www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2497-2502

e-ISSN : 2583-1062
Impact Factor : 7.001

DETECTING PHISHING THREATS IN REAL-TIME: THE MACHINE LEARNING APPROACH

Jagarlamudi Krishna Thrishagna[1]

[1]3rd Year CSE-AI&ML GMRIT

ABSTRACT

Among the huge threats to cybersecurity are phishing attacks, in which individuals and organizations are targeted to steal sensitive information. Although phishing was first used in 1996, it has developed to be the deadliest and most serious online crime. Mainly using email deception as the primary method for deceptive emails, phishing then makes use of spoof websites to get the necessary information from the target audience. Advanced attackers are now developing more complicated ways of conducting their attacks, which makes the conventional detection methodologies such as rule-based systems and blacklists behind the times. In an effort to improve the detection of phishing attempts, this research explores the use of machine learning. We have tried several machine learning algorithms with the purpose of discovering whether they can effectively detect phishing emails, URLs, and websites. All these include decision trees, linear regression, support vector machines, and neural networks. Our research revealed that the application of such massive datasets and sophisticated feature extraction methods may drastically improve the accuracy of detection as well as reduce false positives when implementing machine learning models.It aims for ensemble approaches, which are seen to use many models in a bid to improve performance. From this study, it is realized that machine learning can help detect more dynamic phishing attempts. It makes cybersecurity systems more resilient to the nature of new attacks. The work, therefore, contributes to the growing processes for more dependable and effective strategies to protect digital environments against the specific attack of phishing**.**

Keywords- Phishing Detection, Machine Learning, Cybersecurity, Deep Learning, Neural Networks, Ensemble Methods, Feature Extraction.

1. INTRODUCTION

Phishing remains one of the most serious and emerging cybersecurity problems that poses grave threats by deceiving people and companies into providing private information since its dawn in 1996 phishing attacks have graduated from mere e-mail scams to complicated schemes that almost completely resemble genuine websites and services thus causing severe damage to financial fronts and an individuals reputation this development therefore has been fueled by the development of many detection techniques meant to counter these dangers techniques include fuzzy logic search engine-based strategies and the analysis of webpage linkages and anti-phishing toolsnot with standing advancements problems with scalability runtime efficiency and real-time detection still hinder the creation of efficient phishing detection systems there have been more research interests in recent times in using machine learning techniques notably random forest xgboost and lightgbm to enhance detection accuracy and minimize false positives additionally effective group education incorporating multiple models in order to achieve improved performance as opposed to conventional and deep learning techniques the recent breakthroughs include advanced algorithms and cloud-based solutions in the systems such as dephides and phishnot which aim to achieve maximum accuracy and processing efficiency the continuously evolving phishing techniques justify the need for flexible and scalable detection systems keeping abreast of sophisticated phishing attacks further research will be directed toward improving the quality of datasets streamlining current procedures and exploring new approaches

2. LITERATURE SURVEY

phishing being a major online threat since 1996 has escalated into sophisticated attacks that pose significant financial and reputational damage other detection methods have been developed targeting different website features such as urls html and domain properties gowtham et al 2017 designed a system to identify suspicious domains at the cost of a true positive rate of 9953 but with scalability issues li et al 2019 employed the stacking model which achieved 986 without mentioning any runtime efficiency other techniques include anti-phishing kits fuzzy logic models and search engine-based techniques which performed well but with practical issues such as realizing real-time processing and adapting to evolving phishing tactics. [1]

The research paper is about phishing url detection focusing on the login url instead of the home page some challenges like relatively high false-positive rates are discussed and placed against the efforts some classifiers like random forest xgboost,lightgbm have made in recent years the techniques of phishing detection have improved markedly from the efforts made toward tackling the problem head-on and addressing it through advanced approach systems. [2]

The study examines the effectiveness of ensemble machine learning methods in detecting phishing websites, highlighting their superior performance compared to traditional and deep learning algorithms. Ensemble methods, particularly Random Forest, show high detection accuracy and computational efficiency even with reduced feature sets. The study emphasizes the importance of feature selection techniques for optimizing feature subsets and enhancing detection capabilities. Deep learning models require resources and can be complex to implement, making ensemble machine learning methods a valuable tool in cyber threat combat. [3]

