]
# Feature extraction
new_url_features = [extract_features_from_url(url) for url in new_urls]
new_urls_df = pd.DataFrame(new_url_features)
# Load model & predict
model = joblib.load('model.pkl')
predictions = model.predict(new_urls_df)
probabilities = model.predict_proba(new_urls_df)
# Display results
for url, pred, prob in zip(new_urls, predictions, probabilities):
    status = 'legitimate' if pred == 0 else 'phishing'
    severity_score = prob[1] * 100
    print(f"URL: {url}, Status: {status}, Severity Score: {severity_score:.2f}%")
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

VI. RESULTS

A screenshot showing the results of the trained machine learning model. The confusion matrix shows classification performance.

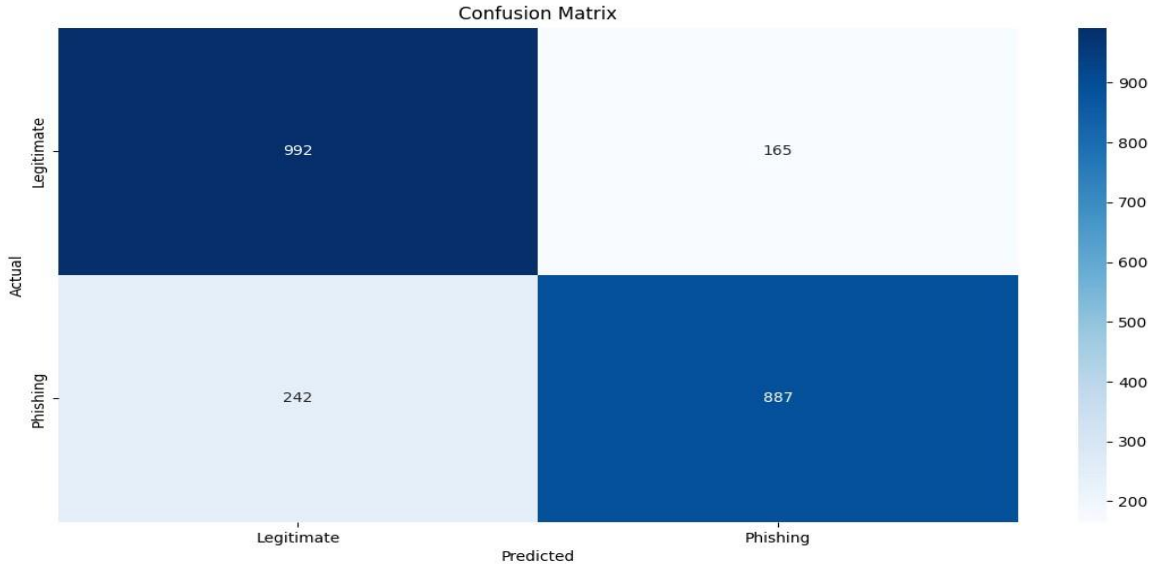

Figure 3: Confusion matrix

The confusion matrix illustrates the performance of the trained phishing detection model. It includes the following metrics:

- **True Positives (TP)**: Number of phishing URLs correctly classified as phishing.
- **True Negatives (TN)**: Number of legitimate URLs correctly classified as legitimate.
- **False Positives (FP)**: Number of legitimate URLs incorrectly classified as phishing.
- **False Negatives (FN)**: Number of phishing URLs incorrectly classified as legitimate.

A confusion matrix with high TP and TN values alongside low FP and FN values indicates that the model is effective in distinguishing phishing from legitimate URLs. However, a higher FN rate can be critical, as undetected phishing attempts can result in significant harm to users.

```
Model Accuracy: 82.20%
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.86      0.83      1157
           1       0.84      0.79      0.81      1129

    accuracy                           0.82      2286
   macro avg       0.82      0.82      0.82      2286
weighted avg       0.82      0.82      0.82      2286
```

Flask API Functionality

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

API endpoint response in Postman.
Input: A phishing URL sent as a POST request.
Output: JSON response indicating status (e.g., "phishing") and severity score.

Input: A phishing URL is sent as a POST request to the Flask API.

Output: The API returns a JSON response indicating the URL status as "phishing" along with a severity score. The API is functional and integrates effectively with the trained model to provide real-time feedback on submitted URLs. The severity score adds granularity by quantifying the confidence level of phishing detection

Input: A legitimate URL sent as a POST request.

