International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

An Efficient Machine Learning Based Approach For Phishing Detection

[1]Mohamed Abdelshafea Mousa Abbas, [1]Ruth Ramya,[1] P.Vidyullatha, [2]M Suman, [2]Syed Inthiyaz

[1]Department of CSE, Koneru Lakshmaiah Education Foundation, Guntur, AP, India-522302
[2]Department of ECE, Koneru Lakshmaiah Education Foundation, Guntur, AP, India-522302

Abstract— Phishing is a breach of statistics safety through which attackers can advantage get admission to sensitive individual credentials through manner of using counterfeit net web sites closely equal to legitimate net web sites. Phishing starts of evolved with a fraudulent emails or exceptional communique that is designed to attack on a victim. If the victim clicks on immediately to the given url through manner of the cyber-attack that the attacker can get the extraordinary statistics or the essential statistics of the patients and misuse the statistics. There are one in all a type sorts of set of guidelines that can be used to come across the given url , whether or not or now no longer it is good url or the awful url . Among the ones all of these algorithms some algorithms will the ideal stop end result or the maximum percentage of the phishing attack detector. Some of the algorithms with a view to supply the almost accurate outcomes are, Random Forest Algorithm, Decision Tree Algorithm. The message exactly seems like the precise message which have become sent from the attackers but appears exactly similar to the message from an authorized enterprise agency or a company. This assignment can be accomplished through manner of using the Machine Learning using some libraries.

Keywords- Phishing; Personal Information; Machine Learning; vicious Links.

I. INTRODUCTION

Phishing Attacks has have come a large trouble now a days, that the sufferers were given without problems trapped with inside the arms of the attackers which became very unlawful and unhappy aspect for the sufferer. The essential factor that must be taken into consideration is that the customers should or capable of apprehend that which URL is malicious and which URL is secure to apply and may use the hyperlink. So, to apprehend the trouble and act accordingly, this mission will assist the customers to become aware of the URL that is secure or now no longer. To resolve the trouble, in this mission we've evolved a software program which detects that the given url is phishing URL or it's far a now no longer a phishing URL. For this we've evolved a software program the use of the Machine Learning with a number of the libraries with inside the python. The fraud internet site which appears precisely because the unique URL internet site. Experts can become aware of faux web sites however now no longer all the customers can become aware of the faux internet site in the end the person can turn out to be the sufferer of the cyber assault through the attackers and lead those to reachable to the customer's non-public facts and the exclusive facts. Also, the attacker can be capable of souse borrow banks account credentials. Phishing assaults made smooth for the attackers because of loss of person awareness.

II. LITERATURE REVIEW

In [4] the Author have detected phishing web sites through the usage of diverse device gaining knowledge of algorithms after which as compared the accuracy of the one-of-a-kind algorithms. Their experimental effects indicated at Random Forest set of rules having the very best accuracy, don't forget and precision. A class version is proposed in to categorize the phishing assaults. Feature extraction turned into completed from diverse web sites primarily based totally at the UCI Irvine ML repository. Authors of proven a technique of detecting phishing e mail assaults the usage of NLP and ML. In [2] they executed semantic evaluation of textual content for detecting any type of malicious activity. NLP turned into used to parse sentences the authors have proposed algorithms for function choice for phishing detection to enhance the first-rate of the dataset. They as compared their set of rules with different usually used algorithms for class on foundation of accuracy. In [7] they look at confirmed that Tree algorithms didn't paintings properly on faded datasets. Lazy K Star set of rules confirmed the exceptional effects. The whole looks at turned into completed the usage of Weka. In, a version turned into created the usage of Random Forest set of rules to categorize phishing URLs.

Their set of rules confirmed 95% accuracy at the check dataset. The authors [6] have used neural networks to extract capabilities from the URL with nonunique know-how approximately the URL. An accuracy of 94.18% turned into received the usage of Adam optimizer. The authors did a comparative look at among logistic regression strategies with bi grams and deep gaining knowledge of strategies like CNN and CNN-LSTM architectures. The authors suggested an accuracy of 98% the usage of CNN-LSTM architecture.