This paper on research for detecting phishing sites through machine learning deals with the analysis of URLs and domain names. It talks about the disadvantages in most of the already-existing datasets, involving an information deficiency in feature classification and a lack of real-life examples. The authors introduced a new detection method based on six classifier algorithms with eleven predetermined features, simplifying feature extraction and thus reducing processing overhead. In fact, the study asserts the efficiency of Random Forest and Support Vector Machines. [4]

The DEPHIDES study introduced a Phishing Detection System based on Deep Learning, which uses algorithms that are used for analyzing the URL and determining the malicious patterns within it. It makes the detection highly efficient and accurate by using Artificial Neural Networks, Convolutional Neural

Networks, and Recurrent Neural Networks. Moreover, the performance evaluation metrics employed include accuracy, precision, and recall. DEPHIDES aims to improve online security through phishing threat identification. [5]

Phishing attacks have changed drastically since 1996 and have meant much financial as well as reputational risk. Techniques for detection include research-based checking of links in webpages, stacking models, anti-phishing kits, fuzzy logic models, and search engine-based methods. These scalability, runtime performance, and real-time detection issues, however, are critical issues to robust phishing detection systems. [6]

The growing malicious activities on the internet have challenged the need to design more sophisticated detection methods. Gowtham et al. in 2017, suggested a system that identifies suspicious websites, but requires very wide scaling. Li et al. in 2019 applied the multi-layered model with no focus on efficiency of performance. Other approaches include specialized tools, logic-based systems, and web search techniques.However, solutions found in these approaches still need to overcome huge efforts and immediate sensing problems, which makes them need further development. [7]

One of the major threats to security is phishing. The traditional blacklist and rule-based approaches are not adapted to this threat. This paper presents a proposed system, PhishNot, which uses machine learning for effective phishing URL detection based on a reduced set of 14 features. The detection accuracy using the Random Forest algorithm stands at 97.5%. The system has been implemented for practical deployment on the cloud, processing, and scalability in real-world applications. High precision of the system with efficient feature selection and adaptability make it a useful tool for the security of transactions on Ethereum from cyber threats. Future work envisions an extension in the functionality of PhishNot to add more capabilities concerning the detection of other illegal activities on Ethereum. [8]

Over 50% of the total cybercrimes on Ethereum are phishing. Eth-PSD overcomes this by including a balanced dataset with a feature selection approach based on the machine learning model. A voting-based technique has been used for feature identification in the system, and the classification is tested with multiple classifiers. It achieved an accuracy of 98.11% in detecting phishing scam, which outperformed current models. Future work aims to expand Eth-PSD's capabilities to detect other illegal activities on Ethereum. [9]

This study investigated the performance of machine learning models in detecting phishing attacks on web pages. Three models- namely, KNN, SVM, and RFwere tested in the experiment. According to the experiment results, Random Forest performed the best. In fact, the group obtained 98.35% accuracy with a 100% True Positive rate and 90.48% True Negative rate. This work proposed the development of a browser extension, PhishNet, based on the rules from the Random Forest model to identify phishing websites in real-time. Further studies should be conducted to enhance accuracy and explore other machine learning tools. [10]

Machine learning has now designed a multilayer stacked ensemble learning model for the detection of phishing websites. Comparing to baseline models, such multilayer stacked ensemble learning achieves better high levels of accuracy in the detection of phishing attacks. It attained accuracy from the datasets of 96.79% up to 98.90% while deployed. Using multiple classifiers and one meta-learner architecture, it has posed a better solution for phishing detection. The success of the model in critical datasets such as UCI and Mendeley, therefore, underlines the promise of ensemble learning in enhancing cybersecurity. [11]

Phishing is a major threat to cybersecurity and there is provided a taxonomy of methods designed for phishing detection. The methods have some challenges such as high computational complexity, manual parameter tuning that brings about arduous efforts. The research also revealed that current available datasets are limited in terms of diversity and quantity. Future work lies in hybrid model improvement of DL algorithms, quality enhancement of datasets, and the use of

explainable neural networks to improve interpretability. The paper recommends scalable and flexible phishing detection systems, which adapt to changing tactics at reduced computational cost without loss in accuracy. [12]

Phishing is probably one of the most persistent cybersecurity threats, which uses different deceptions to masquerade as legitimate websites and steal sensitive information. Three approaches are detection listbased, similarity-based, and machine learning-based. List-based methods are simple but impractical for zero-hour attacks. Similarity-based methods are accurate but computationally expensive and slow. Machine learning-based methods can detect new attacks; however, the effectiveness of this method may depend on the quality features and diversity in datasets. [13]

