International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

Is It Phishing or Not? A Survey on Phishing Website Detection

Dr. J. Thilagavathi , Mrs. D. HemaMalini , Mrs. S. Chandra Priyadharshini , Mr. N. A. Bhaskaran

Professor [1], Assistant Professor [234]

thilagavathi@actechnology.in , hemamalini.d@actechnology.in ,

chandrapriyadharshini.s@actechnology.in , nabhaskaran@actechnology.in

Department of AI&DS, Arjun College of Technology, Thamaraikulam, Coimbatore-Pollachi Highway, Coimbatore, Tamilnadu-642 120

Abstract: The vital requirement for robust detection and prevention measures is highlighted by the fact that phishing is an ongoing and highly successful security threat that presents considerable dangers to both people and targeted brands. An Extensive Analysis of Detection Techniques: The present status of phishing website detection is going to be thoroughly investigated and evaluated as part of this study. A thorough comprehension of current detection technologies and their efficacy is the goal of this work. Methods for detection that fall under the categories of list-based, similarity-based, and machine learning-based techniques will be closely examined in this research. Additionally, it explores the datasets that were used to assess these strategies, offering valuable insights into their advantages and disadvantages.In order to improve the effectiveness of detection methods, this project identifies areas that need more study and development in the domain of phishing website detection, thereby filling in current research gaps.With the use of a Voting Classifier (MLP+XGB+Decision tree classifier), this research aims to improve the identification of phishing websites. To guarantee practical usability in cybersecurity apps, a user-friendly Flask framework with SQLite integration makes registration and signin for user testing a breeze.

Index terms -*Phishing, security threat, phishing website, phishing detection, URL, blacklists, machine learning, page similarity, datasets, social engineering.*

1. INTRODUCTION

Phishing is a dangerous security threat that exploits sophisticated psychological and social engineering techniques to trick individuals into clicking links of malicious websites and submit highly valuable

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

sensitive information, such as personal or corporate information and account credentials. Phishing attacks [7, 8, 10, 14, 24] are far from being technologically complex and their deployment requires little effort. Nevertheless, they are generally very effective. Attackers create wellcrafted phishing websites with a look and feel of the legitimate sites they are trying to impersonate, thus making it very challenging for individuals to identify phishing sites. In addition, to avoid being detected, attackers have refined over the years their tactics and evasion techniques, as demonstrated in [1].

Phishing attacks have several direct and indirect impacts. They affect the individuals being phished, whose identity and accounts might be compromised, thus leading to money being stolen as well as to a potential crisis of trust towardsonline services. These attacks also affect the companies and organizations being impersonated, whose brands might be abused, thus leading to potential data breaches, financial losses and reputation damages.

A study by Enisa [2] reveals that phishing attacks are among the most common cyber incidents European smallmedium enterprises are likely to be exposed to. In the Cybersecurity threat trends report [3] Cisco suggests that in 2020 phishing accounts for around 90% of data breaches. Moreover, 86% of organizations had at least one user try to connect to a phishing site. In fact, as discussed in [4], individuals tend to fall prey of phishing attacks especially because of the insufficient attention paid in assessing the legitimacy of a website and the lack of appropriate education. According to the Phishing activity trends report [5] by Anti-Phishing Working Group (APWG), the total number of phishing websites observed in the first quarter of 2022 exceeds one million.

In the years, the detection of phishing websites has been widely investigated and a large body of the literature has addressed this challenging problem. Our survey aims at providing a broad and comprehensive review of the state of the art in the area of phishing website detection by focusing on the most relevant solutions proposed in the literature.

To gain the trust of the individuals, attackers make the link and website appear legitimate using various tricks, such as typosquatting and combosquatting techniques. For example, they craft the patterns of the Uniform Resource Locator (URL) – shown in address bar of the browser – by inserting unnecessary punctuation marks (e.g., dash), misspelled words (e.g., paymet) or specific words (e.g., brand name being targeted) in incorrect positions. Sometimes, attackers replace English characters with identical looking characters from different alphabets. In fact, although malicious sites might be hosted on compromised servers, attackers might choose to register specific domains with appropriately crafted names. Moreover, attackers tend not to use phishing URLs multipletimes due to the low cost of generating

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

new ones, thus making the detection of phishing websites even more challenging. Let us recall that a URL [11, 12] is a human-readable string of characters – parsed by client programs in a standard way – uniquely identifying a resource on the web [6].