In [8], the authors suggested of phishing detection version in Chinese Websites. The overall performance of the version turned into studied through mining the semantic capabilities of phrases in Chinese net pages. Different device gaining knowledge of algorithms like Random Forest, Ada boost and Bagging have been as compared at the dataset. In [3], the authors have evolved an extension to Google Chrome for phishing internet site detection the usage of device gaining knowledge of algorithms. The UCI ML Repository has been used for this purpose. The downside of this extension is that the wide variety of malicious web sites is growing each day with new web sites developing every day and the education set for the look at is just too small. In [5] The authors of have defined approximately the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

one-of-a-kind URL capabilities which includes number one area, sub area, and rating of web sites for phishing detection. In [6] the authors of evolved a device known as Phish Score which does lexical evaluation at the URL to stumble on phishing. They used the relatedness of the URLs of their look at for growing the device. In, [7] the authors have defined approximately net spoofing assaults classes and attempted to apply area call capabilities to decide phishing URLs. The authors of have studied the accuracy of various classifiers for prediction of unsolicited mail emails.

In [2] they have accrued the masses of phrases from the e-mail through textual content mining which turned into finally applied for class the usage of not unusual place ML algorithms like random forests, choice bushes set of rules, SVM set of rules, BART set of rules and neural networks. But amongst all those algorithms, Random Forest set of rules turned into discovered to be correct amongst those algorithms.

In [4] the authors have collected masses of web sites with a purpose to be categorized as suspicious and phishing. In [7] They have used records mining strategies like Random Forest set of rules, Neural Network set of rules, Decision Tree set of rules and Neural Networks to categorize the internet site into one of the exceptional and maximum appropriate classes. Finally, the effects of the paintings had been as compared with different works on each comparable and diverse dataset. The challenge of this paintings is that there's no description of the phishing capabilities considered for device gaining knowledge of purpose.

III. Problem Discussion

Phishing Attacks has have become a large hassle now a days, that the sufferers were given without difficulty trapped with inside the arms of the attackers which changed into very unlawful and unhappy issue for the sufferer. The predominant factor that must be taken into consideration is that the customers should or capable of recognize that which url is malicious and which url is secure to apply and may use the link. So, to recognize the hassle and act accordingly, this task will assist the customers to become aware of the url that is secure or now no longer. To remedy the hassle, in this task we've got evolved a software program which detects that the given url is phishing url or it's miles a now no longer a phishing url. For this we've got evolved a software program the use of the Machine Learning with the aid of using the use of a number of the libraries with in inside the python. The faux internet site which appears precisely because the authentic url internet site. Experts can become aware of faux web sites however now no longer all of the customers can become aware of the faux internet site in the long run the consumer can come to be the sufferer of the cyber assault with the aid of using the attackers and lead them to handy the customer's non-public information and the private information. Also, the attacker can be capable of thieves' banks account credentials. Phishing assaults made clean for the attackers because of loss of consumer awareness.

By the use of this phishing assault detector, we will be capable of expect that which url is secure to apply and which url is malicious Urls.

IV. Methodology

Determining whether a given internet site URL is phishing or valid is a binary category hassle which may be solved with the assist of labelled records on which supervised gaining knowledge of may be applied [4]. Data series for this hassle calls for latest internet site URLs belonging to each class - phishing and valid. The underneath can be accompanied through coaching of the dataset through extracting applicable functions which enables in distinguishing phishing web sites from valid web sites. The functions want to be processed if you want to supply as enter to the system gaining knowledge of algorithm. Then the version is skilled the use of the schooling set and its accuracy is decided at the trying out set. The glide chart depicting the technique is summarized with inside the underneath figure1.