This paper introduced a new machine learning approach of phishing detection that focused on the Hybrid Ensemble Feature Selection technique. Using a Cumulative Distribution Function gradient algorithm, HEFS improves accuracy in phishing detection by relevant features identified from datasets. Compared to the performance of other machine learning models, Decision Trees, Random Forest, and Neural Networks each showed improved accuracy with HEFS. This approach provides a valid defense against phishing attacks, eliminates false alarms, and provides better protection against cyber threats. [14]

Introducing a new method, Hybrid Ensemble Feature Selection (HEFS), which is supported by the Cumulative Distribution Function gradient (CDF-g) algorithm to upgrade phishing detection. A proper selection of key indicators can be obtained due to heuristic selection, such as URL features. Thus, both precision and performance can be improved. An experiment, incorporating different models of machine learning, like Decision Trees, Random Forest, and Neural Networks, proves that HEFS has good better detection ability with less false alarms than others. The current study thus recommends a high-level system for identifying loss in financial and data security. [15]

3. METHODOLOGY

1. Dataset and Data Preprocessing

 Dataset: Kaggle will be used as the source for this study's dataset. The dataset will comprise 11,000+ URLs with 33 features by which these URLs can be classified as either phishing or not. **Data Preprocessing:**

- **No null handling of missing values:** The missing and null values within the dataset were removed. o **Data Transformation**: The data set was transformed into a feature vector for the algorithms to process. The transformations, such as encoding, were not required due to most features being categorical.
- **Dataset Split:** The data was split 70 / 30 to train and test as an approach to model evaluation

2. Feature Selection

- The paper applies the Canopy Clustering technique of feature selection. This facilitates a decrease in the dimensionality of this dataset while retaining only the most prominent features useful in detecting phishing.
- The Canopy Feature Selection method is very efficient in pruning the dataset by highlighting the most important features, enhancing the performance of the classifiers.

3. Applied Machine Learning Algorithms

The study uses a combination of several machine learning models to classify phishing URLs, each offering unique strengths:

A) Decision Tree (DT)

A tree-based classifier that recursively partitions the dataset into subsets according to feature values is computationally efficient and well-suited for the large range of features that the Decision Tree technique creates a tree structure for categorization of the URLs. Depending upon whether the URLs have special characters or their length, each node of the tree makes decisions. Using entropy and information gain, the decision tree classifies the data into as many sets as possible. It continues to choose the attribute that offers maximum information gain at each split till it reaches a decision at the leaf nodes.

B) Support Vector Machine (SVM)

A supervised learning model that looks for the best hyperplane that can separate two classes in an ndimensional space: phishing and legitimate ones. SVM uses a hyperplane in a high-dimensional space to differentiate between phishing and authentic URLs. The dataset is a binary classification problem that will have legitimate or phishing, making SVM the best choice in coming up with an optimal border that maximizes the distance between the two classes. As mentioned in the feature, the SVM model maps every URL onto a point in an n-dimensional space. It takes the kernel approach when the separation happens to be non-linear. On one side of the hyperplane, all the URLs are labeled as phishing and on the other side are the valid ones.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2497-2502
e-ISSN : 2583-1062
Impact Factor : 7.001

C) Random Forest (RF):

Collect a dataset with both phishing and legitimate URLs; then extract relevant features, like URL length, special characters, and HTTPS usage. Preprocess the data, clean it, and split the data into the training set and the testing set. Train RandomForestClassifier on the training set and then evaluate its performance on the test set in terms of accuracy as well as other metrics. Optionally fine-tune the hyperparameters for better results.

$$x.y = x_1y_1 + x_2y_2 = \sum_{i=1}^{2}(x_iy_i)$$

D) K-Nearest Neighbors (KNN)

A non-parametric technique that classifies data points based on the majority class among the k-nearest neighbors. It is simple but suffers when dealing with big, high-dimensional datasets. KNN is simple to use and also classifies URLs based on their similarity to other URLs in the database. The majority class among the k-nearest neighbors is what the algorithm considers when it makes a forecast. KNN evaluates

$$F(x) = \frac{1}{B}\sum_{i=1}^{B} f_i(x)$$

the feature space distance between URLs in the paper. It describes a URL that is mostly surrounded by recognizable phishing URLs. The authors tweak a hyperparameter, the number of neighbors (k), to maximize efficiency.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$$

E) Hybrid Model(LR+SVM+DT)

