Phishing Attack Detection Using Hybrid Machine Learning Models and URL Analysis

Ms.Meenakshi Naduvinamani1., Akhila Chilukoori[2]

1 Assistant Professor, Department of CSE, Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India
2, B.Tech CSE (20RG1A0573),
Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

***Abstract*:** To show how the effectiveness of phishing detection models can decrease over time, we trained a baseline model using older datasets and tested it on new URLs. The results suggested declining accuracy; we then carried out an extensive analysis on current phishing domains in order to discover new trends and tactics used by attackers. Creation of a brand new dataset dubbed Phishing Index Login URL-90,000 (PILU-90k) was of utmost necessity in supporting our research. The dataset contains a total of 60,000 legitimate URLs (being index and login pages) and 30,000 were phishing ones.Using this dataset, a Logistic Regression model connected with TF-IDF-feature extraction was built. This model had an impressive accuracy rate in recognizing login URLs, at 98.50%.

Keywords: *Phishing-detection models**,** PILU-90K dataset**,** Logistic Regression, TF-IDF feature extraction, Cybersecurity, URL Classification*

INTRODUCTION

Phishing attacks represent a considerable risk to both individuals and organizations, frequently resulting in financial losses, identity theft, and various forms of cybercrime. Conventional techniques for identifying phishing websites often fall short, as attackers can swiftly generate new URLs and domains to bypass detection.

Machine learning thus provides a viable way to deal with this problem. By examining a wide array of features and behaviors associated with websites, machine learning algorithms can reasonably and proficiently detect phishing attempts. This paper seeks to establish automated and effective methodologies for identifying phishing websites through machine learning, while also addressing the inherent challenges and limitations of this approach.

Normally, phishers create excessive websites meant to mislead the users so that they divulge sensitive information under the false pretense that they view legitimate sites. In a world filled with threats of fraud, machine learning can prove its prowess in identifying any fraudulent website when the characteristics and behavior of a website are analyzed upon. Several methods through which machine learning can aid in the detection of phishing websites include:

1. Feature engineering: Machine learning algorithms can be trained to recognize phishing websites by evaluating various attributes, such as domain names, URL structures, SSL certificates, and content.
2. Natural Language Processing (NLP): Phishing websites frequently contain deceptive content aimed at misleading users. NLP techniques can analyze the content of these websites to identify patterns that are typically associated with phishing attempts.
3. Behavioral analysis: Machine learning algorithms can assess a website's behavior, including its

interactions with users and requests made to external servers, which can reveal patterns commonly associated with phishing sites.
4. Supervised learning: The various machine learning algorithms can be trained on labeled datasets containing both phishing and legitimate sites to enable the algorithm to diagnose fresh phishing sites.

1.2 Types of URL's

Legitimate URLs: These are authentic web addresses belonging to reputable organizations, businesses, or services. Legitimate URLs lead users to genuine websites without any malicious intent. Users can trust these URLs to access legitimate content, such as company homepages, product pages, or informational resources. Examples of legitimate URLs include "www.google.com," "www.amazon.com," or "www.microsoft.com."

Phishing URLs: Phishing URLs are deceptive web addresses created by malicious actors to trick users into divulging sensitive information or performing malicious actions. Phishing URLs often mimic the appearance of legitimate URLs to deceive users into believing they are visiting trusted websites. These URLs lead users to fraudulent web pages designed to steal personal information, such as login credentials, credit card details, or other sensitive data. Examples of phishing URLs include variations like "www.paypal-security.com" (mimicking PayPal) or "www.bankofamerica-login.net" (mimicking Bank of America).

RELATED WORK

In 2019, Loukas et al. proposed a comprehensive taxonomy and survey of cyber-physical intrusion detection approaches for vehicles. They categorized existing techniques based on various criteria, such as target vehicle category, architecture, deployment, and features. This survey provided a valuable overview of the landscape of intrusion detection solutions for vehicles, highlighting their strengths and limitations.

B. Mohandes et al. (2018) advanced the concept of cyber-physical sustainability through their integrated analysis of smart power systems, focusing on electric vehicles. This study emphasized the significance of sustainable practices in the automotive industry, particularly in the context of cybersecurity, as electric vehicles become more prevalent and connected.

L. Pan et al. (2017) explored various cybersecurity attacks targeting modern vehicular systems. Their research concentrated on the detection accuracy of different attack types, providing insights into the vulnerabilities present in vehicular networks and the necessity for improved defense mechanisms.

