International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 8 Issue 5, Sep-Oct 2024 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Phish Guard Phishing Website using Machine Learning Algorithms

Abhishek Jadhao[1], Lakshmi Mahindre[2], Komal Rahangdale[3], Vinita Singh[4], Prof. Rina Shipurkar[5], Prof. Usha Kosarkar[6]

[1,2,3,4]School of Science, G. H. Raisoni University, Amravati, Maharashtra, India
[5]Assistant Professor, G. H. Raisoni University, Amravati, Maharashtra, India
[6]Assistant Professor, G H Raisoni College of Engineering & Management, Nagpur, Maharashtra, India

ABSTRACT

Phishing attacks pose a significant threat to individuals and organizations, leading to substantial financial and reputational damage. Traditional detection methods, such as blacklists and signature-based techniques, often fall short in identifying sophisticated phishing attempts. This research proposes a comprehensive system that leverages machine learning and deep learning techniques to detect and delete phishing threats in emails and websites. The system integrates multiple modules to analyze email structures, text content, and URLs, ensuring a robust defense against phishing attacks. By employing advanced algorithms like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, the system achieves high accuracy in identifying phishing attempts. Experimental results demonstrate the system's effectiveness in real-world scenarios, significantly reducing the risk of phishing attacks. This study contributes to the field of cybersecurity by providing a scalable and efficient solution for phishing detection and mitigation, paving the way for safer online interactions. The anonymous and uncontrollable framework of the Internet is more vulnerable to phishing attacks. Existing research works show that the performance of the phishing detection system is limited. There is a demand for an intelligent technique to protect users from the cyber-attacks. In this study, the author proposed a URL detection technique based on machine learning approaches. A recurrent neural network method is employed to detect phishing URL. Researcher evaluated the proposed method with 7900 malicious and 5800 legitimate sites, respectively. The experiments' outcome shows that the proposed method's performance is better than the recent approaches in malicious URL detection. It is one of the familiar attacks that trick users to access malicious content and gain their information. In terms of website interface and uniform resource locator (URL), most phishing webpages look identical to the actual webpages. Various strategies for detecting phishing websites, such as blacklist, heuristic, Etc., have been suggested.

How to cite this paper: Abhishek Jadhao | Lakshmi Mahindre | Komal Rahangdale | Vinita Singh | Prof. Rina Shipurkar | Prof. Usha Kosarkar "Phish Guard Phishing Website using Machine Learning Algorithms" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-8 | Issue-5, October 2024, pp.625-634, URL: www.ijtsrd.com/papers/ijtsrd69425.pdf

IJTSRD69425

KEYWORDS: *machine learning, phish attack, anti phishing tool, cybersecurity solutions, url scanning*

I. INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing . In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion . Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. The general method to

detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack. Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high.

II. RELATED WORK

Phishing attacks are categorized according to Phisher’s mechanism for trapping alleged users. Several forms of these attacks are keyloggers, DNS toxicity, Etc. The initiation processes in social engineering include online blogs, short message services (SMS), social media platforms that use web 2.0 services, such as Facebook and Twitter, file-sharing services for peers, Voice over IP (VoIP) systems where the attackers use caller spoofing IDs . Each form of phishing has a little difference in how the process is carried out in order to defraud the unsuspecting consumer. E-mail phishing attacks occur when an attacker sends an e-mail with a link to potential users to direct them to phishing websites.

A. CLASSIFICATION OF PHISHING ATTACK TECHNIQUE

Phishing websites are challenging to an organization and individual due to its similarities with the legitimate websites . Fig 1 presents the multiple forms of phishing attacks. Technical subterfuge refers to the attacks include Keylogging, DNS poisoning, and Malwares. In these attacks, attacker intends to gain the access through a tool / technique. On the one hand, users believe the network and on the other hand, the network is compromised by the attackers. Social engineering attacks include Spear phishing, Whaling, SMS, Vishing, and mobile applications. In these attacks, attackers focus on the group of people or an organization and trick them to use the phishing URL . Apart from these attacks, many new attacks are emerging exponentially as the technology evolves constantly.

