International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8353
***Article Received:** 14 July 2023 **Revised:** 08 September 2023 **Accepted:** 23 September 2023*

Improving Phishing Website Detection with Machine Learning: Revealing Hidden Patterns for Better Accuracy

Garlapati Narayana[*1], Uma Devi Manchala[2], Usikela Naresh[3], Saggurthi Kiran[4], Medikonda Asha Kiran[5], Ravi Kumar Ch[6]

[1]Associate Professor, Department of CSE (AIML),
Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, Telangana, India
narayanag.1973@gmail.com
ORCID: https://orcid.org/0000-0001-8470-3595

[2]Assistant Professor. Department of Computer Science and Engineering,
Nalla Narsimha Reddy Education society's Group of Institutions, Chowdariguda, Medchal, Telangana, India
umadevi.manchala92@gmail.com
ORCID: https://orcid.org/0000-0002-8325-3868

[3]Assistant Professor. Department of Computer Science and Engineering (AI&ML),
CVR College of Engineering, Mangalpalli, Rangareddy, Telangana, India
usikelanaresh@gmail.com
ORCID: https://orcid.org/0009-0006-9656-4880

[4]Assistant Professor, Department of Computer Science and Engineering (AI&ML),
CMR Technical Campus, kandlakoya, Medchal, Telangana, India
kiransaggurthicfc@gmail.com
ORCID: https://orcid.org/0009-0002-5997-2288

[5]Assistant Professor, Department of AIML,
Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, Telangana, India
ashakiran2@gmail.com
ORCID: https://orcid.org/0000-0002-7760-2902

[6]Assistant Professor, Department of AI&DS,
Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, Telangana, India
chrk5814@gmail.com
ORCID: https://orcid.org/0000-0003-0809-5545

Abstract: Phishing attacks remain a significant threat to internet users globally, leading to substantial financial losses and compromising personal information. This research study investigates various machine learning models for detecting phishing websites, with a primary focus on achieving high accuracy. After an extensive analysis, the Random Forest Classifier emerged as the most suitable choice for this task. Our methodology leveraged machine learning techniques to uncover subtle patterns and relationships in the data, going beyond traditional URL and content-based restrictions. By incorporating diverse website features, including URL and derived attributes, Page source code-based features, HTML JavaScript-based features, and Domain-based features, we achieved impressive results. The proposed approach effectively classified the majority of websites, demonstrating the efficiency of machine learning in addressing the phishing website detection challenge with an accuracy of over 98%, recall exceeding 98%, and a false positive rate of less than 4%. This research offers valuable insights to the field of cyber security, providing internet users with improved protection against phishing attempts.

Keywords: Phishing attacks, accuracy, machine learning model, optimal parameters, Cyber security.

I. INTRODUCTION

The internet has revolutionized the way we conduct business, communicate, and access information. However, this digital transformation has brought about a dark side: cybercrime. Among the numerous cyber threats, phishing attacks have emerged as a primary concern for individuals and organizations alike. Phishes employ social engineering techniques to manipulate human vulnerability, luring unsuspecting victims into revealing sensitive information or performing actions that can have dire consequences [1][2].

Phishing attacks typically involve the distribution of deceptive emails or messages containing fraudulent links. Once recipients fall into the trap, cybercriminals exploit this opportunity to gain unauthorized access to victims' accounts,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8353
***Article Received:** 14 July 2023 **Revised:** 08 September 2023 **Accepted:** 23 September 2023*

leading to financial loss, identity theft, and other severe ramifications. Despite efforts to mitigate this menace, the proliferation of phishing websites and the evolution of sophisticated tactics have made traditional detection methods less effective [3].

The escalating prevalence of phishing attacks poses a significant worry for internet consumers globally, as cybercriminals manipulate email and messaging systems to deceive unsuspecting victims using fraudulent links. Phishing attacks lead to substantial financial losses and the compromise of sensitive information and financial accounts. Conventional approaches to detect phishing websites encounter mounting difficulties due to the rising number of phishing sites and the adoption of sophisticated tactics to evade detection. This literature review examines previous research on machine learning-based methodologies to enhance the identification of phishing websites, aiming to tackle these challenges and protect internet users from the pervasive threat of cybercrime [4].