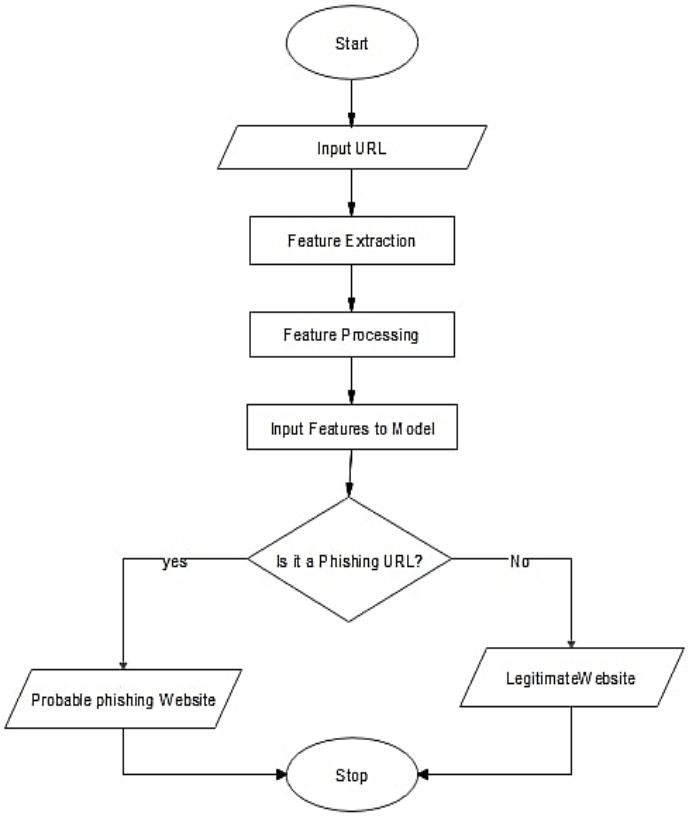


Fig. 1 This are the results of the Classification Phase; Here the url was classified as malicious or non-malicious. Support vector machines classified and identified numerous websites as legitimate or phishing websites based on their inputs. Heuristics were combined in this research paper to help identify phishing websites.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

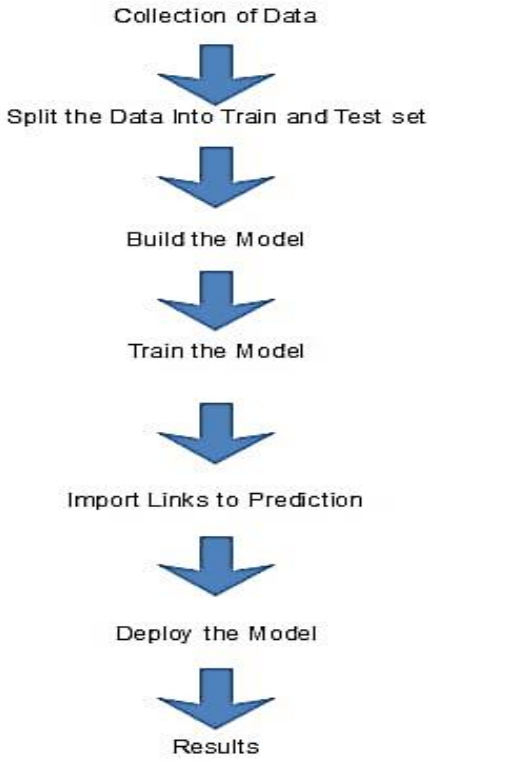


Fig. 2 Results for deployment of the model.

A. Dataset Description

A relevant dataset is obtained and pre-processed in order to assess the suggested artificial intelligence approach for phishing detection. The dataset comprises of a variety of cases of phishing and genuine websites, enabling a thorough assessment of the method of detection. In order to acquire datasets, real-world data must be gathered from a variety of sources such as open databases, research on cybersecurity datasets, and partnerships with industry partners. The dataset contains a wide variety of phishing websites that use a variety of attack methods, including spear phishing, phishing emails, and social engineering. The preprocessing procedures guarantee the consistency and quality of the dataset, allowing for the detection model's accurate evaluation. The dataset underwent the following preparation methods:

Data Cleaning: The dataset is examined for any incomplete, duplicate, or irrelevant instances. Noise or inconsistencies in the data, such as missing values or erroneous entries, are addressed through data cleaning procedures.

Data Normalisation: To eliminate biases caused by different feature ranges, characteristics in the data set are normalised to a single scale. To guarantee equity in feature representation, normalization approaches like min-max scaling and z-score normalization are used.

Balancing the Dataset: As phishing instances are generally fewer in number compared to legitimate instances, class imbalance is a common issue. To address this, techniques such as oversampling or under sampling are employed to balance the dataset, ensuring equal representation of phishing and legitimate instances during training and evaluation.