In 2016, M.J. Kang and J.W. Kang developed an advanced intrusion detection system leveraging deep neural networks for in-vehicle network security. Their approach utilized machine learning techniques to enhance the detection of anomalies, demonstrating the potential of AI-driven solutions in safeguarding automotive systems.

B. Groza and S. Murvay (2013) contributed to the field with their research on secure broadcast communication protocols within

Controller Area Networks (CAN). Their work focused on ensuring secure communication channels among vehicles, addressing the challenges posed by potential cyber threats.

In 2012, A. Monot et al. proposed a method for combining runnable sequencing with task scheduling in multicore automotive ECUs, enhancing the efficiency of automotive systems through optimized task management. C. Ling and D. Feng also addressed security concerns by developing an algorithm for detecting malicious messages on CAN buses, showcasing the necessity of effective message validation mechanisms.

The work of T. Hoppe et al. (2011) analyzed security threats to automotive CAN networks, providing practical examples and countermeasures that can be implemented to enhance network security. H. Oguma et al. (2008) introduced an innovative attestation-based security architecture designed for in-vehicle communication, aimed at ensuring the authenticity and integrity of transmitted data. Lastly, T.Y. Moon et al. (2007) designed a diagnostic gateway system for LIN, CAN, and FlexRay communication protocols, integrating robust diagnostic functions to enhance vehicle reliability and safety.

Disadvantages

• Yet, it is an irrefutable fact that the existing work on phishing URL detection using machine learning still focuses on more complicated datasets for their phases of learning.

• Data unavailability: Such phenomena as too little training data affect the accuracy of virtually any machine learning model.

• Poorly labeled data: One of the most dangerous flaws common to phishing detection models based on machine learning is that the models depend on and learn directly from the available data under the false assumption that those data are correctly labeled, thus affecting the quality of detection.

Year	Author	Problem Statement	Techniques Used	Dataset	Performance Analysis Parameters	Limitations
2019	G. Loukas, E. Karapistoli, E. Panaousis, P. Sarigiannidis, T. Vuong	Cyber-physical intrusion detection approaches for vehicles	Taxonomy and survey	Multiple sources	Accuracy, detection rates	Focused mainly on taxonomies, lacks novel implementation
2018	B. Mohandes, R. Al Hammadi, W. Sanusi, T. Mezher, S. El Khatib	Advancing cyber-physical sustainability in electric vehicles	Integrated analysis of smart power systems	Case study on electric vehicles	Sustainability metrics	Limited to power systems, broader scope not explored
2017	L. Pan, X. Zheng, H.X.	Cyber security attacks on	Cybersecurity	Case study on electric	Attack detection	Focuses mainly on attacks, not on

	Chen, T. Luan, L. Batten	modern vehicular systems	analysis	vehicles	accuracy	defense mechanisms
2016	M.J. Kang, J.W. Kang	Intrusion detection system for in-vehicle networks	Deep neural network-based intrusion detection	Case study on electric vehicles	Accuracy, false-positive rate	Dataset specifics and applicability not discussed
2013	B. Groza, S. Murvay	Secure broadcast communication in CAN networks	Secure broadcast protocols	Case study on electric vehicles	Security, communication overhead	Assumes static topology, limited real-world testing
2012	A. Monot, N. Navet, B. Bavoux, F. Simonot-Lion	Combining runnable sequencing with task scheduling	Runnable sequencing, task scheduling	Case study on electric vehicles	Execution time, resource usage	Scalability for larger systems not discussed
2012	C. Ling, D. Feng	Detection of malicious messages on CAN buses	Detection algorithm	Case study on electric vehicles	Accuracy, false-positive rate	Scalability for larger systems not discussed
2011	T. Hoppe, S. Kiltz, J. Dittmann	Security threats to automotive CAN networks	Security threat analysis	Case study on electric vehicles	Practical examples and selected countermeasures	Limited to short-term countermeasures
2008	H. Oguma, X. Yoshioka, M. Nishikawa, R. Shigetomi, A. Otsuka, H. Imai	New attestation-based security architecture for in-vehicle communication	Attestation-based security architecture	30 participants, 50 phishing URLs	Accuracy, false-positive rate	tilizing resource-efficient KPS and a local Master ECU for verification.
2007	T.Y. Moon, S.H. Seo, J.H. Kim, S.H. Hwang, J. Wook Jeon	Gateway system with diagnostic functions for LIN, CAN, and FlexRay	Gateway system design	30 participants, 50 phishing URLs	Phishing attack success rate, user awareness	Small sample size, limited scope of toolbars

Table.1 Literature survey from 2007to 2019.