The LSD model, a hybrid model that integrates Decision Tree, Support Vector Machine, and Logistic Regression employing both soft and hard voting techniques, is presented in this study. This group method increases the accuracy of categorization.Out of the three classifiers, the hybrid model combines their predictions. While the majority class is selected in hard voting, the probabilities of the classifiers are averaged in soft voting. According to the article, this hybrid model uses the advantages of each separate model to surpass them.

$$P_{\text{final}} = \frac{1}{N}\sum_{i=1}^{N} P_i$$

4. Training and Model Evaluation

- **Cross-Validation**: To test for overfitting, the authors will apply k-fold cross-validation-most probably a 10-fold one-so that the models are good generalizers of completely unseen data and so that one can evaluate stability and performance.
- **Grid Search Hyperparameter Tuning**: This technique used to fine-tune parameters like max_depth in Decision Trees, n_estimators in Random Forest, and also C and gamma in SVMs so that each model may be well-adjusted for its particular kind of task

5. Evaluation Metrics

The models were evaluated using standard classification metrics, including:

- **Accuracy**: The number of correct URL identification, both phishing and legitimate ones.
- **Precision**: Determines the percentage of true phishing URLs among those predicted as phishing..
- **Recall (Sensitivity)**: Measures the model's ability to recall actual phishing URLs.

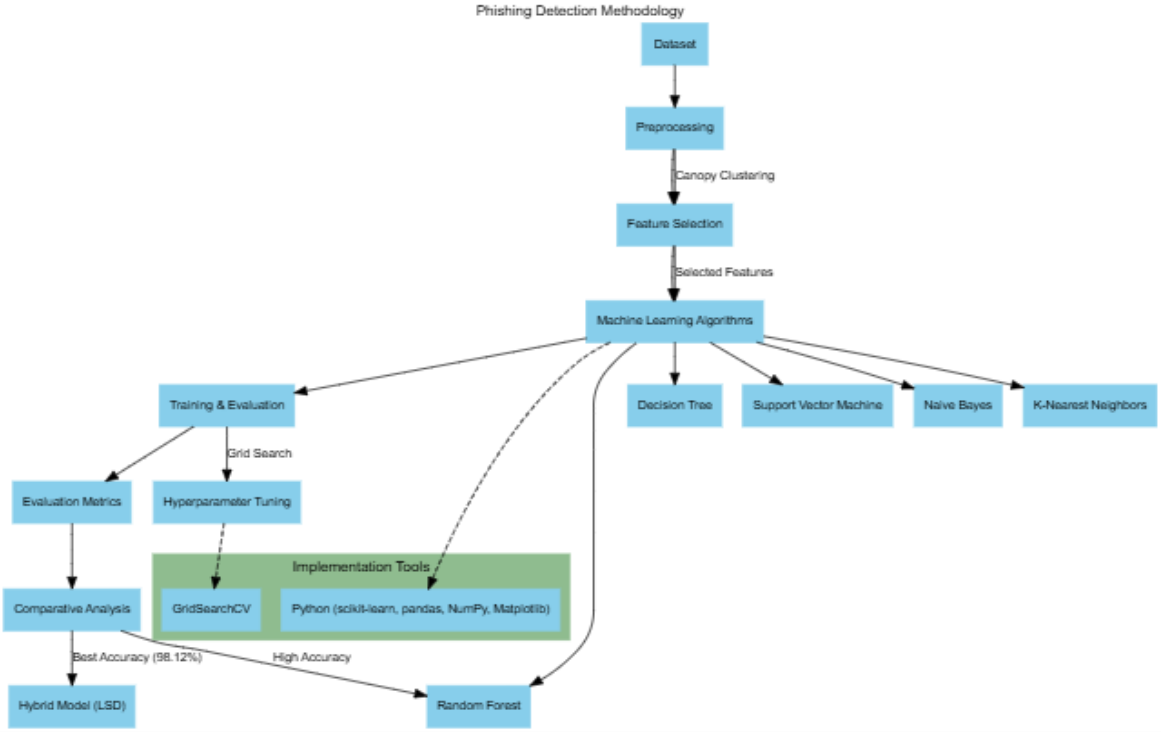
6. Comparative Analysis

All models were compared after training and evaluation. Propose LSD hybrid model showed superior performance with the highest accuracy of 98.12%. The Accuracy for each individual model such as Random Forest, SVM, and Naive Bayes failed compared to the proposed LSD hybrid model.