The dataset description includes the following information:

- The overall number of cases in the data set, comprising both phishing and genuine websites, is referred to as the "number of instances."
- Features: The group of features used to detect phishing, which includes behavioral, domain-based, content-based, and URL-based features. These characteristics could include things like URL width, domain age, the inclusion of questionable keywords, the outcomes of an HTML analysis, and interaction between users' patterns.
- Class Distribution: The distribution of instances across the phishing and legitimate classes in the dataset, indicating the class imbalance and the magnitude of the challenge.
- Data Partitioning: The proportions in which the dataset is divided into training, validation, and testing sets. The sizes of these sets are determined to ensure an adequate representation of instances for model training, tuning, and evaluation.

B. Feature Engineering

In the context of phishing detection, analyzing the HTML structure and embedded JavaScript code of web pages is crucial for identifying potential phishing attempts. This section focuses on the techniques employed to extract relevant features from HTML and JavaScript to enhance the effectiveness of the phishing detection model.

- HTML Analysis:
  HTML analysis involves parsing the HTML code of a webpage to extract features that can indicate the presence of phishing elements or malicious behavior. Some key techniques used for HTML analysis include:
- Form Field Detection:
  Phishing attacks often involve the use of forms to capture user credentials or sensitive information. By analyzing the HTML structure, the presence of form fields and their attributes (e.g., input type, name, action URL) can be identified. Suspicious or unexpected form elements can raise a red flag for potential phishing activity.
- Hidden Element Detection:
  Phishers may hide certain elements on the webpage to deceive users or trick automated systems. Analyzing the HTML can reveal the presence of hidden elements (e.g., hidden input fields, CSS styles) that may indicate malicious intent.
- Redirection Detection:
  Phishing attacks often involve redirecting users from legitimate webpages to malicious ones. By examining the HTML code, any redirection scripts, meta tags, or JavaScript-based redirects can be identified, helping to detect potential phishing attempts.
- JavaScript Analysis:
  JavaScript is widely used to enhance interactivity and dynamic behavior on webpages. However, it can also be leveraged by phishers to execute malicious code or perform unauthorized actions. Analyzing JavaScript code can provide valuable insights into potential phishing attempts. Some techniques used for JavaScript analysis include:
- Script Source Analysis:
  Examining the sources of JavaScript files or embedded scripts can help identify any suspicious or malicious URLs. This involves checking the domains, checking for obfuscated code, or comparing with known malicious script repositories.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

- Function Call Analysis:
  Analyzing function calls within JavaScript code can provide insights into potentially malicious activities. For example, detecting calls to functions related to user input interception, data exfiltration, or unauthorized network requests can indicate phishing behavior.
- Event Handler Analysis:
  Phishers often exploit event handlers (e.g., onclick, on submit) to execute malicious actions when users interact with the webpage. Analyzing the JavaScript code can help identify any suspicious or unexpected event handlers associated with form submissions, link clicks, or other user interactions.
- DOM Manipulation Analysis:
  JavaScript can be used to dynamically manipulate the Document Object Model (DOM) of a webpage. Analyzing the JavaScript code can reveal any suspicious or unauthorized DOM manipulations that may be indicative of phishing activity.

In [1] According to the author, phishing attacks aren't the only ones increasing in frequency these days. But regardless of how well you prepared, it also becomes more sophisticated. An attacker may occasionally force you to the ground. Don't worry; we'll provide you some tactics that will always give you the upper hand against phishing attempts.

- Self-Education:
  If you can, invest in strong spam filtering. Avoid clicking on any unfamiliar links that are sent to you via email from unknown users. Downloading software, scripts, documents, or attachments from unidentified sources is not advised. Be wary of prone websites. Make sure everything is proper by paying attention to the letters on the website. If you detect a minor adjustment, such as a letter changing, just disregard it. I'll use Google.com and Microsoft.com as examples. As a final piece of advice, I'd advise searching for the website. In certain ways, this may make it easier for you to fall victim to a DNS poisoning attack.
- Keep Backups Up to Date*:*
  Always have backups on hand. Update your backup frequently. This is the best security against all types of cyber-attacks, not just phishing attempts. You can restore everything from your backups and run the business while the attacker tries to shut it down by blocking your data.
- MFA — Multi-Factor Authentication*:*
  As stated, you must provide more than one piece of identification to confirm your identity. Up until the attacker obtains all your credential factors, you are secure. This might be useful. Even the attacker stole your password to some extent. Always keep your login credentials secure, and it's best to change them periodically.
- Keep Change Credentials Over Time*:*
  It's best to regularly change the login credentials whether you are being targeted or not. This would undoubtedly lower the likelihood that social engineering attempts would succeed.
- Adhere the best practices*:*
  You must be knowledgeable about and adhere to the finest cyber security practices in your daily life. Several widespread procedures serve as a deterrent to all such phishing attempts:
- Follow the password guidelines. Continue to update your cellphones, tablets, and PCs. Use encryptions and antivirus software. Observe all email security recommendations. Use a VPN whenever necessary.