2. LITERATURE SURVEY

Phishing is a critical threat to Internet users. Although an extensive ecosystem serves to protect users, phishing websites are growing in sophistication, and they can slip past the ecosystem's detection systems—and subsequently cause real-world damage—with the help of evasion techniques. Sophisticated client-side evasion techniques, known as cloaking, leverage JavaScript to enable complex interactions between potential victims and the phishing website, and can thus be particularly effective in slowing or entirely preventing automated mitigations. Yet, neither the prevalence nor the impact of client-side cloaking has been studied.In this paper [1], we present CrawlPhish, a framework for automatically detecting and categorizing client-side cloaking used by known phishing websites. We deploy CrawlPhish over 14 months between 2018 and 2019 to collect and thoroughly analyze a dataset of 112,005 phishing websites in the wild. By adapting state-of-the-art static and dynamic code analysis, we find that 35,067 of these websites have 1,128 distinct implementations of client-side cloaking techniques. Moreover, we find that attackers' use of cloaking grew from 23.32% initially to 33.70% by the end of our data collection period. Detection of cloaking by our framework exhibited low false-positive and false-negative rates of 1.45% and 1.75%, respectively. We analyze the semantics of the techniques we detected and propose a taxonomy of eight types of evasion across three high-level categories: User Interaction, Fingerprinting, and Bot Behavior.Using 150 artificial phishing websites [30, 36, 38], we empirically show that each category of evasion technique is effective in avoiding browser-based phishing detection (a key ecosystem defense). Additionally, through a user study, we verify that the techniques generally do not discourage victim visits. Therefore, we propose ways in which our methodology can be used to not only improve the ecosystem's ability to mitigate phishing websites with client-side cloaking, but also continuously identify emerging cloaking techniques as they are launched by attackers.

Phishing was a threat in the cyber world a couple of decades ago and still is today. It has grown and evolved over the years as phishers are getting creative in planning and executing the attacks. Thus, there is a need for a review of the past and current phishing approaches. A systematic, comprehensive and easy-to-follow review of these approaches is presented here. The relevant mediums and vectors of these approaches are identified for each approach. The medium is the platform which the approaches reside and the vector is the means of propagation utilised by the phisher to deploy the attack. The paper [7] focuses primarily on the detailed discussion of these

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

approaches. The combination of these approaches that the phishers utilised in conducting their phishing attacks is also discussed. This review will give a better understanding of the characteristics of the existing phishing techniques which then acts as a stepping stone to the development of a holistic anti-phishing system. This review creates awareness of these phishing techniques and encourages the practice of phishing prevention among the readers. Furthermore, this review will gear the research direction through the types of phishing, while also allowing the identification of areas where the anti-phishing effort is lacking. This review will benefit not only the developers of anti-phishing techniques but the policy makers as well.

Internet technology is so pervasive today, for example, from online social networking to online banking, it has made people's lives more comfortable. Due the growth of Internet technology, security threats to systems and networks are relentlessly inventive. One such a serious threat is "phishing", [44, 45, 46, 47] in which, attackers attempt to steal the user's credentials using fake emails or websites or both. It is true that both industry and academia are working hard to develop solutions to combat against phishing threats. It is therefore very important that organisations to pay attention to end-user awareness in phishing threat prevention. Therefore, aim of our paper is twofold [8]. First, we will discuss the history of phishing attacks and the attackers' motivation in details. Then, we will provide taxonomy of various types of phishing attacks. Second, we will provide taxonomy of various solutions proposed in literature to protect users from phishing based on the attacks identified in our taxonomy. We conclude our paper discussing various issues and challenges that still exist in the literature, which are important to fight against with phishing threats.

In the era of electronic and mobile commerce, massive numbers of financial transactions are conducted online on daily basis, which created potential fraudulent opportunities. A common fraudulent activity that involves creating a replica of a trustful website to deceive users and illegally obtain their credentials is website phishing. Website phishing is a serious online fraud, costing banks, online users, governments, and other organisations severe financial damages. One conventional approach to combat phishing is to raise awareness and educate novice users on the different tactics utilized by phishers by conducting periodic training or workshops. However, this approach has been criticised of being not cost effective as phishing tactics are constantly changing besides it may require high operational cost. Another anti-phishing approach is to legislate or amend existing cyber security laws that persecute online fraudsters without minimising its severity. A more promising anti-phishing approach is to prevent phishing attacks using intelligent machine learning (ML) technology. Using this technology, a classification system is integrated

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

in the browser in which it will detect phishing activities and communicate these with the end user. This paper [9] reviews and critically analyses legal, training, educational and intelligent anti-phishing approaches. More importantly, ways to combat phishing by intelligent and conventional are highlighted, besides revealing these approaches differences, similarities and positive and negative aspects from the user and performance prospective. Different stakeholders such as computer security experts, researchers in web security as well as business owners may likely benefit from this review on website phishing.

Malicious URL, a.k.a. malicious website, is a common and serious threat to cybersecurity. Malicious URLs host unsolicited content (spam, phishing, drive-by exploits, etc.) and lure unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), and cause losses of billions of dollars every year. It is imperative to detect and act on such threats in a timely manner. Traditionally, this detection is done mostly through the usage of blacklists. However, blacklists cannot be exhaustive, and lack the ability to detect newly generated malicious URLs. To improve the generality of malicious URL detectors, machine learning techniques have been explored with increasing attention in recent years. This article [11] aims to provide a comprehensive survey and a structural understanding of Malicious URL Detection techniques using machine learning. We present the formal formulation of Malicious URL Detection as a machine learning task, and categorize and review the contributions of literature studies that addresses different dimensions of this problem (feature representation, algorithm design, etc.). Further, this article provides a timely and comprehensive survey for a range of different audiences, not only for machine learning researchers and engineers in academia, but also for professionals and practitioners in cybersecurity industry [9, 45], to help them understand the state of the art and facilitate their own research and practical applications. We also discuss practical issues in system design, open research challenges, and point out some important directions for future research.