Fig 1. Multiple forms of phishing attacks.

B. PHISHING DETECTION APPROACHES-

Phishing detection schemes which detect phishing on the server side are better than phishing prevention strategies and user training systems. These systems can be used either via a web browser on the client or through specific host-site software presents the classification of Phishing detection approaches. Heuristic and ML based approach is based on supervised and unsupervised learning techniques. It requires features or labels for learning an environment to make a prediction. Proactive phishing URL detection is similar to ML approach. However, URLs are processed and support a system to predict a URL as a legitimate or malicious . Blacklist and Whitelist approaches are the traditional methods to identify the phishing sites . The exponential growth of web domains reduces the performance of the traditional method .

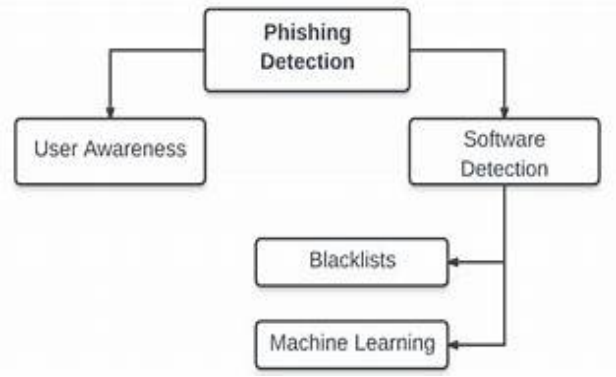

Fig 2. Anti—Phishing approaches

The existing methods rely on new internet users to a minimum. Once they identify phishing website, the site is not accessible, or the user is informed of the probability that the website is not genuine. This approach requires minimum user training and requires no modifications to existing website authentication systems. The performance of the detection systems is calculated according to the following:

- Number of True Positives (TP): The total number of malicious websites.
- Number of True Negatives (TN): The total number of legitimate websites.
- Number of False Positives (FP): The total number of incorrect predictions of legitimate websites as a malicious website.
- Number of False Negatives (FN): The total number of incorrect predictions of malicious websites as a legitimate website

C. RESEARCH QUESTIONS-

Researcher framed the Research Questions (RQ) according to the objective of the study and its background. They are as follows:

- RQ1—How URL detectors identify the phishing URLs or websites?
- RQ2—How to apply ML methods to classify malicious and legitimate websites?
- RQ3—How to evaluate a URL detector performance?

On the one hand, RQ1 and RQ2 assist to develop a ML based phishing detection system for securing an network from phishing attacks. On the other hand, RQ3 specifies the importance of the performance evaluation of a phishing technique. To address RQ1, authors found some recent literature related to URL detection using Artificial Intelligence (AI) techniques. The following part of this section presents the studies in detail with Table 2.

Authors in the study proposed a URL-based anti-phishing machine learning method. They have taken 14 features of the URL to detect the website as a malicious or legitimate to test the efficiency of their method. More than 33,000 phishing and valid URLs in Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers were used to train the proposed system. The phishing detection method focused on the learning process. They extracted 14 different features, which make phishing websites different from legitimate websites. The outcome of their experiment reached over 90% of precision when websites with SVM Classification are detected.

The study explored multiple ML methods to detect URLs by analyzing various URL components using machine learning and deep learning methods. Authors addressed various methods of supervised learning for the identification of phishing URLs based on lexicon, WHOIS properties, PageRank, traffic rank information and page importance properties. They studied how the volume of different training data influences the accuracy of classifiers. The research includes Support Vector Machine (SVM), K-NN, random forest classification (RFC) and Artificial Neural Network (ANN) techniques for the classification.

Based on the output without and with the functionality selection a comparative study of machine learning algorithms is carried out in the study . Experiments on a phishing dataset were carried out with 30 features including 4898 phished and 6157 benign web pages. Several ML methods were used to yield a better outcome. A method for selecting functions is subsequently employed to increase model performance. Random forests algorithm achieved the highest accuracy prior to and after the selection of features and dramatically increase building time. The results of the experiment shown that using the selection approach with machine learning algorithms can boost the effectiveness of the classification models for the detection of phishing without reducing their performance.