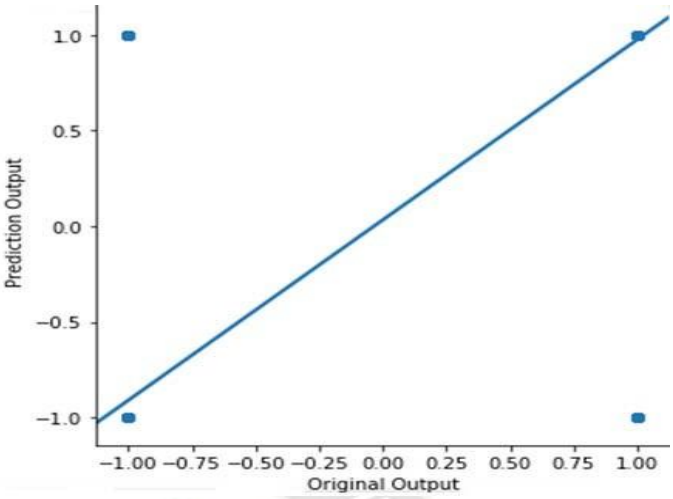
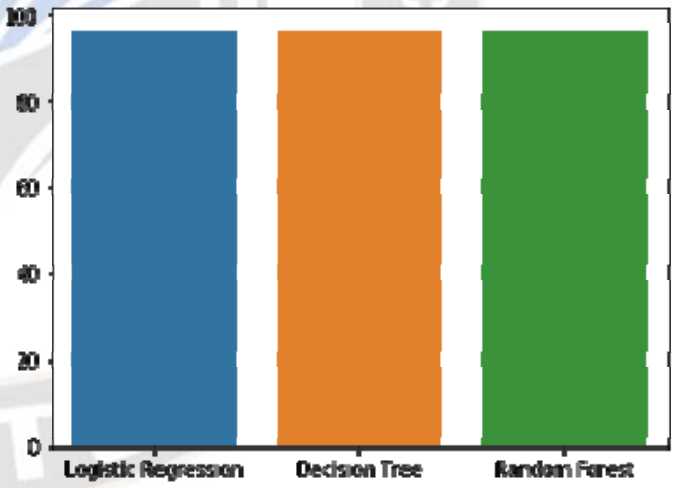


Fig. 3. Linear regression plot of original output versus predicted output

V. RESULTS

The forest randomization algorithm foresaw this. It slightly deviates from the result that the phished websites were supposed to produce.



The true positive, false positive, true negative, false negative count and accuracy results of 9064 test websites using Decision Tree, Logistic Regression and Random Forest Algorithms in machine learning is shown in table 1.

Algorithm	TN	TP	FP	FN	Accuracy
Logistic Regression	6447	2287	325	17	96.23 %
Decision Tree	6393	2341	326	16	96.23 %

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

Random Forest	6392	2374	297	13	96.58 %

Table 1. Performance comparison using Decision Tree, Logistic Regression and Random Forest Algorithms.

A. Machine Learning Algorithm

Logistic Regression:

Formula:

$p(y=1|x) = 1 / (1 + e^{(-z)})$

$p(y=1|x)$: The likelihood that the class z is positive: Features and coefficients combined linearly.

For classification in binary tasks, logistic regression (LR) is a common machine learning approach that is suitable for phishing detection. Based on a number of independent factors, it is an algorithm that forecasts the likelihood of a binary result.

Using a collection of input features, logistic regression works by predicting the likelihood of a positive category (phishing). The linear sum of the input features is transformed into a score of probability ranging from 0 to 1 using the model of logistic regression using a logistic function, commonly referred to as the sigmoid function.