3. METHODOLOGY

i) Proposed Work:

The project comprehensively reviews phishing detection, emphasizing list-based, similarity-based, and machine learning approaches, discussing methods, datasets, challenges, and emphasizing the significance of textual properties and human factors. It advocates integrating AI and machine learning [11] to bolster detection, promoting collaboration for knowledge exchange, and highlighting the essential role of education and awareness in preventing phishing effectively.The project extends its capabilities with a sophisticated Voting Classifier, combining MLP, XGBoost, and Decision Tree

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

Classifier, enhancing the accuracy of phishing website detection. This ensemble method demonstrates improved performance by leveraging the strengths of diverse classifiers. To ensure practical usability, the project integrates a user-friendly Flask framework with SQLite, streamlining signup and signin processes for user testing in the realm of cybersecurity applications [9, 45]. This combination of advanced detection techniques and a seamless user interface enhances the overall effectiveness of phishing website detection.

ii) System Architecture:

The system architecture for detecting suspicious web pages typically involves a multi-layered approach that incorporates various detection methods, such as blacklisting, whitelisting, textual analysis, visual similarity comparison, and machine learning [11]. The goal is to accurately classify web pages as either legitimate or phishing. The architecture begins with the input of web pages that need to be evaluated for potential phishing.

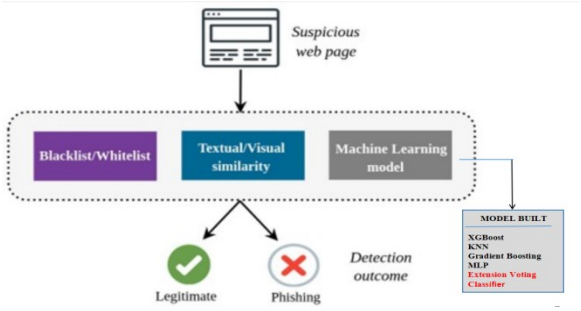


Fig 1 Proposed architecture

iii) Dataset collection:

Load the Phishtank and Curlie datasets into your project. These datasets likely contain information about URLs [11, 12], domains, and potentially other features related to phishing detection.As already discussed, blacklists and whitelists are populated using different approaches that take into account the behaviors of attackers as well as of the individuals. For the evaluation of these approaches, collections of phishing and legitimate websites taken from various sources are considered. For example, popular sources of malicious URLs are represented by PhishTank [35] – a community based phishing website reporting and verification system – and by the Safe Browsing lists provided by Google. Similarly, Alexa – a service providing top-ranked domains retired in May 2022 – and DMOZ – an open directory of the web discontinued in 2017 and now replaced by Curlie [36] – used to be the sources of benign URLs.

	phish_id	url	phish_detail_url
0	8265749	http://www.paxful-terms.online.sti-int.top	http://www.phishtank.com/phish_detail.php?phis...
1	8265747	https://secondary.obec.go.th/mathayom/evaluati...	http://www.phishtank.com/phish_detail.php?phis...
2	8265746	https://f5lrvwt.r.us-east-1.awstrack.me/L0/htt...	http://www.phishtank.com/phish_detail.php?phis...
3	8265744	https://swisspasshilfeservice.sviluppo.host/lo...	http://www.phishtank.com/phish_detail.php?phis...
4	8265742	https://anibis.ecommepce.online/fr/851984	http://www.phishtank.com/phish_detail.php?phis...

Fig 2 NSL KDD dataset

iv) Data Processing:

Data processing involves transforming raw data into valuable information for businesses. Generally, data

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

scientists process data, which includes collecting, organizing, cleaning, verifying, analyzing, and converting it into readable formats such as graphs or documents. Data processing can be done using three methods i.e., manual, mechanical, and electronic. The aim is to increase the value of information and facilitate decision-making. This enables businesses to improve their operations and make timely strategic decisions. Automated data processing solutions, such as computer software programming, play a significant role in this. It can help turn large amounts of data, including big data, into meaningful insights for quality management and decision-making.

v) Feature selection:

Feature selection is the process of isolating the most consistent, non-redundant, and relevant features to use in model construction. Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow. The main goal of feature selection is to improve the performance of a predictive model and reduce the computational cost of modeling [39].

Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model. The main benefits of performing feature selection in advance, rather than letting the machine learning model figure out which features are most important.

vi) Algorithms:

Support Vector Machine (SVM) - SVM is a powerful supervised learning algorithm that is used for classification tasks. It finds an optimal hyperplane in a high-dimensional space to separate data into different classes. In phishing detection, SVM can effectively classify websites into phishing or legitimate based on features extracted from the URLs or webpage content.

```
from sklearn.svm import SVC

svm = SVC()

svm.fit(X_train, y_train)
```

Fig 3 SVC

Random Forest -Random Forest is an ensemble learning algorithm that operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) from individual trees. It's effective for phishing detection by aggregating predictions from multiple decision trees to determine if a website is phishing or not based on various features.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(max_depth=2, random_state=0)

rf.fit(X_train, y_train)
```

Fig 4 Random forest

Logistic Regression -Despite its name, logistic regression is used for binary classification tasks. It predicts the probability that a given input point belongs to a certain class. In phishing detection, logistic regression models can be trained to determine the likelihood of a website being a phishing site based on specific features.