In this study authors proposed URLNet, a CNN-based deep-neural URL detection network. They argued that current methods often use Bag of Words(BoW) such as features and suffered some essential limitations, such as the failure to detect sequential concepts in a URL string, the lack of automated feature extraction and the failure of unseen features in real—time URLs. They developed a CNNs and Word CNNs for character and configured the network. In addition, they suggested advanced techniques that were particularly effective for handling uncommon terms, a problem commonly exist in malicious URL detection tasks. This method can permit URLNet to identify embeddings and use sub word information from invisible words during testing phase.

Authors suggested a URL detector for high precision phishing attacks. They argued that the technique could be scaled to various sizes and proactively adapted. For both legitimate and malicious URLs a limited data collection of 572 cases had been employed. The characteristics were extracted and then weighed as cases to use in the prediction process. The test results were highly reliable with and without online phishing threats. For the improvement of the accuracy, Genetic algorithm (GA) has been used.

TABLE NO 1.COMPARISON STUDY OF LITERATURE

S. No.	Authors	Contributions	Limitations
1	Jain A.K., and Gupta B.B [2]	Employed both NB and SVM algorithms to identify the malicious websites.	Both SVM and NB are slow learners and does not store the previous results in the memory. Thus, the efficiency of the URL detector may be reduced.
2	Purbay M., and Kumar D. [3]	Utilized multiple ML methods for classifying URLs.	They compared the performance of different types of ML methods. However, there were no discussions about the retrieval capacity of the algorithms.
3	Gandotra E., and Gupta D. [4]	Applied multiple classification algorithms for detecting malicious URLs.	The outcome of the experiments demonstrated that the performance of the system was better rather than other ML methods. However, It lacks in handling larger volume of data.
4	Hung Le et al., [5]	Proposed a deep learning based URL detector. Authors argued that the method can produce insights from URL.	Deep learning methods demand more time to produce an output. In addition, it processes the URL and matches with library to generate an output.
5	Hong J. et al., [6]	Developed a crawler to extract URLs from data repositories. Applied lexical features approach to identify the phishing websites.	The performance evaluation was based on crawler-based dataset. Thus, there is no assurance for the effectiveness of the URL detector with real time URLs.
6	Kumar J. et al., [7]	Proposed a URL detector based on blacklisted dataset. Also, a lexical feature approach was employed to classify malicious and legitimate websites.	Authors employed an older dataset which can reduce the performance of the detector with real—time URLs.
7	Hassan Y.A. and Abdelfettah B. [8]	Suggested a URL detector for classifying websites and predict the phishing websites. They used GA technique to improve the performance.	The performance of GA based URL detector was better; nonetheless, the predicting time was huge with complex set of URLs.
8	Rao RS and Pais AR. [9]	Authors employed page attributes include logo, favicon, scripts and styles.	The method employed a server for updating the page attributes that reduces the performance of the detecting system.
9	Aljofey A et al. [10]	A CNN based detecting system for identifying the phishing page. A sequential pattern is used to find URLs.	The existing research shows that the performance of CNN is better for retrieving images rather than text.
10	AlEroud A and Karabatis Gv[11]	Generative adversarial network is used in the research to bypass a detection system.	Neural Network based detection system can identify the impression of an adverse network by learning the environment.

III. RESEARCH METHODOLOGY

RQ3 stated that how ML method can be employed to identify a malicious or legitimate URL. To present a solution, authors