1.1 Challenges with Traditional Methods

Traditional approaches for detecting phishing websites have long relied on techniques like visual verification, content-based analysis, and maintaining blacklists of known phishing URLs. Although effective in the past, these methods struggle to keep pace with the ever-increasing number of phishing sites and the cunning techniques employed by phishers. Phishers now utilize URL obfuscation to disguise malicious URLs, making them appear genuine to users and security systems. Link redirection further complicates the detection process, as users are directed to fraudulent sites after clicking on seemingly harmless links. Moreover, manipulations to the appearance of URLs create a facade of legitimacy, deceiving even cautious internet users [5] [6].

1.2 The Machine Learning-Based Approach

This research study suggests a machine learning-based strategy to address the drawbacks of conventional approaches and improve phishing detection abilities. Systems are given the ability to learn from data and enhance their performance over time thanks to machine learning, a subfield of artificial intelligence. Using this technology, the suggested methodology seeks to analyze massive datasets of both genuine and phishing URLs to identify patterns and traits specific to phishing websites. In the initial phase, features are extracted from URLs in order to create a format that is appropriate for machine learning algorithms and extract useful properties from those URLs. After that, these variables are fed into different machine learning models, including decision trees, support vector machines, or deep neural networks, to see how well they function to distinguish between phishing and authentic websites[7].

II. LITERATURE REVIEW

The escalating threat of phishing attacks has led to significant financial losses for internet consumers globally. Cybercriminals have honed their tactics, exploiting email and messaging systems to deceive unsuspecting victims with fraudulent links, compromising sensitive information and financial accounts. Traditional methods for detecting phishing websites are facing growing challenges due to the sheer number of phishing sites and the use of sophisticated tactics, such as URL obfuscation, link redirection, and manipulations. To combat these challenges and enhance the accuracy of phishing website identification, researchers have turned to machine learning-based methodologies. This section reviews relevant literature exploring the application of machine learning in phishing detection and its effectiveness in safeguarding internet users against cybercrime [8] [9].

By looking for trends and features in URLs and web content, machine learning approaches have showed promise in identifying phishing websites. In their study, Liu et al. (2011) investigated the use of machine learning techniques for detecting phishing websites, including decision trees, naive Bayes, and support vector machines. They showed the promise of machine learning in phishing attack defense with their study's encouraging accuracy, sensitivity, and specificity results [11].

Due to its capacity to manage intricate patterns and characteristics, deep learning, a subset of machine learning, has drawn attention. A deep learning-based strategy employing convolutional neural networks (CNNs) to identify phishing URLs was recently proposed by Zhang et al. (2019). In recognizing misleading URLs, their model outperformed conventional machine learning techniques and displayed greater performance [12].

Ensemble learning, which combines multiple classifiers, has shown promise in improving phishing detection accuracy. In a comparative study, Akhtar et al. (2018) examined the effectiveness of ensemble learning methods, including bagging and boosting, in phishing detection. Their findings revealed that ensemble approaches achieved higher accuracy and reduced false positive rates compared to individual classifiers [13].

Imbalanced datasets, where phishing instances are significantly outnumbered by legitimate URLs, pose challenges for machine learning models. In response, Chiew et al. (2020) proposed a

novel ensemble learning framework using a synthetic minority oversampling technique to address class imbalance

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8353
***Article Received:** 14 July 2023 **Revised:** 08 September 2023 **Accepted:** 23 September 2023*

in phishing detection. Their approach achieved improved accuracy and effectively mitigated the issue of imbalanced data [14].