```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression(random_state=0)

lr.fit(X_train, y_train)
```

Fig 5 Logistic regression

K-Nearest Neighbors (KNN) -KNN is a simple and intuitive algorithm used for classification. It classifies a data point based on the majority class of its k-nearest neighbors in the feature space. In phishing detection, KNN can assess the similarity of a website to known phishing or legitimate websites based on extracted features.

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=3)

knn.fit(X_train, y_train)
```

Fig 6 KNN

Decision Tree -A decision tree is a flowchart-like structure where each internal node represents a feature, each branch represents a decision based on that feature, and each leaf node represents a class label. Decision trees can be trained to classify websites into phishing or legitimate based on features such as URL structure or content.

```
from sklearn.tree import DecisionTreeClassifier


dt = DecisionTreeClassifier(
    criterion='gini',
    max_depth=10,
    min_samples_split=5,
    min_samples_leaf=2,
    max_features='sqrt',
    random_state=42,
    class_weight=None,
    splitter='best',
    min_impurity_decrease=0,
    ccp_alpha=0.0
)

# Fit the DecisionTreeClassifier to your training data
dt.fit(X_train, y_train)
```

Fig 7 Decision tree

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

Adaboost - Adaboost (Adaptive Boosting) is an ensemble learning method that constructs a strong classifier by combining multiple weak classifiers. It focuses more on the difficult-to-classify instances, making it suitable for phishing detection by enhancing the classification of suspicious websites.

```
from sklearn.ensemble import AdaBoostClassifier

ada = AdaBoostClassifier()

ada.fit(X_train, y_train)
```

Fig 8 Adaboost

Naive Bayes -Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with the assumption of independence between features. It is efficient and commonly used in text classification tasks. In phishing detection, Naive Bayes can predict the likelihood of a website being phishing or not based on extracted features.

```
from sklearn.naive_bayes import GaussianNB

gnb = GaussianNB()

gnb.fit(X_train, y_train)
```

Fig 9 Naïve bayes

Gradient Boosting -Gradient Boosting is an ensemble learning technique that builds a strong model by sequentially adding weak models. It aims to minimize the loss function. For phishing detection, gradient boosting can be employed to improve classification accuracy by combining weak learners.

```
from sklearn.ensemble import GradientBoostingClassifier

gb = GradientBoostingClassifier()

gb.fit(X_train, y_train)
```

Fig 10 Gradient boosting

XGBoost -XGBoost (Extreme Gradient Boosting) is an efficient and scalable implementation of gradient boosting. It's known for its speed and performance in machine learning competitions. XGBoost can enhance phishing detection accuracy by building an ensemble of weak models.

```
import xgboost as xgb

xg = xgb.XGBClassifier(objective="binary:logistic"

xg.fit(X_train, y_train)
```

Fig 11 XGboost

Convolutional Neural Network (CNN) -CNN is a deep learning model commonly used for image and

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

pattern recognition. In phishing detection, CNN can analyze website content or visual elements to identify patterns associated with phishing.

```
model = Sequential()
model.add(Conv1D(32, 3, padding="same",input_shape = (X_train.shape[1], 1)
model.add(MaxPool1D(pool_size=(4)))
model.add(Dropout(0.2))
model.add(Conv1D(32, 3, padding="same", activation='relu'))
model.add(MaxPool1D(pool_size=(4)))
model.add(Dropout(0.2))
model.add(Flatten())
model.add(Dense(units=50))
model.add(Dense(units=1,activation='softmax'))
```

Fig 12 CNN

Long Short-Term Memory (LSTM) -LSTM is a type of recurrent neural network (RNN) well-suited for sequential data analysis. In phishing detection, LSTM can be used to analyze the sequential nature of URL or webpage content to make predictions.

```
model = Sequential()
model.add(LSTM(64,return_sequences=True,input_shape = (1, X_train.shape[2])))
model.add(Dropout(0.2))
model.add(LSTM(64,return_sequences=True))
model.add(Dropout(0.2))
model.add(LSTM(64,return_sequences=True))
model.add(Flatten())
model.add(Dense(units=50))

model.add(Dense(units=2,activation='softmax'))
```

Fig 13 LSTM

Deep Neural Network (DNN) -DNN is a neural network with multiple layers between the input and output layers. It can capture complex patterns and features from the input data. In phishing detection, DNN can be used to process various features for effective classification.

```
model = Sequential()
model.add(SimpleRNN(64,return_sequences=True,input_shape = (1, X_train.shape[2])))
model.add(Dropout(0.2))
model.add(SimpleRNN(64,return_sequences=True))
model.add(Dropout(0.2))
model.add(SimpleRNN(64,return_sequences=True))
model.add(Flatten())
model.add(Dense(units=50))
model.add(Dense(units=2,activation='softmax'))
```

Fig 14 DNN

Multi-Layer Perceptron (MLP) -MLP is a class of feedforward artificial neural network. It consists of multiple layers of interconnected nodes, allowing for complex learning. In phishing detection, MLP can learn and classify based on extracted features.

```
from sklearn.neural_network import MLPClassifier

mlp = MLPClassifier()

mlp.fit(X_train, y_train)
```

Fig 15 MLP

Perceptron -A perceptron is a simple type of artificial neural network and serves as the foundation for more complex models. It's a linear binary classification model that can be used for basic phishing detection tasks.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

```
from sklearn.linear_model import Perceptron

pre = Perceptron(tol=1e-3, random_state=0)

pre.fit(X_train, y_train)
```