PROBLEM STATEMENT

In spite of the emergence of other culturally sensitive, albeit generally unconvincing, prophecies, the nature of the long war waged by phishing still afflicts the

penetration of the methodologies employed by the phishing detection systems. With each advancement made by computer criminals, their attacks evolve, thereby leaving traditional detection systems flatfooted. Therefore, they fail to distinguish between real and fake websites, thereby exposing users to a broad range of phishing scams, which usually almost insufferably lead to theft of sensitive information. Victims bear quite a considerable financial loss and reputation loss; thus, there is an urgent need for much-needed sophisticated and modular detection techniques capable of general and accurate identification of phishing websites, hence protecting user data and enhancing trust in the online space. Innovative solutions are hence paramount for the successful dealing of these advanced threats as and when they arise.

ARCHITECTURE

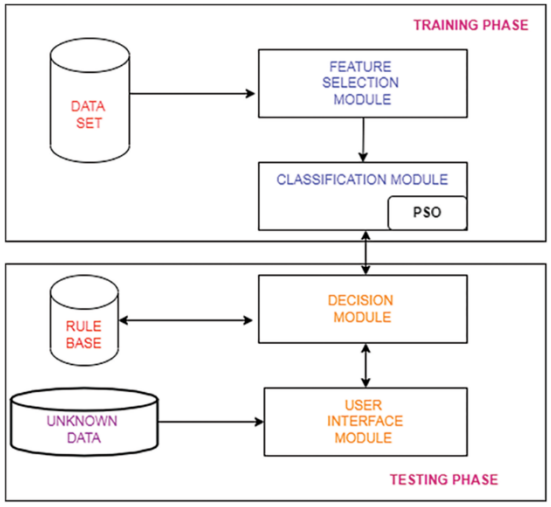

Fig. Architecture of HLSD

ALGORITHM:

Decision Trees are the most widely used supervised learning algorithms for classification and regression tasks. They mechanism is reduction of a dataset through repeated partitioning based on the most Youngest Significance Feature, building tree structures showing the decision at various internal nodal points and the outcome or class label in leaf nodal points. The objective here is maximum class separation or minimum error for regression problems. It has satisfactory interpretability and is relatively simple for visualization, thus gaining a preference as a proper means for explaining prediction. But they are susceptible to overfitting, especially with deep trees that catch noise in their data. Regularization strategies like pruning could minimize this effect.

Naive Bayes is probabilistic classification algorithm based on Bayes' Theorem. It presupposes that the input features of the data are conditional independent, making a simplification while providing solutions within a computationally feasible time frame although sometimes incompatible with real-world conditions. Nevertheless, the simplifying assumption Naive Bayes appears to work well in many applications, ranging from text classification to spam filtering; the independence assumption is reasonable in this instance. This algorithm is fast, efficient on small data sets, and capable of dealing with high dimensions, yet said independence condition on which it rests can sometimes be a disadvantage to its accuracy if the features are highly inter-correlated.

Random Forest builds several decision trees during training and predicts on them by averaging their predictions. During a random construction, several decision trees are caused that make each one stabilize the model's accuracy. Each tree in the forest is trained on a random subset of the data using a feature randomly chosen for splitting, reducing variance and helping to prevent overfitting in otherwise large trees. Random Forests are incredibly adaptable and yield good results on a wide variety of tasks, classifying and regressing.

K-nearest neighbor is a basic algorithm for both classification and regression. It is an instance-based learning algorithm that does not create an explicit model to use while classifying new data points. The new data points are classified according to their similarity with the 'k' closest neighbors in the training data. Usually, this is dictated by the distance metric, most commonly the Euclidean distance. K-nearest neighbors' approach is simple and easy to implement, working quite well when small and well-behaved datasets are concerned. Nevertheless, as the data grows, it is computationally expensive since every point in the data needs its distance calculated for any future prediction. Lastly, it is sensitive not only to the given value of 'k' but also to the feature scaling, both of which can have a dramatic effect on performance.

PROPOSED MODEL:

Phishing URL-based cyberattack detection is an essential area of research for protecting user privacy and averting financial losses. This study proposes a novel approach that effectively identifies phishing URLs using machine learning techniques.

