Volume 8, Issue 2, pp. 32-37, 2024.
International Journal of Scientific Engineering and Science
ISSN (Online): 2456-7361

Developing a Phishing Detection System Utilizing a Hybrid Machine Learning Approach Focused on URL Analysis

K Sudhakar[1*], G Lakshmi Pujitha[2], K Venu[3], CH JayaSri[4], A Sadhwika[5], K Jasmine[6]
[1*]Department of Information Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India-534202
[2]Department of Information Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India-534202
[3]Department of Information Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India-534202
[4]Department of Information Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India-534202
[5]Department of Information Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India-534202
[6]Department of Information Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India-534202

Abstract— *Cybercrimes of all kinds are coordinated these days via the internet. Thus, phishing attempts are the main focus of this research. The main technique used in phishing attacks is email distortion. Mock sites are used in conjunction with challenging correspondences to collect the requisite data from the relevant parties. Despite the fact that several research on the prevention, detection, and awareness of phishing assaults have been published, there is still no comprehensive and reliable way for doing so. Machine learning is therefore crucial to the battle against cybercrimes such as phishing. The phishing URL-based dataset, a compilation of phishing and authentic URL attributes gathered from over 11,000 domain datasets, serves as the foundation for the proposed study. After preprocessing, a number of machine learning techniques have been applied to guard against phishing URLs and safeguard users. In order to effectively and accurately defend against phishing attacks, this study makes use of a variety of machine learning models, including decision trees, logistic regression, random forests, naive Bayes, gradient boosting classifiers, K-neighbors' classifiers, and support vector classifiers. Additionally, a hybrid LSD model that combines decision trees, support vector machines, and logistic regression with both soft and hard voting is proposed. The grid searches hyper parameter optimization methodology and the canopy feature selection method with cross-fold validation are used in the suggested LSD model. To further illustrate the impacts and efficacy of the models, an array of evaluation criteria was applied to evaluate the suggested technique. Precision, accuracy, recall, F1-score, and specificity were among these requirements. The results of the comparative analysis demonstrate that the suggested approach outperforms the alternative models and gives the best results.*

Keywords— *Protocol, cyber security, social networks, logistic regression, support vector machines, unified resource locators (URL), ensemble classifiers, voting classifiers, logistic regression, logistic regression, and decision trees (LSD).*

I. INTRODUCTION

The internet is essential to many facets of daily life. A network of computers connected via phone lines, fiber optic lines, wireless connections, satellite connections, and other types of telecommunication's linkages makes up the Internet. Many different kinds of cybercrimes are currently coordinated online. Therefore, the primary focus of this study is on phishing attacks. Phishing uses email deception as the foundation for deceptive emails, then uses spoof websites to get the necessary information from the targeted individuals. Phishing detection techniques come in a variety of forms, including machine learning and list-based techniques. Additionally, to enhance prediction outcomes, a grid search parameter based on the canopy feature selection technique was employed in cross-fold validation. Phishing assaults are therefore a serious and deadly cybercrime on the internet, and there isn't a perfect solution in place at the moment to stop them. When it comes to protecting against cybercrimes involving phishing assaults, machine learning is essential. Hybrid machine learning algorithms will be employed to precisely identify phishing assaults.

The goal of this project is to apply a hybrid machine learning technique with a URL analysis focus to create a reliable and effective phishing detection system. By utilizing the advantages of several machine learning approaches, the system seeks to improve the precision of phishing website identification. The suggested system endeavors to offer a comprehensive response to the dynamic problems posed by phishing assaults through the integration of diverse features and models, hence enhancing cybersecurity and safeguarding users.

Phishing attacks are still a serious danger to cybersecurity because they take advantage of consumers' confidence in URLs to trick them into disclosing personal information. Because phishing assaults are dynamic, current detection techniques frequently have trouble telling a fraudulent URL apart from a valid one. The urgent need for a strong phishing detection system that leverages the advantages of many machine learning approaches to efficiently evaluate URL features and offer dependable defense against ever-evolving and complex phishing schemes is addressed by this research.