Fig 16 Perception

Passive Aggressive -The Passive Aggressive algorithms are a family of online learning algorithms, often used for classification tasks. They work well in scenarios where data streams in and models need to adapt to changing patterns, potentially useful in evolving phishing detection scenarios.

```
from sklearn.linear_model import PassiveAggressiveClassifier

passive = PassiveAggressiveClassifier()

passive.fit(X_train, y_train)
```

Fig 17 Passive aggressive

Voting Classifier -The Voting Classifier is an ensemble method that combines the predictions of multiple base estimators (e.g., different models) and predicts the class label by majority voting. It can help enhance the overall classification accuracy in phishing detection by leveraging diverse models.

```
from sklearn.ensemble import VotingClassifier


clf1 = MLPClassifier()
clf2 = xgb.XGBClassifier()
clf3 = DecisionTreeClassifier()

eclf11 = VotingClassifier(estimators=[('mlp', clf1), ('xg', clf2), ('dt', clf3)]


eclf11.fit(X_train, y_train)
```

Fig 18 Voting classifier

4. EXPERIMENTAL RESULTS

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

$$Recall = \frac{TP}{TP + FN}$$

Accuracy: Accuracy is the proportion of correct predictions in a classification task, measuring the overall correctness of a model's predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives, making it suitable for imbalanced datasets.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

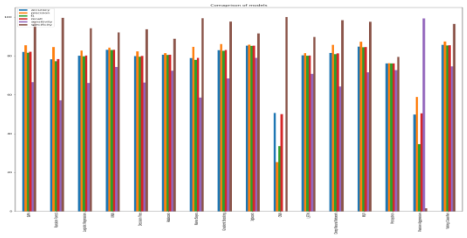


Fig 19 Performance graph

Algorithm name	accuracy	precision	f1	recall
SVM	82.05	85.50891	81.61685	82.09679
Random Forest	78.3	84.54446	77.29888	78.36371
Logistic Regression	80.15	82.76042	79.76179	80.19227
KNN	83.15	84.23692	83.02297	83.1767
Decision Tree	79.95	82.38691	79.58061	79.99107
Adaboost	80.6	81.46032	80.47504	80.62473
Naive Bayes	78.9	84.71087	78.00241	78.96131
Gradient Boosting	83	86.10069	82.64002	83.04395
Xgboost	85.25	85.81936	85.19581	85.26892
CNN	50.6	25.3	33.59894	50
LSTM	80.35	81.45244	80.13305	80.23615
Deep Neural Network	81.55	85.68736	80.92865	81.34511
MLP	84.75	87.2867	84.4429	84.59358
Perceptron	76.15	76.25267	76.10478	76.10876
Passive Aggressive	49.8	58.88869	34.52553	50.38686
Extension Voting Classifier	**85.65**	**87.44321**	**85.44211**	**85.51892**

Fig 20 Performance Evaluation

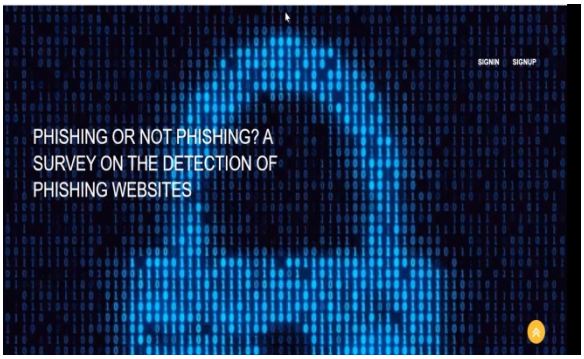
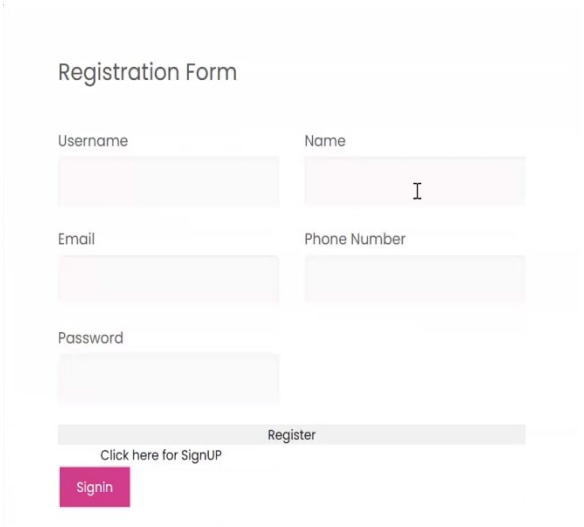