Output: JSON response indicating status (e.g., "legitimate ") and severity score.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The API correctly identifies legitimate URLs, showcasing its reliability and minimal false positives. This demonstrates practical applicability for end-users and integration into other systems.

VII. CONCLUSION AND RECOMMENDATION

The developed Browser Plugin Service for Real-Time Detection of Website Phishing Attacks stands as a pivotal tool in the rapidly expanding realm of internet usage. By consolidating diverse phishing information, it addresses pressing need for a simple yet innovative solution for mitigating phishing attacks. Hence, as the numbers of internet users grow vastly daily, it necessitates the need to protect naive or highly susceptible users from phishing attacks and empower them with detailed knowledge on preventive measures. However, a PhishBlocker" is highly recommended for all internet users to be protected from phishing attacks. Also, further research is equally required to expand the scope of this research and keep users updated on internet vulnerabilities.

8.0 Further Research

1. Expansion of Dataset is required to incorporate more diverse and updated phishing URLs to improve the model's adaptability to emerging phishing techniques.
2. Multilingual Support is required to extend the system to detect phishing attempts in non-English URLs, as phishing campaigns often target multilingual populations.

REFERENCES-

[1] Abufardeh, S., & Falah, B. (2023). The State of Phishing Attacks and Countermeasures. *Sameer Abufardeh & Bouchaib Falah International Journal of Computer Science & Security (IJCSS)*, *17*(4), 54–70. https://www.cscjournals.org/manuscript/Journals/IJCSS/Volume17/Issue4/IJCSS-1702.pdf

[2] Alnemari, S. (2018). Applied Sciences. *Early Writings on India*, 124–134. https://doi.org/10.4324/9781315232140-14

[3] Al-sabbagh, A., Hamze, K., & Khan, S. (2024). *An Enhanced K-Means Clustering Algorithm for Phishing Attack Detections*. 1–18.

[4] Alzboon, M. S., & Alzboon, L. (2025). *Guardians of the Web : Harnessing Machine Learning to Combat Phishing Attacks Guardianes de la Web : Aprovechando el Aprendizaje Automático para Combatir los Ataques de Phishing*. https://doi.org/10.56294/gr202591

[5] Asiri, S., Xiao, Y., Alzahrani, S., & Li, T. (2024). PhishingRTDS: A real-time detection system for phishing attacks using a Deep Learning model. *Computers and Security*, *141*. https://doi.org/10.1016/j.cose.2024.103843

[6] Bethany, M., Galiopoulos, A., Bethany, E., Karkevandi, M. B., Vishwamitra, N., & Najafirad, P. (2024). *Large Language Model Lateral Spear Phishing: A Comparative Study in Large-Scale Organizational Settings*. 1–18. http://arxiv.org/abs/2401.09727

[7] Beu, N., Jayatilaka, A., Zahedi, M., Babar, A., Hartley, L., Lewinsmith, W., & Baetu, I. (2023). Falling for phishing attempts: An investigation of individual differences that are associated with behavior in a naturalistic phishing simulation. *Computers and Security*, *131*, 103313. https://doi.org/10.1016/j.cose.2023.103313

[8] Bezerra, A., Pereira, I., Rebelo, M. Â., Coelho, D., Oliveira, D. A. De, Costa, J. F. P., & Cruz, R. P. M. (2024). A case study on phishing detection with a machine learning net. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-024-00579-w

[9] Birlea, M. C. (2020). *Phishing Attacks: Detection And Prevention*. http://arxiv.org/abs/2004.01556

[10] Desouza, K. C. (2024). Machine Learning Techniques. In *Managing Knowledge with Artificial Intelligence* (Issue January). https://doi.org/10.5040/9798216977254.0008

[11] Djatsa, F. (2020). Threat Perceptions, Avoidance Motivation and Security Behaviors Correlations. *Journal of Information Security*, *11*(01), 19–45. https://doi.org/10.4236/jis.2020.111002

[12] Esteban, D., Vivas, D., Yecid, W., Pena, G., Cristancho, S. P., & Rojas, A. E. (2024). *A Controlled Phishing Attack in a University Community : A Case Study A Controlled Phishing Attack in a University Community : A Case Study*. *September*. https://doi.org/10.58346/JISIS.2024.I2.007