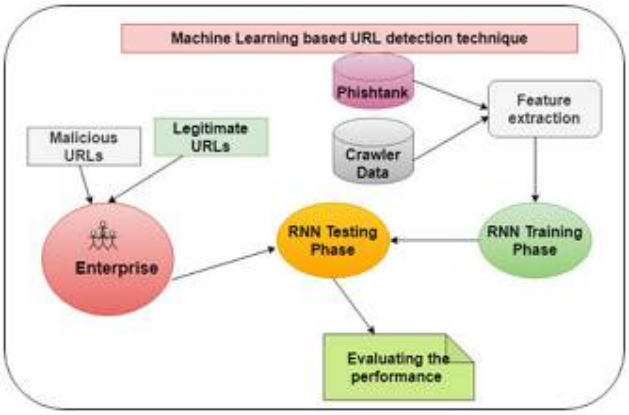

FIG NO.1 RESEARCH FRAMEWORK

Let ∑mn=0xn be the set of URLs where m is the maximum limit for the number (n) of URLs. Let M, L ∈ x_n be the malicious and legitimate, accordingly. Suppose M and L contains the properties Pm and Pl, respectively. The proposed framework employs RNN—LSTM to identify the properties Pm and Pl in an order to declare an URL as malicious or legitimate. The following equations from 1 to 4 presents the method for identifying the malicious URL. The term "recurring neural network" implies two broad groups of networks of a similar general structure, where one is a finite, and the other is an infinite input. Both network groups contains time dynamic behaviour. A recurrent network of finite input is a directed acyclic graph that can be replaced by a purely feedforward neural network, whereas a recurrent network of infinite input is a directed cyclical graph that cannot be modified. The modified version of RNN is LSTM. It is a deep learning method, which prevents the gradient problem of RNN. Multiple gates are employed for improving the performance of LSTM. In comparison with RNN, LSTM prevents back propagation. Each input of LSTM generates an output that becomes an input for the following layer or module of LSTM. Eqs 1 to 4 illustrates the concept of the proposed study.

∑(M+L)=xn

1. Input=∑mn=0xn
2. Malicious=Output_RNN(Input(Pm))
3. Legitimate=Output_RNN(Input(Pl))

Cell state (CS)—It indicates the cell space that accommodate both long term and short-term memories.

Hidden state (HS)—This is the output status information that user use to determine URL with respect to the current data, hidden condition and current cell input. The secret state is used to recover both short-term and long-term memory, in order to make a prediction.

Input gate (IT)—The total number of information flows to the cell state.

Forget gate (FT)—The total number of data flows from the current input and past cell state into the present cell state.

Output gate (OT)—The total number of information flows to the hidden state.

Input: Data Repositories
Output: Raw Data

```
procedure DATA COLLECTION
    W ← ExtractData(Repositories)
    W1 ← FilterInvalidURL(W)
    N ← Count(W1)
    return W1,N
end procedure
```

ALGORITHM—DATA COLLECTION

illustrates the steps of data pre—process. url is one of the elements of URL dataset. In this process, the raw data is pre—processed by scanning each URL in th dataset. A set of functions are developed in order to remove the irrelevant data. Finally, D2 is the set of features returned by the pre—process activity.

ALGORITHM—DATA PRE-PROCESS

represents the processes of data transformation. "Num" is the vector returned by the data transformation process. During this process, each feature of D2 is converted as a vector. Each data in D2 is processed using the Generate Vectors function. A vector is generated and passed as an input to the training phase.

Input: Features(D2)
Output: Vectors

```
procedure DATA TRANSFORMATION(D2)
    while d ← D2 do
        Num ← GenerateVectors(d)
    end while
    return Num
end procedure
```

ALGORITHM—DATA TRANSFORMATION

provides the processes involved in the training phase. Each URL is processed with the support of vector. LSTMLib is one of the functions in the LSTM to predict an output using the vectors. The library is updated with the extracted features that contains the necessary data related to malicious and normal web pages. Thus, the iterative process is used to scan each vector and suspicious URL and generate a final outcome. Lastly, op is the prediction returned by the proposed method during the training phase.