Fig 21 Home page



International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

Fig 22 Signin page

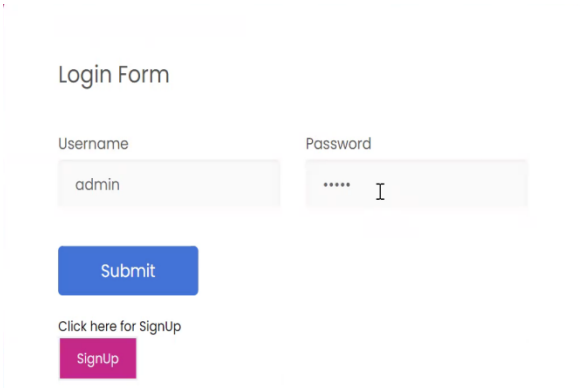


Fig 23 Login page

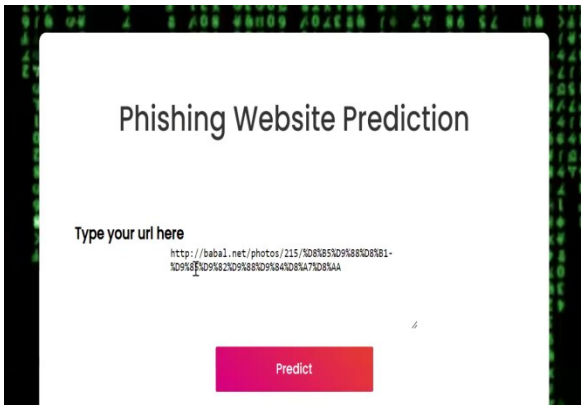


Fig 24User input

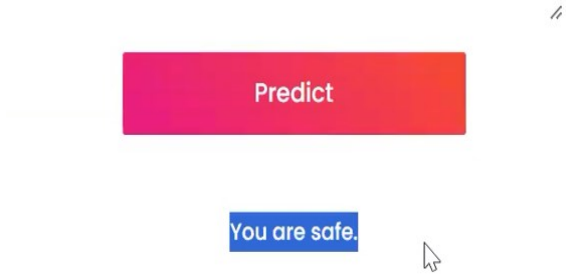


Fig 25 Predict result for given input

5. CONCLUSION

The project thoroughly examines the current advancements and methodologies in detecting phishing websites. It aims to provide an in-depth understanding of the existing landscape of phishing detection. The project covers a wide array of detection approaches, shedding light on the diversity in techniques used to identify phishing attempts [40, 41, 42]. Additionally, it discusses the datasets utilized for evaluation, providing insights into the empirical foundations of the field. Furthermore, it addresses the research gaps in phishing detection that necessitate further exploration. The project underscores the significance of feature selection, emphasizing the need to carefully choose features based on their individual strengths and weaknesses. It advocates for features that possess high discriminating power and align with attacker strategies. The Voting Classifier, an extension of the project, achieves an accuracy of 85.65%, showcasing superior performance and robustness in the detection of phishing websites. The ensemble technique combines multiple classifiers, enhancing the overall effectiveness of the system in identifying various phishing tactics and improving generalization across diverse scenarios. Integration of a user-friendly Flask interface with secure authentication enhances the overall user experience during system testing. This ensures a practical and accessible environment for inputting data and

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

evaluating the performance of the phishing detection system. The Flask interface simplifies the testing process and contributes to a seamless user interaction, making the evaluation of the system's effectiveness more efficient and user-friendly.

6. FUTURE SCOPE

The project advocates for dedicated research efforts to delve into the challenges and gaps within phishing detection. This involves a comprehensive analysis of existing limitations to drive the development of more effective countermeasures. Highlighting the role of individuals as a vulnerable link, the project emphasizes the integration of education into phishing countermeasures. Educating users to identify and respond to phishing attempts effectively is considered a fundamental approach to enhance prevention. The project underscores the need to enhance the accuracy of phishing detection [32, 34, 37] by continually updating and expanding lists of known phishing websites. This continuous development ensures that detection mechanisms stay up-to-date with the evolving landscape of phishing. Recognizing the dynamic nature of phishing tactics, the project advocates for the exploration and development of new techniques and approaches. This proactive approach aims to anticipate and effectively detect emerging phishing tactics and strategies. The project promotes the integration of advanced technologies, such as artificial intelligence and machine learning, to bolster the effectiveness of phishing detection systems. These technologies can offer more sophisticated and adaptive detection capabilities. The project calls for the evaluation and improvement of existing detection methods using larger and more diverse datasets. This approach ensures that the detection methods are thoroughly tested and validated under varied real-world conditions, enhancing their overall effectiveness and applicability.

REFERENCES

[1] P. Zhang, A. Oest, H. Cho, Z. Sun, R. Johnson, B. Wardman, S. Sarker, A. Kapravelos, T. Bao, R. Wang, Y. Shoshitaishvili, A. Doupé, and G.-J. Ahn, ‘‘CrawlPhish: Large-scale analysis of client-side cloaking techniques in phishing,’’ in Proc. IEEE Symp. Secur. Privacy (SP), May 2021, pp. 1109–1124.

[2] ENISA. (2021). Cybersecurity for SMEs—Challenges and Recommendations. [Online]. Available: https://www.enisa.europa.eu/publications/enisa-report-cybersecurity-for-smes

[3] Cisco. (2021). Cyber Security Threat Trends: Phishing, Crypto Top the List. [Online]. Available: https://umbrella.cisco.com/info/2021-cybersecurity-threat-trends-phishing-crypto-top-the-list

[4] M. Alsharnouby, F. Alaca, and S. Chiasson, ‘‘Why phishing still works: User strategies for combating phishing attacks,’’ Int. J. Hum.-Comput. Stud., vol. 82, pp. 69–82, Oct. 2015.