More than 11,000 phishing URLs were used to train and evaluate various machine learning models. The models included decision trees, linear regression, naive Bayes, random forest, gradient boosting machines, support vector classifiers, K-nearest neighbors, and one proposed hybrid model combining linear regression, support vector classifiers, and decision trees with soft and hard voting.

Cross-fold validation and grid search parameter tuning were performed in tandem with canopy feature selection to optimize the model performance of the proposed hybrid model. This offered model parameters that were tuned for maximum accuracy.

The standard performance measures are employed in evaluating the effectiveness of the proposed methodology: accuracy, precision, recall, specificity, and F1-score. These measures provide a holistic view of the model's performance in detecting phishing URLs with reduced false positives and false negatives.

In the final analysis, a promising direction for the detection of phishing URLs using advanced machine learning methods is proposed along with strict evaluation techniques.

METHODOLOGY

1. Development and Analysis of Phishing URL Detection System

Objective: Develop a strong phishing URL detection backed by the dataset of more than 11,000 phishing URL attributes to study which features can best differentiate between a phishing and a genuine URL.

2. Implementation and Optimization of Machine Learning Models

Objective: This experiment will utilize Decision Tree, Linear Regression, Naive Bayes, Random Forest, Gradient Boosting Machine, Support Vector Classifier, K-Neighbors Classifier, and a proposed hybrid model LSD to test its functionality against others. The hybrid model will be optimized by using cross-fold validation and tuning of hyperparameters with a grid search and canopy feature selection to improve prediction accuracy.

3. Assessment of Detection Effectiveness

Objective: To measure the degree of detection effectiveness via the metrics of accuracy, precision, recall, specificity, and F1, maintaining alignment with expected performance criteria among real-world standards.

4. Results Analysis and Provide Feedback for Future Enhancement

Objective: Analyze the results from the detection system so that valid insights and recommendations for bettering the system can be generated and the system will mold itself to new phishing threats and maintain accurate detection rates.