Input: Vector(Num), URL
Output: URL-Type

```
procedure TRAINING PHASE(Num)
    while num ← Num do
        if num = Feature(URL) then
            op = Phishing URL
        else
            op = Legitimate URL
            if op = LSTMLib(feature) then
                op = Phishing URL
            else
                op = Legitimate URL
            end if
        end if
    end while
    return op
end procedure
```

ALGORITHM—TRAINING PHASE

indicates the testing phase of the proposed URL detection. The proposed processes each element from LSTMMemory function is compared with the vector of URL and decide an output. The f is the element of the feedback which is collected from the crawler that indicates the page rank of a website. The page rank indicates the value of a website and the lowest ranking website will be declared as malicious or suspicious to alert the users.

```
Input: URL
Output: Type of URL
procedure TESTING PHASE(URL)
    while url ← URL do
        if element ← LSTMMemory = Feature(URL) then
            op = Phishing URL
        else
            op = Legitimate URL
            feedback = phishtank(op)
            if element ← LSTMMemory = f ← feedback then
                op = Phishing URL
            else
                op = Legitimate URL
            end if
        end if
    end while
    return op
end procedure
```

ALGORITHM—TESTING PHASE

shows the snippet of epoch settings in the training phase. The epoch value is used to indicate the execution time of a method. The learning rate can be increased to improve the performance of a method.

IV. PROPOSED WORK

MACHINE LEARNING ALGORITHM-

Three machine learning classification model Decision Tree, Random forest and Support vector machine has been selected to detect phishing websites.

A. DECISION TREE ALGORITHM

One of the most widely used algorithm in machine learning technology. Decision tree algorithm is easy to understand and also easy to implement. Decision tree begins its work by choosing best splitter from the available attributes for classification which is considered as a root of the tree. Algorithm continues to build tree until it finds the leaf node. Decision tree creates training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label. In decision tree algorithm, gini index and information gain methods are used to calculate these nodes**.**

B. RANDOM FOREST ALGORITHM

Random forest algorithm is one of the most powerful algorithms in machine learning technology and it is based on concept of decision tree algorithm. Random forest algorithm creates the forest with number of decision trees. High number of tree gives high detection accuracy. Creation of trees are based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree. Among randomly selected features, random forest algorithm will choose best splitter for the classification and like decision tree algorithm; Random forest algorithm also uses gini index and information gain methods to find the best splitter. This process will get continue until random forest creates n number of trees. Each tree in forest predicts the target value and then algorithm will calculate the votes for each predicted target. Finally random forest algorithm considers high voted predicted target as a final prediction.

C. SUPPORT VECTOR MACHINE ALGORITHM

Support vector machine is another powerful algorithm in machine learning technology. In support vector machine algorithm each data item is plotted as a point in n-dimensional space and support vector machine algorithm constructs separating line for classification of two classes, this separating line is well known as hyperplane. Support vector machine seeks for the closest points called as support vectors and once it finds the closest point it draws a line connecting to them. Support vector machine then construct separating line which bisects and perpendicular to the connecting line. In order to classify data perfectly the margin should be maximum. Here the margin is a distance between hyperplane and support vectors.

V. ROPOSED RESEARCH MODEL

A. Objective

The primary goal is to develop an effective system for detecting phishing websites to enhance cybersecurity measures.

B. Data Collection

1. Data Sources:-

Phishing Dataset: Utilize publicly available datasets like the Phishing Websites Data Set from the UCI Machine Learning Repository or datasets from Kaggle.

Legitimate Websites: Scrape data from well-known legitimate websites to create a balanced dataset.

Real-Time Data: Integrate APIs (e.g., Google Safe Browsing) to get real-time data on phishing URLs.

2. Data Attributes :-

URL characteristics (length, entropy)

Domain age and registration details

Presence of HTTPS

Use of special characters or IP addresses

Page content features (e.g., keywords, meta tags)

C. Feature Extraction

1. Static Features:-

URL-Based Features: Analyze the structure of URLs (e.g., presence of subdomains, length).

Domain Features: Examine WHOIS information, age of domain, and registration details.

2. Dynamic Features:-

Content Analysis: Use NLP techniques to analyze the content of the webpage (e.g., identifying phishing-related keywords).

JavaScript Analysis: Inspect scripts for malicious behavior.

D. Model Selection

1. Machine Learning Algorithms:-

Supervised Learning: Train classifiers like Random Forest, Decision Trees, Support Vector Machines (SVM), and Neural Networks.

Ensemble Methods: Consider using ensemble techniques (e.g., Bagging, Boosting) to improve accuracy.

2. Deep Learning Approaches:-

Convolutional Neural Networks (CNN): For image-based phishing detection.