[5] Anti-Phishing Working Group—APWG. (2022). Phishing Activity Trends Report-1Q. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q1_2022.pdf

[6] T. Berners-Lee, R. Fielding, and L. Masinter. Uniform Resource Identifier (URI): Generic Syntax, RFC 3986, Jan. 2005. [Online]. Available: https://www.rfc-editor.org/rfc/rfc3986.txt

[7] K. L. Chiew, K. S. C. Yong, and C. L. Tan, ''A survey of phishing attacks: Their types, vectors and technical approaches,'' Exp. Syst. Appl., vol. 106, pp. 1–20, Sep. 2018.

[8] B. B. Gupta, N. A. G. Arachchilage, and K. E. Psannis, ''Defending against phishing attacks: Taxonomy of methods, current issues and future directions,'' Telecommun. Syst., vol. 67, no. 2, pp. 247–267, 2018.

[9] I. Qabajeh, F. Thabtah, and F. Chiclana, ''A recent review of conventional vs. automated cybersecurity anti-phishing techniques,'' Comput. Sci. Rev., vol. 29, pp. 44–55, Aug. 2018.

[10] E. Benavides, W. Fuertes, S. Sanchez, and M. Sanchez, ''Classification of phishing attack solutions by employing deep learning techniques: A systematic literature review,'' in Developments and Advances in Defense and Security (Smart Innovation, Systems and Technologies), vol. 152, A. Rocha and R. P. Pereira, Eds. Berlin, Germany: Springer, 2020, pp. 51–64.

[11] D. Sahoo, C. Liu, and S. C. H. Hoi, ''Malicious URL detection using machine learning: A survey,'' 2017, arXiv:1701.07179.

[12] C. M. R. Da Silva, E. L. Feitosa, and V. C. Garcia, ''Heuristic-based strategy for phishing prediction: A survey of URL-based approach,'' Comput. Secur., vol. 88, Jan. 2020, Art. no. 101613.

[13] G. Varshney, M. Misra, and P. K. Atrey, ''A survey and classification of web phishing detection schemes,'' Secur. Commun. Netw., vol. 9, no. 18, pp. 6266–6284, Dec. 2016.

[14] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, ''A comprehensive survey of AI-enabled phishing attacks detection techniques,'' Telecommun. Syst., vol. 76, no. 1, pp. 139–154, Jan. 2021.

[15] A. Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar, ''SoK: A comprehensive reexamination of phishing research from the security perspective,'' IEEE Commun. Surveys Tuts., vol. 22, no. 1, pp. 671–708, 1st Quart., 2020.

[16] Z. Dou, I. Khalil, A. Khreishah, A. Al-Fuqaha, and M. Guizani, ''Systematization of knowledge (SoK): A systematic review of software-based web

phishing detection,'' IEEE Commun. Surveys Tuts., vol. 19, no. 4, pp. 2797–2819, 4th Quart., 2017.

[17] A. K. Jain and B. B. Gupta, ''Phishing detection: Analysis of visual similarity based approaches,'' Secur. Commun. Netw., vol. 2017, pp. 1–20, Jan. 2017.

[18] M. Khonji, Y. Iraqi, and A. Jones, ''Phishing detection: A literature survey,'' IEEE Commun. Surveys Tuts., vol. 15, no. 4, pp. 2091–2121, 4th Quart, 2013. [19] Google Safe Browsing. Accessed: Oct. 10, 2022. [Online]. Available: https://safebrowsing.google.com/

[20] S. Bell and P. Komisarczuk, ''An analysis of phishing blacklists: Google Safe Browsing, OpenPhish, and PhishTank,'' in Proc. Australas. Comput. Sci. Week Multiconf., Feb. 2020, pp. 1–11.

[21] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, ''An empirical analysis of phishing blacklists,'' in Proc. 6th Conf. Email AntiSpam (CEAS), 2009, pp. 1–10.

[22] A. Oest, Y. Safaei, A. Doupé, G.-J. Ahn, B. Wardman, and K. Tyers, ''PhishFarm: A scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists,'' in Proc. IEEE Symp. Secur. Privacy (SP), May 2019, pp. 1344–1361.

[23] A. Oest, Y. Safaei, P. Zhang, B. Wardman, K. Tyers, Y. Shoshitaishvili, and A. Doupé, ''PhishTime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists,'' in Proc. 29th USENIX Secur. Symp., 2020, pp. 379–396.

[24] N. A. Azeez, S. Misra, I. A. Margaret, L. Fernandez-Sanz, and S. M. Abdulhamid, ''Adopting automated whitelist approach for detecting phishing attacks,'' Comput. Secur., vol. 108, Sep. 2021, Art. no. 102328.

[25] Y. Cao, W. Han, and Y. Le, ''Anti-phishing based on automated individual white-list,'' in Proc. 4th ACM Workshop Digit. Identity Manag., Oct. 2008, pp. 51–60.

[26] W. Han, Y. Cao, E. Bertino, and J. Yong, ''Using automated individual white-list to protect web digital identities,'' Exp. Syst. Appl., vol. 39, no. 15, pp. 11861–11869, Nov. 2012.

[27] A. K. Jain and B. B. Gupta, ''A novel approach to protect against phishing attacks at client side using auto-updated white-list,'' EURASIP J. Inf. Secur., vol. 2016, no. 1, pp. 1–11, Dec. 2016.