[13] Gallo, L., Gentile, D., Ruggiero, S., Botta, A., & Ventre, G. (2024). The human factor in phishing: Collecting and analyzing user behavior when reading emails. *Computers and Security*, *139*. https://doi.org/10.1016/j.cose.2023.103671

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[14] Jayaraj, R., Pushpalatha, A., Sangeetha, K., Kamaleshwar, T., Udhaya Shree, S., & Damodaran, D. (2024). Intrusion detection based on phishing detection with machine learning. *Measurement: Sensors*, *31*(June 2023), 101003. https://doi.org/10.1016/j.measen.2023.101003

[15] Khan, M. H., & Muntaha, S. T. (2024). *Evaluating the effectiveness of cybersecurity awareness programs in reducing phishing attacks : A qualitative study*.

[16] Kolla, J., Praneeth, S., Baig, M. S., & Karri, G. R. (2022). *A comparison study of machine learning techniques for phishing detection*. *4*(1), 21–33.

[17] Lindez-Macarro, M. E., Gallego-Losada, R., Montero-Navarro, A., & Rodríguez-Sánchez, J. L. (2025). A bibliometric analysis of financial fraud exploiting the elderly in the digital age. *International Journal of Bank Marketing*, *43*(5), 943–978. https://doi.org/10.1108/IJBM-11-2023-0634

[18] Lohiya, R., & Thakkar, A. (2024). A Compendium on Risk Assessment of Phishing Attack Using Attack Modeling Techniques. *Procedia Computer Science*, *235*, 1105–1114. https://doi.org/10.1016/j.procs.2024.04.105

[19] Marshall, N., Sturman, D., & Auton, J. C. (2024). Exploring the evidence for email phishing training: A scoping review. *Computers and Security*, *139*(December 2023), 103695. https://doi.org/10.1016/j.cose.2023.103695

[20] Mohamed, G., Visumathi, J., Mahdal, M., & Anand, J. (2022). *An Effective and Secure Mechanism for Phishing Attacks Using a Machine Learning Approach*.

[21] Morrow, E. (2024). Scamming higher ed: An analysis of phishing content and trends. *Computers in Human Behavior*, *158*(March), 108274. https://doi.org/10.1016/j.chb.2024.108274

[22] Multidisiplin, J., & Volume, I. (2024). *An Analysis of Phishing Attacks : Information Technology Security : Cybercrime and Its Solutions*. *3*(05), 696–712.

[23] Mutlutürk, M., Wynn, M., & Metin, B. (2024). Phishing and the Human Factor: Insights from a Bibliometric Analysis. *Information (Switzerland)*, *15*(10). https://doi.org/10.3390/info15100643

[24] Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S., & Porras, J. (2023a). Mitigation strategies against the phishing attacks: A systematic literature review. In *Computers and Security* (Vol. 132). https://doi.org/10.1016/j.cose.2023.103387

[25] Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S., & Porras, J. (2023b). Mitigation strategies against the phishing attacks: A systematic literature review. *Computers and Security*, *132*, 103387. https://doi.org/10.1016/j.cose.2023.103387

[26] Russell, E., Jackson, T. W., Fullman, M., & Chamakiotis, P. (2024). Getting on top of work-email: A systematic review of 25 years of research to understand effective work-email activity. *Journal of Occupational and Organizational Psychology*, *97*(1), 74–103. https://doi.org/10.1111/joop.12462

[27] Shaik, C. (2020). Counter Challenge Authentication Method: A Defeating Solution to Phishing Attacks. *International Journal of Computer Science, Engineering and Applications*, *10*(1), 1–8. https://doi.org/10.5121/ijcsea.2020.10101

[28] Siddharth, Srivastava, R., Raj, H., Dwivedi, S., Dwivedi, S., Singh, R., & Chaurasiya, N. (2025). Detection of phishing attacks using machine learning. *Emerging Trends in Computer Science and Its Application*, *July*, 105–109. https://doi.org/10.1201/9781003606635-@

[29] Spithoven, R., & Anthonie, D. (2024). Who will take the bait? Using an embedded, experimental study to chart organization-specific phishing risk profiles and the effect of a voluntary microlearning among employees of a Dutch municipality. *Journal of Cybersecurity , 2024*, *June*. https://doi.org/https://doi.org/10.1093/cybsec/tyae010 Research