Advantages of Logistic Regression for Phishing Detection:

1. Simplicity: Logistic regression is relatively simple to implement and interpret compared to more complex machine learning algorithms.
2. Efficiency: Logistic regression algorithms have low computational overhead, making them efficient for large-scale datasets.
3. Feature Importance: Logistic regression provides coefficients for each input feature, allowing for the identification of important features contributing to the prediction.
4. Probability Estimation: Logistic regression models can provide probability estimates, allowing for a better understanding of the confidence in the predictions.
5. Robustness to Noise: Logistic regression is less sensitive to noisy data compared to other algorithms like decision trees.

Limitations of Logistic Regression for Phishing Detection:

1. Linear Decision Boundary: Logistic regression assumes a linear decision boundary, which may limit its ability to capture complex relationships between features in some cases.
2. Feature Engineering: Logistic regression heavily relies on feature engineering to represent the data effectively. Choosing relevant features and performing appropriate transformations is crucial for achieving good performance.
3. Lack of Nonlinearity: Logistic regression cannot capture nonlinear relationships between features unless additional feature engineering techniques, such as polynomial features or interaction terms, are applied.
4. Assumption of Independence: Logistic regression assumes that the input features are independent of each other, which may not hold true in all cases.

B. Random Forest

Random forest is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy and robustness of predictions. It is widely used in various domains, including phishing detection. Random forest operates by constructing a multitude of decision trees using random subsets of the training data and random subsets of the input features.

Working Principle of Random Forest:

Using a different portion of the training information and a randomly chosen portion of the input features, the random forest technique creates a collection of decision trees. Each tree in the ensembles is trained using a distinct bootstrap sample (randomly picked examples with replacements) from the original dataset during the training phase. To ensure variety among the trees, a selected group of traits is also taken into account for every split in the tree.

Advantages of Random Forest for Phishing Detection:

1. High Accuracy: Comparing to individual decision trees, the accuracy of predictions made by random forest, which combines predictions from several decision trees, is improved. Random forest's ensemble structure aids in lowering overfitting and enhancing generalisation.
1. Resistance to Noisy and Oddities: Random Forest is resistant to data points with noise and outliers. Multiple trees are taken into account, which lessens the effect of outliers and produces forecasts that are more reliable.
2. Characteristic Importance: Random Forest offers a way to quantify feature importance, making it possible to pinpoint the elements that are most crucial for phishing detection. This knowledge can aid in feature choice and in comprehending the fundamental traits of phishing assaults.
3. High-dimensional data handling: Random Forest works well even with these kinds of datasets. It is capable of handling many input features well without the need for sophisticated feature design or dimensionality reduction approaches.
4. Effectiveness: Random forests can be parallelized, making it possible to efficiently compute in systems with remote computing or multi-core processors. This qualifies it for datasets of a significant size.

Limitations of Random Forest for Phishing Detection:

1. Interpretability: The random forest model is not as easily interpretable as logistic regression or decision trees. Understanding the decision-making process of a random forest model can be more complex due to the ensemble of trees and the randomness involved in the construction process.
2. Model Development Time: randomly generated forest models may require more time to develop than other algorithms, particularly when working with large numbers of trees or challenging datasets. Parallelization strategies, however, can reduce the training time.
3. Memory Usage: Due to the long-term preservation of different decision trees within an ensemble, random forest models may use more memory. If you have restricted computing resources, you might want to take this into account.

C. Support Vector Machines (SVM)

SVM is a potent machine learning technique that is frequently employed for binary sorting tasks, making it a good choice for phishing detection. In a feature with high dimensions space, SVM seeks to identify an ideal hyperplane that maximally divides points of data from various classes.

Working Principle of Support Vector Machines: The SVM algorithm maps the input data into a higher-dimensional feature space using a kernel function. In this transformed space, SVM seeks to find a hyperplane that maximizes the margin, i.e., the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

distance between the hyperplane and the nearest data points of each class. These data points, called support vectors, play a crucial role in defining the decision boundary.

Advantages of Support Vector Machines for Phishing Detection:

1. Effective for High-Dimensional Spaces: SVM still performs effectively when there are more dimensions (features) than instances. This qualifies it for phishing detection jobs including numerous pertinent features.
2. Robust to Overfitting: The SVM seeks out a decision border with the greatest margin, which aids in lowering overfitting. Well-suited to generalizing to unseen data, it reduces the possibility of misclassifying fresh phishing attempts.
3. Handling Nonlinear Data: SVM may successfully handle nonlinear interactions between features by employing the proper kernel function (e.g., polynomial in radial basis function). This adaptability makes it possible to detect intricate patterns in phishing scams.
4. Feature Importance: By analyzing the size of the weights given to various features, SVM can shed light on the significance of a given feature. This knowledge aids in determining the key indicators for phishing detection.
5. Margin-Based Classification: To improve generalisation and robustness, SVM focuses on maximising the margin between classes. By setting larger punishments for misunderstanding occurrences of the minority class, it may manage unbalanced datasets.