[28] L.-H. Lee, K.-C. Lee, H.-H. Chen, and Y.-H. Tseng, ''POSTER: Proactive blacklist update for anti-phishing,'' in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Nov. 2014, pp. 1448–1450.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

[29] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, ''PhishNet: Predictive blacklisting to detect phishing attacks,'' in Proc. IEEE INFOCOM, Mar. 2010, pp. 1–5.

[30] R. S. Rao and A. R. Pais, ''An enhanced blacklist method to detect phishing websites,'' in Information Systems Security (Lecture Notes in Computer Science), vol. 10717, R. K. Shyamasundar, V. Singh, and J. Vaidya, Eds. Berlin, Germany: Springer, 2017, pp. 323–333.

[31] M. Sharifi and S. H. Siadati, ''A phishing sites blacklist generator,'' in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., Mar. 2008, pp. 840–843.

[32] C. Whittaker, B. Ryner, and M. Nazif, ''Large-scale automatic classification of phishing pages,'' in Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS), 2010, pp. 1–14.

[33] G. Xiang, B. A. Pendleton, J. Hong, and C. P. Rose, ''A hierarchical adaptive probabilistic approach for zero hour phish detection,'' in Computer Security—ESORICS (Lecture Notes in Computer Science), vol. 6345, D. Gritzalis, B. Preneel, and M. Theoharidou, Eds. Berlin, Germany: Springer, 2010, pp. 268–285.

[34] G. Sonowal and K. S. Kuppusamy, ''PhiDMA—A phishing detection model with multi-filter approach,'' J. King Saud Univ.-Comput. Inf. Sci., vol. 32, no. 1, pp. 99–112, 2020.

[35] PhishTank. Accessed: Nov. 4, 2022. [Online]. Available: https://www.phishtank.org

[36] Curlie. Accessed: Nov. 4, 2022. [Online]. Available: https://curlie.org/

[37] S. Afroz and R. Greenstadt, ''PhishZoo: Detecting phishing websites by looking at them,'' in Proc. IEEE 5th Int. Conf. Semantic Comput., Sep. 2011, pp. 368–375.

[38] J.-L. Chen, Y.-W. Ma, and K.-L. Huang, ''Intelligent visual similaritybased phishing websites detection,'' Symmetry, vol. 12, no. 10, Oct. 2020, Art. no. 1681.

[39] K. T. Chen, J. Y. Chen, C. R. Huang, and C. S. Chen, ''Fighting phishing with discriminative keypoint features,'' IEEE Internet Comput., vol. 13, no. 3, pp. 56–63, May 2009.

[40] T.-C. Chen, S. Dick, and J. Miller, ''Detecting visually similar web pages: Application to phishing detection,'' ACM Trans. Internet Technol., vol. 10, no. 2, pp. 1–38, May 2010.

[41] J. Chen and C. Guo, ''Online detection and prevention of phishing attacks,'' in Proc. 1st Int. Conf. Commun. Netw. China, Oct. 2006, pp. 1–7.

[42] K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, ''Utilisation of website logo for phishing detection,'' Comput. Secur., vol. 54, pp. 16–26, Oct. 2015.

[43] M. Dunlop, S. Groat, and D. Shelly, ‘‘GoldPhish: Using images for content-based phishing analysis,’’ in Proc. 5th Int. Conf. Internet Monitor. Protection, 2010, pp. 123–128.

[44] A. Y. Fu, L. Wenyin, and X. Deng, ‘‘Detecting phishing web pages with visual similarity assessment based on Earth mover’s distance (EMD),’’ IEEE Trans. Dependable Secure Comput., vol. 3, no. 4, pp. 301–311, Oct. 2006.

[45] M. Hara, A. Yamada, and Y. Miyake, ‘‘Visual similarity-based phishing detection without victim site information,’’ in Proc. IEEE Symp. Comput. Intell. Cyber Secur., Mar. 2009, pp. 30–36.

[46] C.-Y. Huang, S.-P. Ma, W.-L. Yeh, C.-Y. Lin, and C.-T. Liu, ‘‘Mitigate web phishing using site signatures,’’ in Proc. TENCON IEEE Region Conf., Nov. 2010, pp. 803–808.

[47] W. Khan, A. Ahmad, A. Qamar, M. Kamran, and M. Altaf, ‘‘SpoofCatch: A client-side protection tool against phishing attacks,’’ IT Prof., vol. 23, no. 2, pp. 65–74, Mar. 2021.

[48] I. F. Lam, W. C. Xiao, S. C. Wang, and K. T. Chen, ‘‘Counteracting phishing page polymorphism: An image layout analysis approach,’’ in Advances in Information Security and Assurance (Lecture Notes in Computer Science), vol. 5576, J. H. Park, H. H. Chen, M. Atiquzzaman, C. Lee, T. Kim, and S. S. Yeo, Eds. Berlin, Germany: Springer, 2009, pp. 270–279.

[49] Y. Lin, R. Liu, D. M. Divakaran, J. Ng, Q. Chan, Y. Lu, Y. Si, F. Zhang, and J. Dong, ‘‘Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages,’’ in Proc. 30th USENIX Secur. Symp., 2021, pp. 3793–3810.

[50] W. Liu, X. Deng, G. Huang, and A. Y. Fu, ‘‘An antiphishing strategy based on visual similarity assessment,’’ IEEE Internet Comput., vol. 10, no. 2, pp. 58–65, Mar. 2006