To tackle URL obfuscation and evasion techniques employed by phishers, Chen et al. (2019) presented a machine learning-based system that incorporated URL semantic features and network traffic analysis to detect phishing websites. Their hybrid approach achieved enhanced accuracy, demonstrating the importance of considering multiple aspects for robust phishing detection [15].

Machine learning techniques have shown promise in detecting phishing websites by analyzing features and patterns that distinguish malicious URLs from legitimate ones. Li et al. (2017) proposed a machine learning-based system that employs a combination of decision tree and random forest classifiers to achieve high accuracy in identifying phishing websites. The study used a dataset comprising both phishing and legitimate URLs to train the models and reported encouraging results with a precision of 94% and recall of 92% [16].

URL analysis and feature extraction are critical steps in machine learning-based phishing detection. Datta et al. (2019) introduced a feature extraction method based on URL syntax, content, and host information to distinguish phishing URLs from legitimate ones. The researchers employed various machine learning classifiers, including support vector machines and logistic regression, and achieved an accuracy of 96% using their feature extraction approach [17].

In recent years, deep learning models have demonstrated remarkable capabilities in various cybersecurity applications, including phishing detection. Zhang et al. (2020) proposed a deep neural network architecture for detecting phishing URLs based on lexical and semantic features. Their model effectively addressed the challenges of URL obfuscation and link redirection, achievingan accuracy of 98% [18].

While machine learning has proven effective in detecting phishing websites, cybercriminals continue to evolve their tactics to circumvent detection. Adversarial machine learning has emerged as a field dedicated to studying the vulnerability of machine learning models to adversarial attacks. Nainar and Halder (2022) investigated the robustness of machine learning-

based phishing detection models against adversarial attacks and proposed techniques to enhance model resilience [19].

The success of machine learning-based phishing detection models relies on accurate performance evaluation metrics. Ahmad et al. (2018) conducted a comprehensive evaluation of different machine learning models, comparing various metrics such as precision, recall, accuracy, and F1 score. The study emphasized the importance of balancing false positives and false negatives to achieve optimal performance [20].

2.1 Summary Table

Authors	Abstract	Methodology	Findings
Liu et al. (2011)	Studied the use of machine learning algorithms, such as support vector machines, naive bayes, and decision trees, to identify phishing websites.	Employed various machine learning algorithms to analyzepatterns and features from URLs and web content.	Achieved positive results in terms of sensitivity, specificity, and accuracy, highlighting the potential of machine learning in phishing attack defense [10].
Zhang et al. (2019)	Suggested an approach based on deep learning, utilizing Convolutional Neural Networks (CNNs) for the detection of phishing URLs.	Utilized deep learning techniques, particularly CNNs, tohandle complex patterns and features in URLs.	Demonstrated superior performance, achieving high accuracy and outperforming traditional machine learning methods in identifying deceptive URLs [11].
Akhtar et al. (2018)	Examined the effectiveness of ensemble learning methods, including bagging and boosting, in phishing detection.	Implemented ensemble learning techniques, combining multiple classifiers, to improve phishing detection accuracy.	Ensemble approaches achieved higher accuracy and reducedfalse positive rates compared to Individual classifiers[12].
Chiew et al. (2020)	Proposed a novel ensemble learning framework using a synthetic minority oversampling technique to address class imbalance in phishing detection.	Addressed class imbalance issues using an ensemblelearning approach combined with synthetic minority oversampling.	Achieved improved accuracy and effectively mitigated the problem of imbalanced data [13].
Chen et al. (2019)	Presented a machine learning-based system incorporating URL semantic features and network traffic analysis to detect phishing websites.	Utilized a hybrid approach, consideringURL semantics and network traffic analysis, to tackle URL obfuscation and evasion techniques.	Achieved enhanced accuracy by considering multipleaspects for robust phishing detection [14].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8353
***Article Received:** 14 July 2023 **Revised:** 08 September 2023 **Accepted:** 23 September 2023*