RESULTS

Fig:1-Home page

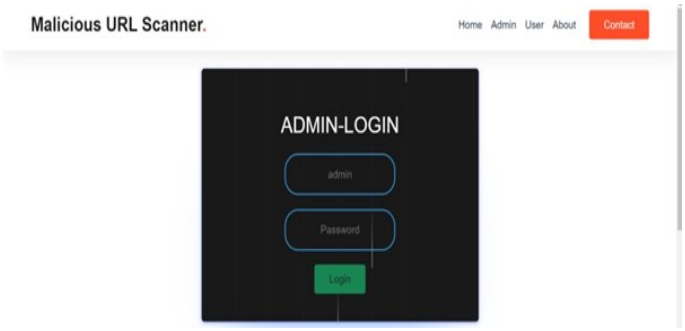

Fig:3-Login Page

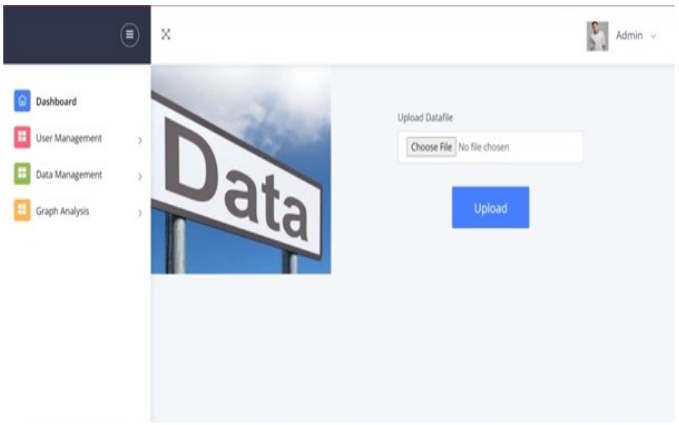

Fig:3-DataUploadPage

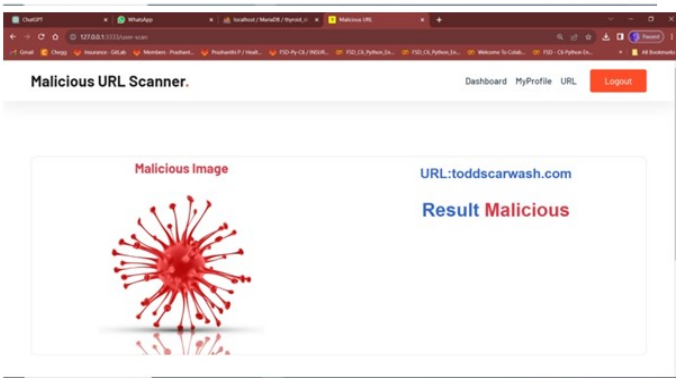

Fig:5 Malicious URL Scanner Page

Fig:4 Basic Info

CONCLUSION

The proposed phishing URL detection system provides directions for solution development towards sustaining the war against cybercrime and, especially protecting user privacy. The system uses deep datasets containing over 11,000 attributes regarding phishing URLs to help ascertain the major characteristics plaint offenders that categorize phishing websites from legitimate ones. By utilizing various machine learning classification algorithms, including sophisticated ones such as the LSD hybrid model, the detection is augmented in respect to accuracy and resistance to determination, respectively.

Employing several optimization approaches in their respective order- such as cross-fold validation and

grid search parameter tuning- the model is able to hit high-performance metrics, enabling real-world utilization in actual applications. Besides that, the evaluation procedure-is based on established indicators of performance, like accuracy, precision, recall, specificity, and F1-score-furnishes a set of reliable lenses to assess any capability of the system.

The analysis of the results opens up new windows for both promotion of the system while assisting as a medium to consider areas needing improvement. As the phishing schemes keep on being developed for the future, the work stresses the need for constant amendments and improvements to be done in detection methodologies. Ultimately, this work contributes to the advancement of the safe online environment that empowers users to browse safely and confidently.

REFERENCES

[1] A. Monot ; N. Navet ; B. Bavoux ; F. Simonot-Lion, "Multisource Software on Multicore Automotive ECUs—Combining Runnable Sequencing With Task Scheduling", IEEE Trans. Industrial Electronics, vol. 59, no. 10. Pp. 3934-3942, 2012.

[2] T.Y. Moon; S.H. Seo; J.H. Kim; S.H. Hwang; J. Wook Jeon, "Gateway system with diagnostic function for LIN, CAN and FlexRay", 2007 International Conference on Control, Automation and Systems, pp. 2844 – 2849, 2007.

[3] B. Groza; S. Murvay, "Efficient Protocols for Secure Broadcast in Controller Area Networks", IEEE Trans. Industrial Informatics, vol. 9, no. 4, pp. 2034-2042, 2013.

[4] B. Mohandes, R. Al Hammadi, W. Sanusi, T. Mezher, S. El Khatib, "Advancing cyber– physical sustainability through integrated analysis of smart power systems: A case study on electric vehicles", International Journal of Critical Infrastructure Protection, vol. 23, pp. 33-48, 2018.

[5] G. Loukas, E. Karapistoli, E. Panaousis, P. Sarigiannidis, T. Vuong, A taxonomy and survey of cyber-physical intrusion detection approaches for vehicles, Ad Hoc Networks, vol. 84, pp. 124-147, 2019.

[6] Hoppe T, Kiltz S, Dittmann J. Security threats to automotive can networks. practical examples and selected short-term countermeasures. Reliab Eng Syst Saf vol. 96, no. 1, pp. 11–25, 2011.

[7] Schulze S, Pukall M, Saake G, Hoppe T, Dittmann J. On the need of data management in automotive systems. In: BTW, vol. 144; pp. 217–26, 2009.

[8] Ling C, Feng D. An algorithm for detection of malicious messages on can buses. 2012 national conference on information technology and computer science. Atlantis Press; 2012.

[9] Oguma H, Yoshioka X, Nishikawa M, Shigetomi R, Otsuka A, Imai H. New attestation based security architecture for in-vehicle communication. In: Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE. IEEE; pp. 1–6, 2008.

[10] L. Pan, X. Zheng, H. X. Chen, T. Luan, L. Batten, "Cyber security attacks to modern vehicular systems", Journal of Information Security and Applications, vol. 36, pp. 90-100, October 2017.

[11] Kang, M. J., & Kang, J. W., "Intrusion detection system using deep neural network for in-vehicle network security", PloS one, vol. 11, no. 6, e0155781, 2016.

[12] Theissler, A., "Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection", Knowledge-Based Systems, vol. 123, pp. 163-173. F. Zhu, J. Yang, C. Gao, S. Xu, T. Yin, "A weighted oneclass support vector machine",