- The Random Forest model also performed well and achieved accuracy close to the hybrid model.
- In comparative analysis, it proved that ensemble models like Random Forest and the LSD hybrid model prove better than single models since ensemble models can combine the strengths of various classifiers.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2497-2502
e-ISSN : 2583-1062
Impact Factor : 7.001

4. METHODOLOGY WORKFLOW



5. CONCLUSION

Phishing attack sophistication levels are rising and evolving to become a great threat to cybersecurity. Innovative means in detection are required. Conventional methods based on blacklists and rule-based systems fail to keep up with the changing strategies of attackers. This research undertakes to examine how machine learning can considerably enhance the detection of phishing methods through techniques such as Random Forest, Support Vector Machines, and Ensemble Methods. Comparing these techniques with traditional methods, these yield improved detection accuracy along with reduction in false positives.

The importance of feature engineering and extraction is thus emphasized, along with the need for realtime scalable detection systems that can adapt to constantly changing phishing tactics. Tackling such problems could be promisingly led by recent breakthroughs in ensemble methods and deep learning. Making use of large datasets and complex models, machine learning dynamically creates efficient solutions in identifying phishing emails, URLs, and websites.

Future research would focus on improving the quality of the dataset, development of better approaches than the current ones, and bringing up new ideas which are capable of outperforming the hackers. Perhaps deep learning techniques, artificial neural networks, and convolutional neural networks will certainly enhance the resilience of defenses against phishing attacks and make cybersecurity systems even more robust and flexible.

6. REFERENCES

[1] Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S. B., & Joga, S. R. K. (2023). Phishing detection system through hybrid machine learning based on URL. 36805-36822.

[2] Sanchez-Paniagua, M., Fernandez, E. F., Alegre, E., Al-Nabki, W., & Gonzalez-Castro, V. (2022). Phishing URL detection: A real-case scenario through login URLs. 42949-42960. [3] Wei, Y., & Sekiya, Y. (2022). Sufficiency of ensemble machine learning methods for phishing websites detection.124103-124113.

[3] Kara, I., Ok, M., & Ozaday, A. (2022). Characteristics of understanding urls and domain names features: the detection of phishing websites with machine learning methods.124420-124428.

[4] Sahingoz, O. K., Buber, E., & Kugu, E. (2024). Dephides: Deep learning based phishing detection system.

[5] Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2022). A predictive model for phishing detection. Journal of King Saud University-Computer and Information Sciences, 34(2), 232-247. [7] Ahammad, S. H., Kale, S. D., Upadhye, G. D., Pande, S. D., Babu, E. V., Dhumane, A. V., & Bahadur, M. D. K. J. (2022). Phishing URL detection using machine learning methods. Advances in Engineering Software, 173, 103288.

[6] Alani, M. M., & Tawfik, H. (2022). Phishnot: A cloud-based machine-learning approach to phishing url detection. Computer Networks, 218, 109407.

[7] Kabla, A. H. H., Anbar, M., Manickam, S., & Karupayah, S. (2022). Eth-PSD: A machine learning-based phishing scam detection approach in ethereum. IEEE Access, 10, 118043-118057. [10] Ojewumi, T. O., Ogunleye, G. O., Oguntunde, B. O., Folorunsho, O., Fashoto, S. G., & Ogbu, N. J. S. A. (2022). Performance evaluation of machine learning tools for detection of phishing attacks on web pages. Scientific African, 16, e01165.

[8] Kalabarige, L. R., Rao, R. S., Abraham, A., & Gabralla, L. A. (2022). Multilayer stacked ensemble learning model to detect phishing websites. IEEE Access, 10, 79543-79552

[9] Do, N. Q., Selamat, A., Krejcar, O., Herrera-Viedma, E., & Fujita, H. (2022). Deep learning for phishing detection: Taxonomy, current challenges and future directions. Ieee Access, 10, 36429-36463. [13] R. Zieni, L. Massari and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites," in IEEE Access, vol. 11, pp. 18499-18519, 2023

[10] Jayaraj, R., Pushpalatha, A., Sangeetha, K., Kamaleshwar, T., Shree, S. U., & Damodaran, D. (2024). Intrusion detection based on phishing detection with machine learning. Measurement: Sensors, 31, 101003.

[11] M. M. Uddin, K. Arfatul Islam, M. Mamun, V. K. Tiwari and J. Park, "A Comparative Analysis of Machine Learning-Based Website Phishing Detection Using URL Information," 2022 ,220-224.