Recurrent Neural Networks (RNN): To analyze sequences in URL patterns.

E. Implementation Framework

1. Tech Stack:-

Backend: Python (Flask/Django) for server-side implementation.

Frontend: HTML, CSS, JavaScript frameworks (e.g., React) for user interface.

Database: SQL/NoSQL databases to store website data and user queries.

2. API Integration:-

Incorporate third-party APIs for real-time checking and threat intelligence feeds.

F. Evaluation Metrics

1. Performance Metrics:-

Accuracy: Percentage of correctly identified phishing vs. legitimate sites.

Precision and Recall: To evaluate the balance between false positives and false negatives.

F1 Score: Harmonic mean of precision and recall to assess overall model performance.

2. Cross-Validation:-

Use k-fold cross-validation to ensure robustness of the model.

G. User Interface Design

1. Web Interface:-

Simple user input form to enter URLs for analysis.

Display results with confidence scores and actionable insights.

2. User Feedback:-

Implement a feedback loop where users can report false positives/negatives to improve the model.

H. Deployment and Monitoring

1. Deployment:-

Deploy the model using cloud platforms (e.g., AWS, Azure) for scalability.

2. Monitoring:-

Continuously monitor the model's performance and update it based on new phishing techniques and data.

I. Ethical Considerations

Ensure user privacy and data security.

Maintain transparency about data usage and model limitations.

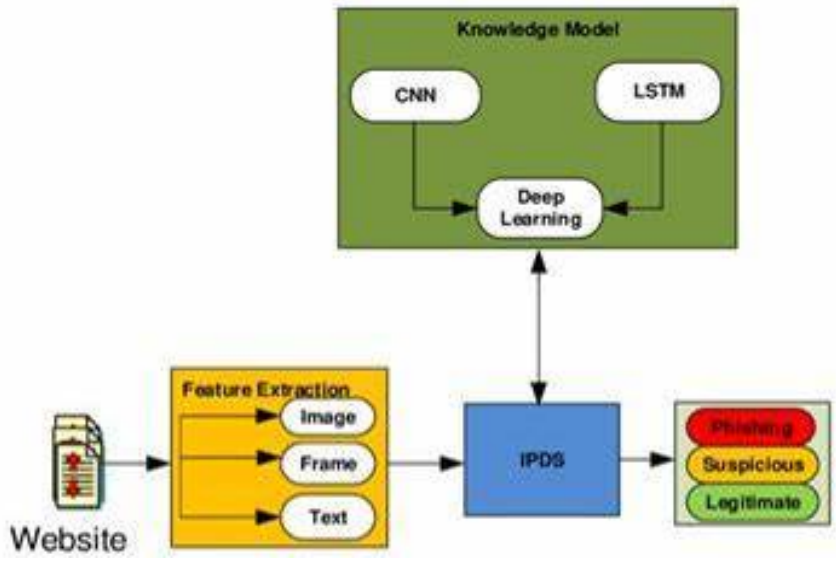

Fig1. Proposed research model

VI. RESULT ANALYSIS –

To evaluate the efficiency of a system, we use certain parameters. For each machine learning model, we calculate the Accuracy, Precision, Recall, F1 Score and ROC curve to determine its performance. Each of these metrics is calculated based on True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

In the case of URL classification, True Positive (TP) is the number of phishing URLs that are correctly classified as phishing. True Negative (TN) is the number of legitimate URLs that are correctly classified as legitimate. False Positive (FP) is the number of legitimate URLs that are classified as phishing. False Negative (FN) is the number of phishing URLs that are classified as legitimate. These values are summarized in Table IV called Confusion Matrix.

	Predicted Phishing	Predicted Legitimate
Actual Phishing	TP	FN
Actual Legitimate	FP	TN

TABLE:- CONFUSION MATRIX FOR PHISHING DETECTION

Precision is the number of URLs that are actually phishing out of all the URLs predicted as phishing. It measures the classifiers exactness. The formula to calculate precision is given by Equation (1) below.