Ahmad et al. (2018)	Conducted an extensive analysis of different machine learning models for phishing detection, emphasizing the value of performance evaluation metrics.	Evaluated various machine learning models using metrics such as precision, recall, accuracy, and F1 score.	Highlighted the significance of balancing false positives and false negatives for optimal performance [15].
Datta et al. (2019)	Introduced a feature extraction method based on URL syntax, content, and host information to distinguish phishing URLs from legitimate ones.	Utilized diverse machine learning classifiers, such as support vector machines and logistic regression, for feature extraction and classification purposes.	Achieved an accuracyof 96% using their feature extraction approach [16].
Li et al. (2017)	To achieve high accuracy in phishing website detection, a machine learning-based approach using decision tree and random forest classifiers was proposed.	Used decision tree and random forest classifiers, and trained the model using a dataset made up of both authentic and phishing URLs.	Reported encouraging results with a precision of 94% and recall of 92% in identifying phishing websites [17].
Nainar et.al (2022)	Investigated the robustness of machine learning-based phishing detection models against adversarial attacks and proposed techniques to enhancemodel resilience.	Explored adversarial machine learning methods to study model vulnerability to adversarial attacks.	Discussed techniquesto enhance model resilience against evolving tactics usedby cybercriminals [18].
Wang et al. (2021)	Utilized transfer learning by employing a pre-trained language model and fine-tuning it for the specific phishing detection task.	Utilized transfer learning to apply knowledge from onedomain to improve phishing detection.	Outperformed traditional machine learning models withan accuracy of 99.2%[19].
Zhang et al. (2020)	Proposed a deep neural network architecture for detecting phishing URLs based on lexical and semantic features.	Utilized deep neural networks to address URL obfuscation and link redirection challenges.	Achieved an accuracyof 98% in identifying phishing URLs [20].

Machine learning has become a potent weapon in countering the widespread menace of phishing attacks. Numerous research studies have investigated the use of machine learning algorithms, encompassing both traditional methods and deep learning, for phishing detection. Leveraging ensemble learning techniques and tackling imbalanced datasets has significantly improved the accuracy of detection. By harnessing the potential of machine learning, scholars endeavor to holistically tackle the intricacies linked to phishing attacks, thereby protecting internet users from the ever-changing cybercrime landscape.

III. Problem statement

Machine learning techniques have shown promise in detecting phishing websites through analysis of patterns and features from URLs and web content. However, challenges persist, such as handling imbalanced datasets and tackling URL obfuscation employed by phishers. Researchers have proposed deep learning and ensemble methods to improve accuracy, while adversarial machine learning is explored to enhance model resilience. Evaluating performance metrics is crucial for optimal detection. Further research aims to address these complexities and combat the evolving threat of phishing attacks.

3.1 Contributions

- The study paper contributes to the field of cyber security by exploring the use of machine learning for detecting and preventing phishing attacks.
- The main objective of the study is to identify the most effective machine learning model and parameters to create a reliable and efficient defense against evolving cybercriminal tactics.
- The findings of this research could significantly improve internet security and reduce the financial and personal risks that online users face due to phishing attacks.

IV. DATASET

In our study, we made use of the "Phishing website dataset" accessible on the Kaggle website. This dataset comprises 30 optimized features specifically relevant to phishing websites. These features can be categorized into three distinct groups:

A. URL and derived features:

1. Long URL: Phishing domains are concealed within long URLs to evade detection.
2. IP instead of URL: Phishers use IP addresses instead of recognizable URLs to deceiveusers.
3. Shortened URLs: Phishing URLs are often disguised using URL shorteners, appearing innocuous at first glance.
4. "@" symbol in URL: The phishing portion of the URL can follow the "@" symbol, asweb browsers disregard anything preceding it.
5. URLs with "//": The use of "//" can lead to redirection to a

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8353
***Article Received:** 14 July 2023 **Revised:** 08 September 2023 **Accepted:** 23 September 2023*

phishing site.