Limitations of Support Vector Machines for Phishing Detection:

6. Complexity of Computation: SVM can be physically expensive, particularly when dealing with huge datasets or sophisticated kernel functions. Due to the potential impact on training and prediction times, effective implementation and optimization strategies are needed.
7. SVM is susceptible to outliers as well as noise in the initial training data, that might affect where the decision boundary is placed. To lessen their influence, pre-processing procedures like data cleansing and outlier detection are required.
8. Choice of Kernel Function: The performance of SVM heavily relies on selecting an appropriate kernel function. The choice of the kernel function and its parameters may require some trial and error or domain knowledge.
9. Lack of Probability Estimates: SVM was originally designed for binary classification and provides decision scores rather than direct probability estimates. Probability calibration techniques, such as Platt scaling or isotonic regression, can be applied to obtain reliable probability estimates.

VI. Conclusion

This mission explains the present protection troubles in today's virtual global with appreciate to phishing and the system via which phishing is carried out. Phishing is a severe security concern which may also cause lack of touchy private records because of smart disguising of phishing mails via way of means of attackers. These paintings specially make a specialty of figuring out capabilities beneficial for detecting phishing web sites primarily based totally on entirely the URL of the internet site and making use of machine learning algorithms to categorize web sites into valid and phishing. The look at involves comparison of effects of seven device getting to know algorithms with Random Forest set of rules emerging as the maximum correct and hence, maximum suitable set of rules for this binary classification. We got visible how phishing is a large risk to the safety and protection of the net and how phishing detection is a vital trouble domain. We have reviewed a number of the traditional approaches to phishing detection; specifically blacklist and heuristic assessment methods, and their drawbacks. We have examined device getting to know algorithms at the Phishing Websites Data set and reviewed their effects. We then decided on the first-rate set of rules primarily based totally on its overall performance and built a Chrome extension for detecting phishing net pages. The extension lets in easy deployment of our phishing detection version to give up users. We have detected phishing web sites using Random Forest set of rules with most quantity of accuracy.

VII. Future Work

While the presented approach for phishing detection using machine learning algorithms shows promising results, there are several areas that can be explored for further improvement and research:

1. Feature Engineering Enhancement: Look into any additional features that can help you identify between trustworthy and phishing websites. A more thorough investigation of URL elements, domain image, page happy, and user behavior patterns may be required.
2. Utilise dynamic analysis tools to record users' in-the-moment interactions with websites. To improve the accuracy of detection, this can entail examining user mouse movements, JavaScript behavior, and the time spent on various browser items.
3. Investigate the application of deep learning models for phishing detection, including convolutional neural networks, or CNNs, and neural networks with recurrent connections (RNNs). These algorithms have demonstrated performance across a variety of fields and are capable of detecting complex patterns in attempts at phishing.
4. To increase the accuracy and resilience of phishing detection, use ensemble methods, which combine numerous predictive models or ensemble methodologies. Results could be more trustworthy if techniques like stacking, which aggregate predictions from different models, were used.
5. Analysis of streaming data: Create models that can manage stream data in actual time to quickly spot changing phishing attacks. This might entail methods like learning online, where the model keeps itself updated as fresh information becomes available.
6. Enhance the comprehension of artificial intelligence models, particularly those with high levels of complexity like random forests. For trust to be established and responsibility to be maintained, it is essential to comprehend why a model produces a particular prediction.
7. Imbalanced Dataset Handling: Investigate cutting-edge methods like synthetic data creation, adaptive increasing,

and cost-sensitive learning to address category imbalance in the dataset. To prevent models from becoming biassed towards the majority class, handling class imbalance is crucial.