1. Recall is the number of URLs that the classifier identified as phishing out of all the URLs that are actually phishing. It is also called sensitivity or true positive rate. It is an important measure and should be as high as possible.
2. F1-Score is the weighted average of precision and recall. It is used to measure precision and recall at the same time.
3. Accuracy is the number of instances that were correctly classified out of all the instances in the test data.

VII. CONCLUSION

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used. The proposed study emphasized the phishing technique in the context of classification, where phishing website is considered to involve automatic categorization of websites into a predetermined set of class values based on several features and the class variable. The ML based phishing techniques depend on website functionalities to gather information that can help classify websites for detecting phishing sites. The problem of phishing cannot be eradicated, nonetheless can be reduced by combating it in two ways, improving targeted anti-phishing procedures and techniques and informing the public on how fraudulent phishing websites can be detected and identified. To combat the ever evolving and complexity of phishing attacks and tactics, ML anti-phishing techniques are essential. Authors employed LSTM technique to identify malicious and legitimate websites. A crawler was developed that crawled 7900 URLs from AlexaRank portal and also employed Phishtank dataset to measure the efficiency of the proposed URL detector. The outcome of this study reveals that the proposed method presents superior results rather than the existing deep learning methods. A total of 7900 malicious URLS were detected using the proposed URL detector. It has achieved better accuracy and F1—score with limited amount of time. The future direction of this study is to develop an unsupervised deep learning method to generate insight from a URL. In addition, the study can be extended in order to generate an outcome for a larger network and protect the privacy of an individual.

The findings underscore the critical need for a multi-layered approach to cybersecurity. User education emerges as a cornerstone in this defense strategy, empowering individuals to recognize and avoid phishing attempts. Additionally, the implementation of robust cybersecurity measures, including multi-factor authentication, secure browsing practices, and regular software updates, is essential in fortifying defenses against these threats.Advanced detection algorithms, particularly those leveraging artificial intelligence and machine learning, have shown promise in identifying and neutralizing phishing websites with greater accuracy and speed. These technologies can analyze vast amounts of data to detect patterns and anomalies indicative of phishing activities, thereby providing a proactive defense mechanism. Despite these advancements, the dynamic and evolving nature of phishing tactics necessitates continuous research and development. Future efforts should focus on enhancing detection methods, improving user awareness programs, and fostering collaboration between cybersecurity professionals and organizations. By staying ahead of the increasingly complex tactics employed by phishers, we can better safeguard our digital

environments. In conclusion, while phishing websites remain a formidable challenge, a comprehensive and adaptive approach to cybersecurity can significantly mitigate the risks. Through ongoing education, technological innovation, and collaborative efforts, we can build a more resilient defense against the ever-present threat of phishing

VII. REFERENCES

[1] Whitten, A., & Tygar, J. D. (1999). "Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In Proceedings of the 8th conference on USENIX Security Symposium-Volume 8(pp. 169-184)".

[2] Jakobsson, M., & Myers, S. (2007). "Phishing and countermeasures: understanding the increasing problem of electronic identity theft. Wiley Publishing".

[3] Kumaraguru, P., & Cranor, L. F. (2008). "Phishing in Indian cyber space. In Proceedings of the 4th annual workshop on Cyber security and information intelligence research (pp. 1-1)"

[4] Wang, X., & Zhang, Y. (2011). "Design and implementation of a phishing website detection system based on visual similarity. In Proceedings of the 2011 international conference on Internet computing and information services (pp. 1-4)".

[5] Wang, W., & Li, J. (2012). "A new phishing website detection method based on visual similarity and URL features. In Proceedings of the 2012 international conference on computer science and electronics engineering (Vol. 3, pp. 518-521)".

[6] Choo, K. K. R., & Smith, R. G. (2010). "Phishing for phools: An examination of the cyberspace deception techniques and their effectiveness. Journal of Financial Crime, 17(3),273-286".

[7] Dhamija, R., Tygar, J. D., & Hearst, M. (2006). "Why phishing works. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 581-590)".

[8] Sheng, S., Holbrook, M., Kumaraguru, P., & Cranor, L. F. (2010). "Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 373-382)".