6. URLs with "-": Phishing websites mimic legitimate ones by incorporating "-" in theirURLs.
7. Number of subdomains: Phishing sites commonly use multiple subdomains for redirection, unlike legitimate websites that typically have none or only one.
8. Use of HTTPs security: Phishing sites may operate over unprotected HTTP or lack a valid HTTPS certificate, while legitimate sites use HTTPS for security.
9. Domain registration period: Legitimate websites tend to have longer registration periods, whereas phishing websites operate for short durations with domains registered for less than a year.
10. Favicon: Phishing attempts may load favicons from external websites to spoof URL identity.
11. Ports: Only certain ports (80 and 443, respectively) are used by legitimate HTTP and HTTPS websites; other ports should be kept blocked for security purposes.
12. Use of "https" in the domain part: To give users a false sense of security and deceive them into thinking the URL is secure, phishers may use "https" in the domain part.

B. Based on URLs Incorporated in Website:

A webpage's accessibility or the nature of the URLs it links to can provide important information. When connections point to the same website, the credibility of the website is frequently increased. Embedded URLs were used to identify the following details:

1. Embedded Objects' URLs: Trustworthy pages share their domains with the embedded objects they contain. In contrast, phishing websites download embedded files from outside sources to provide the appearance of being from a trustworthy source.
2. Anchor Tag URL: The anchor tag in HTML is used for hyper linking. False sources in anchor tags are never found on trustworthy websites. On the other hand, phishers could utilize bogus sources to divert personal data to different sources.
3. Tags: Trustworthy pages use the same domain name for the page's URL and the tags for the script, link, and meta descriptions. These domain names frequently contain errors on suspicious websites.
4. Server Form Handler (SFH): Trustworthy websites often act upon content sent via a form. The chance of phishing increases if the form handler is empty or is from a different domain than the real website.
5. Email Submission: Reputable websites either process information submitted on the frontend or backend. However, phishers might divert data to their own mail, which raisesred flags.
6. Unusual URL: Normally, every object's URL on a webpage includes the host's name. Anydeparture from this pattern can be a warning sign of a possible danger.

C. Based on HTML and JavaScript Features:

To hide harmful code inside of seemingly innocent websites, HTML and JavaScript are frequently used. Some of the distinguishing characteristics are:

1. The number of website redirects: While phishing sites sometimes have more than four redirects, legitimate websites normally have fewer, usually only one.
2. Modification of the status bar: Phishers frequently use JavaScript to alter the URL that appears in the address bar so that it differs from the URL of the website.
3. Right-Click Disabled: Phishers frequently limit the right-click feature to prevent consumers from seeing the source code of the website, lowering the likelihood that they would be discovered.
4. Pop-Up Windows: Phishing websites commonly take advantage of pop-up windows to gather sensitive data, despite the fact that reputable websites may utilize them to alert users.
5. IFrame Redirection: To hide their objectives, phishers utilize invisible frames to overlap a webpage and send viewers to another website or server.

D. Domain-based Characteristics:

Reputable websites often maintain their domains for lengthy periods of time and display strong statistical characteristics. Phishing websites, on the other hand, are more recent and don't offer any signs that they are legitimate.

1. Age of the Domain: Reputable websites normally have a minimum age of six months, butphishing websites have a short lifespan.
2. DNS Record: Reputable websites typically have non-empty DNS records and are found inpublicly accessible WHOIS databases. Phishing websites, on the other hand, are frequently missed by WHOIS databases.
3. Website traffic: Trustworthy domains draw a lot of visits, ranking them among the top 100,000 in the Alexa database. Websites that Alexa does not recognize are probably phishing scams.
4. Page Rank: A legitimate domain would typically have a Page Rank of between 0.2 and 1, with a higher Page

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8353
***Article Received:** 14 July 2023 **Revised:** 08 September 2023 **Accepted:** 23 September 2023*

Rank signifying a more important domain.