II. LITERATURE REVIEW

Phishing Detection Leveraging Machine Learning and Deep Learning: A Review: Attacks using phishing techniques deceive targets into divulging private information. We investigate deep learning and machine learning models using massive amounts of data to combat them. In order to detect phishing assaults, we provide various deployment choices and talk about models built on various types of data.

With its breadth, the book is an invaluable resource for academics, industry experts, students, and cyber security researchers alike. An Analysis of Phishing Blacklists: Google Safe Browsing, OpenPhish, and PhishTank: Blacklists are essential for shielding internet users from phishing scams. Among other things, blacklists' accuracy, speed, frequency of updates, size, and scope all affect how effective they are. This paper presents a measuring study that examines Google Safe Browsing (GSB), OpenPhish (OP), and PhishTank (PT), three important phishing blacklists. We look into the URLs in these blacklists and their uptake, dropout, typical lifespan, and overlap. In comparison to 12,433 in PT and 3,861 in OP, we find that, on average, GSB has 1.6 million URLs during our 75-day assessment period. It appears that after 5 and 7 days, OP eliminates a sizable number of its URLs; after 21 days, none are left, which may reduce the efficacy of the blacklist. As the amount of time since being blacklisted rises, we see fewer URLs in all three blacklists, indicating that phishing URLs are frequently transient. Because none of the three blacklists implement a one-time-only URL policy, consumers are shielded from returning to phishing websites. We find that a sizable portion of URLs from all three blacklists reemerge one day after being removed; this could indicate that the removal was done too soon or that threats have resurfaced. Ultimately, we find 11,603 distinct URLs with a 12% overlap that are located in both PT and OP. Even with a lower average size, OP was able to identify more than 90% of these overlapping URLs before PT.

Physical Attributes Significant in Preserving the Social Sustainability of the Traditional Malay Settlement:

A traditional settlement is characterized by physical characteristics and a population that maintains everyday customs, skills, and other cultural activities. However, several traditional villages in Malaysia are currently going through significant changes as a result of urbanization and economic development. Therefore, the purpose of this paper is to identify the physical characteristics that are important for maintaining social sustainability in the traditional Malay settlement. The present study employed a qualitative methodology to ascertain the attributes of the customary villages located in Kuala Terengganu. In the three traditional communities under investigation, open spaces, house borders, and street patterns were found to be important as critical elements for the preservation of social interaction. Thus, the study came to the conclusion that maintaining the social sustainability in traditional settlement communities requires careful consideration of methodologies and choices regarding physical attributes and space typology. It might be difficult to identify phishing sites since they seem real to consumers. The phishing sites can also generate the SSL certificate, which is typically used to protect and encrypt communication. Utilizing HTTPS phishing websites may have negative effects on users. Using a cloned website that appears authentic and has an SSL certificate, attackers can simply trick users and steal confidential data. The user's confidence in the "green padlock" and "lock icon" that appear on the browser when connecting to a website over HTTPS may also be weakened as a result of this. I have examined the key components of how attackers exploit SSL certificates on fictitious domain websites in my thesis. Consequently, I have investigated the resilience of a system to automatically identify a phishing website by utilizing the essential characteristics of an SSL certificate. Several machine learning methods that make use of the examined attributes of retrieved SSL certificates are used in the suggested system. For the site category decision-making process, I have opted for the decision tree algorithm due to its excellent performance and transparency in the final model. To determine if a website is authentic or phishing, the algorithm generates a set of judgment rules. Compared to other machine learning classifiers, the suggested classifier has obtained about 97% of correctly categorized examples.A Web API is developed that offers the proposed system's user interface via HTTP service, allowing users to be connected to the system. The decision criteria of the decision tree algorithm are used by the API to verify a single domain as either a legitimate or phishing domain. The evaluation findings demonstrate the Web API system's prospective effectiveness and efficiency.