[9] Blythe, J., & Wright, P. (2006)."Phishing and the online banking customer. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 1117-1122)".

[10] Kumar, S., Kumar, D., & Lal, S. (2021). "Phishing Websites Detection Using Machine Learning Techniques: A Review. International Journal of Advanced Trends in Computer Science and Engineering, 10(2), 6-12 "

[11] Rezgui, A., Mosbah, M., & Braham, R. (2020)." Phishing websites detection based on textual and visual features using machine learning techniques. Journal of Information Security and Applications, 52, 102507".

[12] AlShurideh, M., & Alkhayat, M. (2020)." Phishing websites detection using machine learning techniques. Journal of Physics: Conference Series, 1654(1), 012020".

[13] Wang, W., & Li, J. (2019). "Phishing website detection based on HTML feature analysis and machine learning algorithms. Security and Communication Networks, 2019, 1-12".

[14] Dey, S., & Guha, S. (2019). "A novel approach for detection and classification of phishing websites using machine learning techniques. Procedia Computer Science, 157, 576-585".

[15] Patel, S., Kotecha, K., & Patel, A. (2018). "Classification and detection of phishing websites using machine learning techniques. International Journal of Computer Science and Information Technologies, 9(6), 5096-5100".

[16] Kumar, R., Pateriya, R., & Tiwari, R. (2018). "Detection of phishing websites using machine learning techniques. International Journal of Computer Sciences and [27] Engineering,6(6), 239-243".

[17] Ren, S., Chen, X., Guo, B., & Zhang, Z. (2017)." Phishing website detection using machine learning techniques".

[18] The set of phishing URLs are collected from opensource service called PhishTank. This service provide a set of phishing URLs in multiple formats like csv, json etc. that gets updated hourly. To download the data: https://www.phishtank.com/developer_info.php . From this dataset, 5000 random phishing URLs are collected to train the ML models.

[19] The legitimate URLs are obatined from the open datasets of the University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html . This dataset has a collection of benign, spam, phishing, malware & defacement URLs. Out of

all these types, the benign url dataset is considered for this project. From this dataset, 5000 random legitimate URLs are collected to train the ML models.

[20] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "An Analytical Perspective on Various Deep Learning Techniques for Deepfake Detection", *1st International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA),* 10th & 11th June 2022, 2456-3463, Volume 7, PP. 25-30, https://doi.org/10.46335/IJIES.2022.7.8.5

[21] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "Revealing and Classification of Deepfakes Videos Images using a Customize Convolution Neural Network Model", *International Conference on Machine Learning and Data Engineering (ICMLDE)*, 7th & 8th September 2022, 2636-2652, Volume 218, PP. 2636-2652, https://doi.org/10.1016/j.procs.2023.01.237

[22] Usha Kosarkar, Gopal Sakarkar (2023), "Unmasking Deep Fakes: Advancements, Challenges, and Ethical Considerations", *4th International Conference on Electrical and Electronics Engineering (ICEEE)*,19th & 20th August 2023, 978-981-99-8661-3, Volume 1115, PP. 249-262, https://doi.org/10.1007/978-981-99-8661-3_19

[23] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2021), "Deepfakes, a threat to society", *International Journal of Scientific Research in Science and Technology (IJSRST)*, 13th October 2021, 2395-602X, Volume 9, Issue 6, PP. 1132-1140, https://ijsrst.com/IJSRST219682

[24] Usha Kosarkar, Prachi Sasankar(2021), "A study for Face Recognition using techniques PCA and KNN", Journal of Computer Engineering (IOSR-JCE), 2278-0661,PP 2-5,

[25] Usha Kosarkar, Gopal Sakarkar (2024), "Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis", Journal of Multimedia Tools and Applications, 1380-7501, https://doi.org/10.1007/s11042-024-19220-w

[26] Usha Kosarkar, Dipali Bhende, "Employing Artificial Intelligence Techniques in Mental Health Diagnostic Expert System", International Journal of Computer Engineering (IOSR-JCE),2278-0661, PP-40-45, https://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume%202/9.%2040-45.pdf?id=7557