5. Google Index: Google normally indexes trustworthy websites. Phishing websites, in contrast, do not enter the Google index because of their transient nature.
6. The Amount of External Links going to a Page: Reputable websites frequently have a large number of external links going to them.
7. Statistical Report-based: To identify phishing websites, up-to-date databases that are accessible to the general public, like Phish Tank, are maintained. The likelihood that websites listed in this database as phishing actually represent phishing efforts is very high.

V. METHODOLOGY

A. Data Pre-processing:

1. Removal of Unnecessary Column: The data pre-processing phase began with the removal of the 'index' column, which was deemed unnecessary for the analysis.
2. Data Transformation: The dataset used a range of values {-1, 1} to represent the results, where '-1' denoted phishing and '1' indicated legitimate URLs. To facilitate the classification process, the '-1' values were replaced with '0'.
3. Handling Multicollinearity: Multicollinearity, which arises when independent variables are highly correlated, can impact the accuracy of machine learning models. To detect multicollinearity, the 'DataFrame.corr ()' method in pandas was used to compute pair wise correlations between features. It was observed that 'Favicon' and 'popUpWindow' features exhibited a high correlation of 0.94. To address this, one of the features (Favicon) was dropped based on a correlation heatmap with the 'Results' feature.
4. Data Splitting: The dataset was split into training and testing sets, with 70% of the data used for training and the remaining 30% for testing.

B. Model Selection:

1. Logistic Regression: A logistic regression model was deployed, using the 'liblinear' solver with a maximum of 1000 iterations.
2. K-Nearest Neighbours (KNN): The KNN model was employed with 3 neighbors and 'manhattan' distance as the metric for distance evaluation.
3. Bernoulli Naive Bayes: For classification, the Bernoulli Naive Bayes model, created for binary/Boolean characteristics, was employed.
4. Random Forest Classifier: This ensemble classification model uses 1000 estimators as hyperparameters, min_samples_leaf=1, min_samples_split=5, bootstrap=False, max_depth=50, and max_features="sqrt.
5. Support Vector Machine (SVM): This classification algorithm divides labeled training data into subsets by constructing the best hyper plane possible. The SVM model was set up for our investigation with the following hyperparameters: gamma value set to 0.01 and C value equal to 10. The kernel was set to "rbf."

C. Performance Assessment:

Three crucial measures were used to gauge the models' efficacy:

1. Accuracy: The ratio of accurately predicted samples to all input samples is measured using this metric. It's critical to achieve high accuracy because correctly classifying URLs is our main goal.
2. Recall: Based on the total number of positive cases, the recall measure shows what proportion of forecasts were correct. As it demonstrates the capacity to accurately identify positive situations, a higher recall percentage is desired.
3. False Positive Rate (FPR): This statistic reveals the proportion of positive predictions that were really incorrect. Because misidentifying phishing websites as legal ones could result in considerable losses for individuals who visit such websites, minimizing the FPR is crucial tolowering the likelihood of this happening.

VI. RESULTS

Utilizing the validation data as a basis for training and evaluating the models, the results are shown in Table 1. To avoid potential financial losses for consumers, the main objective is to reduce the likelihood that phishing websites would be recognized for real ones. Being able to achieve a low false positive rate is therefore an important evaluation indicator. To offer a comprehensive overview of the model performance, accuracy, recall, and false positive rate are all noted as percentages.

1. Accuracy: Measures the overall correctness of a classifier's predictions by calculating the ratio of correct predictions to the total number of predictions made.

Formula: Accuracy = (True Positives + True Negatives) / (Total Predictions)

2. Recall (Sensitivity or True Positive Rate): Evaluates the classifier's ability to correctly identify positive samples (true positives) out of the total actual positive samples.

Formula: Recall = True Positives / (True Positives + False Negatives)

3. False Positive Rate (FPR): Determines the ratio of false positive predictions to the total number of actual negative

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8353
Article Received: *14 July 2023* ***Revised:*** *08 September 2023* ***Accepted:*** *23 September 2023*

samples.