8. Cross-Dataset Validation: Verify the trained model's performance on diverse datasets gathered from different places and timeframes. This guarantees that the model's efficacy is not restricted to a particular dataset.
9. Real-World Installation and User Interaction: Test the suggested model in a practical situation, for as by integrating it into email clients or web browsers to provide users instantaneous alerts. User comments and conversations regarding the system can offer insightful information for future enhancements.
10. Examine the model's resistance to adversarial attacks, in which attackers purposefully change features to avoid being detected. For the model to remain effective, strategies to make it resistant to such attacks must be developed.
11. Implement tools for continual model updating to adjust to new attack patterns and developing phishing strategies. To keep ahead of attackers, the model must be updated frequently with new data.
12. Interdisciplinary Collaboration: Work with specialists in the field of cybersecurity, psychology, and computer science to develop a comprehensive strategy for phishing detection that considers both technical and psychological aspects.

REFERENCES

[1] Cao, Y., Han, W., & Le, Y. (2008, October). Anti-phishing based on automated individual white-list. In Proceedings of the 4th ACM workshop on Digital identity management (pp. 51-60).

[2] Rao, R. S., & Ali, S. T. (2015, April). A computer vision technique to detect phishing attacks. In 2015 Fifth International Conference on Communication Systems and Network Technologies (pp. 596-601). IEEE.Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).

[3] Zamir, A., Khan, H. U., Iqbal, T., Yousaf, N., Aslam, F., Anjum, A., & Hamdani, M. (2020). Phishing web site detection using diverse machine learning algorithms. *The Electronic Library*, *38*(1), 65-80.

[4] Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In Cyber Security (pp. 467-474). Springer, Singapore.

[5] Abdur, B. A. R. C. S., & Crescent, R. C. B. A. R. REVIEW PAPER ON PHISHING ATTACKS.

[6] Purbay, M., & Kumar, D. (2021). Split behavior of supervised machine learning algorithms on the behalf of phishing URL detection. Lecture notes inside electrical engineering.

[7] Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B., & Bindhumadhava, B. S. (2020, January). Phishing website classification and detection using machine learning. In *2020 international conference on computer communication and informatics (iccci)* (pp. 1-6). IEEE.

[8] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345-357.

[9] Xie, Y., Zhang, Y., & Sun, G. (2017). Phishing website detection based on random forest. Proceedings of the 3rd International Conference on Electronic Information Technology and Intellectualization (ICEITI), 469-472.

[10] Khan, M. K., & Wahid, A. (2019). Phishing detection using support vector machines and feature selection. Journal of Information Security and Applications, 47, 102-114.

[11] Jagadeesan, A., & Rengaramanujam, S. (2020). An efficient machine learning model for phishing website detection using logistic regression and random forest algorithms. Computers & Electrical Engineering, 82, 106628.

[12] Singh, A. P., & Rani, M. (2021). A hybrid machine learning approach for phishing website detection using random forest and support vector machine. Journal of Network and Computer Applications, 183, 103036.

[13] Chen, T., Wang, J., Li, S., & Zhang, H. (2018). An improved random forest classifier for phishing detection. Proceedings of the 5th International Conference on Computer Science and Application Engineering (CSAE), 326-330.

[14] Xia, C., Du, F., Li, Y., & Yang, M. (2019). Phishing website detection based on improved random forest algorithm. Proceedings of the International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI), 313-325.

[15] N. Suresh et al., "Crop Yield Prediction Using Random Forest Algorithm," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 279-282, doi: 10.1109/ICACCS51430.2021.9441871.

[16] Murali Krishna, B., et al. “FPGA implementation by using XBee transceiver” (2016) Indian Journal of Science and Technology, 9 (17), art. no. 93032.

[17] Inthiyaz, S., et . al. “Skin disease detection using deep learning” (2023) Advances in Engineering Software, 175, art. no. 103361.

[18] Sreelakshmi, D., Inthiyaz, S.” Fast and denoise feature extraction based ADMF–CNN with GBML framework for MRI brain image” (2021) International Journal of Speech Technology, 24 (2), pp. 529-544.

[19] Inthiyaz, S., et.al., “Design of bi-trigger sram using schmitt trigger for low power 13t cmos application” (2019) International Journal of Scientific and Technology Research, 8 (12), pp. 1466-1471.

[20] Siva Kumar, M., et.al “Delay estimation of different approximate adders using mentor graphics” (2019) International Journal of Advanced Trends in Computer Science and Engineering, 8 (6), pp. 3584-3587.