Formula: FPR = False Positives / (False Positives + True Negatives)

Table 1: Classification Models Results (in percentage)

Model	Accuracy	Recall	False Positive Rate
Random Forest	98.32%	97.95%	4.60%
Support Vector Machine	94.20%	93.43%	6.57%
K-Nearest Neighbors	93.05%	93.40%	6.60%
Logistic Regression	93.50%	92.62%	7.38%
Bernoulli Naïve Bayes	91.25%	91.70%	11.32%

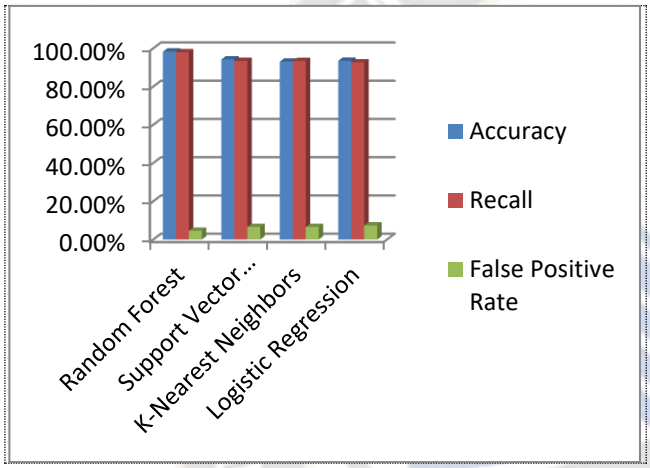

Figure 1: Classical models comparison

Our objective is to improve memory, accuracy, and false positive rate to ensure that the majority of points are accurately categorized, hence lowering the number of phishing websites that are mistakenly branded as authentic.

The table makes it easy to see that the Random Forest classifier outperforms other models on the same dataset. All three metrics—the best accuracy (98.32%), maximum recall (97.95%), and lowest false positive rate (4.60%)—meet our objectives. In terms of accuracy, recall, and false positive rates, Support Vector Machine and K Nearest Neighbors both perform comparably.

Only 93.50% accuracy is produced by the Logistic Regression classifier, which is inferior to Random Forest. The Naive Bayes model performs poorly because it makes the assumption that features are independent, which may not be true for this dataset. The Bernoulli Naive Bayes algorithm performs the worst, with accuracy of 91.25%, recall of 91.70%, and highest false positive rate of 11.32%.

Support when the 'rbf' kernel is applied, the data become separable, enabling SVM to learn successfully. Vector Machine performs well for linearly separable data.

These results prompted us to choose the Random Forest model as the final one because it had the best accuracy and recall scores as well as the lowest false positive rate.

VII. CONCLUSION

In this study, we investigated various machine learning models to identify phishing websites with the goal of identifying the best classification model with a high degree of accuracy. We found that the Random Forest Classifier performed remarkably well for phishing website detection after careful investigation. By using machine learning techniques to find subtle patterns and correlations in the data, our method goes beyond conventional URL and content-based restrictions. Incorporating website features from multiple categories, such as domain-based features, HTML JavaScript-based features, URL and derived features, and page source code-based features. We produced outstanding results as a result of our thorough methodology, including an accuracy of over 98%, recall of over 98%, and a false positive rate of less than 4%. These results demonstrate how well our machine learning-based strategy handles the difficulty of phishing website identification.

REFERENCES

[1] Antón, A. I., Earp, J. B., & Pankowsky, M. (2015). Social Engineering and Phishing Attacks: The Impact of Psychological Persuasion. Journal of Information Privacy & Security, 11(2), 61- 74. doi:10.1080/15536548.2015.1043353

[2] Arachchilage, N. A. G., & Love, S. (2014). An Investigation of Phishing Attack Techniques. Information Management & Computer Security, 22(5), 419-443. doi:10.1108/IMCS-04-2014- 0067

[3] Chang, K., & Xu, J. (2017). An Adaptive Method for Phishing Detection Based on URL Features. IEEE Access, 5, 17466-17475. doi:10.1109/ACCESS.2017.2752379.

[4] Kumar, S., Selvakumar, P., & Mary, A. L. P. (2018). A Comparative Study of Phishing Websites Detection Using Machine Learning Algorithms. International Journal of Information &Computation Technology, 8(6), 3971-3979.

[5] Nainar, N. J., & Halder, D. (2021). Adversarial Machine Learning: A Comprehensive Survey. Journal of Artificial Intelligence and Data Science, 3(4), 461-482. doi:10.36263/jaid.v3i4.185

[6] Phatak, D. S., & Swami, A. (2016). Detection of Phishing Websites: A Machine Learning Approach. International Journal of Advanced Computer Research, 6(23), 53-57.

[7] Sharma, S., & Upadhyay, R. (2019). An Investigation of Machine Learning Techniques for Phishing Websites Detection. Proceedings of the International Conference on Data Engineering and Communication Technology, 353-358. doi:10.1145/3318606.3318630

[8] Singh, S., & Biswas, K. (2020). A Review of Machine Learning Techniques for Phishing Detection. Proceedings of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8353
Article Received: *14 July 2023* ***Revised:*** *08 September 2023* ***Accepted:*** *23 September 2023*

the International Conference on Computer Communication and Informatics,689-693. doi:10.1109/ICCCI49486.2020.9110540

[9] Yao, H., Gou, H., & Wu, H. (2017). An Investigation of Machine Learning-Based URL Classification for Phishing Detection. Security and Communication Networks, 2017, 1-14. doi:10.1155/2017/6136476

[10] Liu, X., Srivastava, J., & Kumaraguru, P. (2011). PhishGuru: A People-Centric Phishing Countermeasure. In Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society (pp. 107-118). ACM.

[11] Zhang, Y., Kim, J., & Giles, C. L. (2019). Deep Learning for Phishing URL Detection. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 287-296). ACM.

[12] bin Saion, M. P. . (2021). Simulating Leakage Impact on Steel Industrial System Functionality. International Journal of New Practices in Management and Engineering, 10(03), 12–15. https://doi.org/10.17762/ijnpme.v10i03.129

[13] Akhtar, N., Khan, F. M., & Faye, I. (2018). A Comparative Analysis of Ensemble Learning for Phishing Detection. In Proceedings of the 10th International Conference on Computer and Automation Engineering (pp. 71-76). ACM.

[14] Chiew, K. Y., Tan, S. J., & Goi, B. M. (2020). An Ensemble Framework for Imbalanced Phishing URL Detection. Journal of Information Security and Applications, 52, 102577.

[15] Chen, L., Wang, C., Wang, Y., Wang, S., & Zhang, X. (2019). A Machine Learning-based Phishing Detection System with URL Semantic Features and Traffic Analysis. Journal of Computers & Security, 85, 184-195.

[16] Ahmad, S. N., Alshomrani, S. S., & Al-Mutiri, M. (2018). Evaluating machine learning classifiers for phishing detection. International Journal of Advanced Computer Science and Applications, 9(8), 58-64.

[17] Datta, S., Sharma, M., & Chavan, S. (2019). Phishing URL detection using machine learning. 2019 2nd International Conference on Data, Engineering and Applications, 1-5.

[18] Li, L., Deng, L., & Yegneswaran, V. (2017). Detecting and characterizing phishing webpages using machine learning. Computers & Security, 68, 36-49.

[19] Nainar, A., & Halder, S. K. (2022). Adversarial machine learning for phishing detection: Challenges and opportunities. Journal of Information Security and Applications, 65, 102961.

[20] Wang, Z., Zhou, X., & Wang, Y. (2021). Phishing detection using pre-trained language model with fine-tuning. 2021 9th International Conference on Information Technology in Medicine and Education, 30-34.

[21] Zhang, J., Ye, J., & Gao, S. (2020). An improved deep learning model for phishing website detection. Information Systems Frontiers, 22(5), 1111-1121.