III. METHODOLOGY

Many researchers have tried to develop tools that shield consumers from cyberattacks by utilizing black lists, white lists, deep learning, and machine learning to stop URL phishing. In earlier research, two categories of phishing detection systems were presented and put into practice: machine-learning-based and list-based phishing identification systems. This section is split into two sections: earlier research based on lists and ones based on machine learning. Blacklists and whitelists are used by list-based phishing identification systems to recognize phishing URLs. Whitelist-based solutions generate secure and trustworthy webpages to generate the necessary data. A suspect website only needs to match the whitelist websites; if it does not appear on the whitelist, the user has deemed it suspicious and potentially dangerous. It is suggested to employ a whitelist-based system that creates a whitelist by tracking and storing the IP address of each website that has a login form where users can enter their personal information. The article talks on machine learning-based phishing detection systems, however list-based systems are mentioned as an earlier method.

1. A single anti-spam service provider provided all of the URLs used for the previous work.
2. Blacklists did not work well to safeguard users in the beginning; the majority of them only identified 20% of phishing attempts made during zero-hour campaigns.
3. These URLs were taken straight out of emails; no additional attack mechanisms were included.
4. 4. Phishing discovered by heuristics did, however, take a while to show up on blacklists.
5. Previous works that doesn’t used any hybrid machine learning models. So it may leads to decreased in the performance of the models.

The phishing and legitimate URL attributes in the proposed study were collected from over 11,000 website datasets and are based on a phishing URL-based dataset that was obtained from a reputable dataset repository. After preprocessing, many machine learning methods have been developed and put to use to prevent phishing. This study uses machine learning models

International Journal of Scientific Engineering and Science
Volume 8, Issue 2, pp. 32-37, 2024. ISSN (Online): 2456-7361

such as decision trees, linear regression, random forests, naive Bayes, gradient boosting classifiers, K-neighbors classifiers, support vector classifiers, and proposed hybrid LSD models that combine decision trees, support vector machines, and logistic regression with both soft and hard voting in order to effectively and accurately defend against phishing attacks. The canopy feature selection technique with cross fold validation and grid search hyper parameter optimization strategies are used in the proposed LSD model.

Benefits:

1. We are employing ensemble-based learning to address the overfitting issue while concentrating on improving the prediction accuracy.
2. By utilizing many models, the prediction is based on the majority of predictions rather than being biased towards one particular model; as a result, the final ensemble prediction is influenced by the predictions from every model.
3. Using a hybrid machine learning model to increase the precision and effectiveness of phishing detection by combining decision tree, logistic regression, and support vector machine methods.
4. Using grid search hyper-parameter optimization techniques and canopy feature selection methods with cross-fold validation to enhance the suggested LSD model's performance.

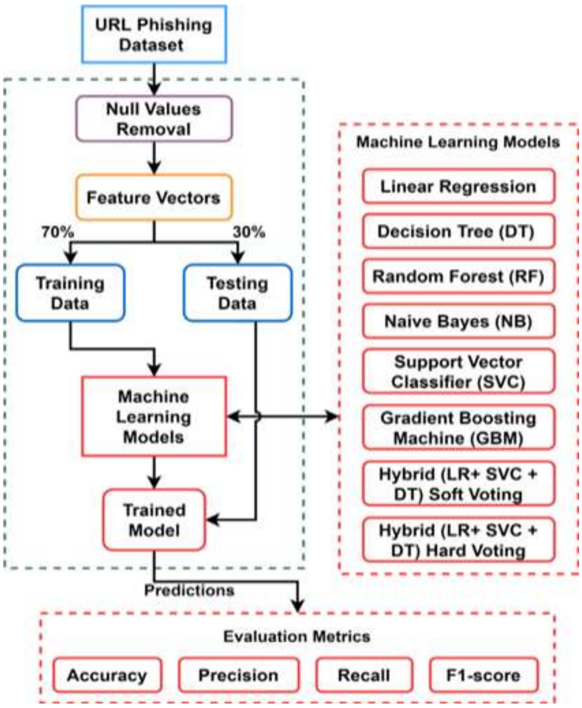

Fig. 1. System Architecture

Modules:

We have created the following modules in order to carry out the aforementioned project.

- Data loading: we will import the dataset by utilizing this module.
- Data Preprocessing: We will examine the data with the help of this module.
- Data division into train and test: This module will separate data into train and test.
- Model creation: Model construction - Support Vector Machine; Random Forest; Decision Tree; Linear Regression (LR) - Bayesian ignorance Gradient boosting, hybrid LSD, soft, hard, and hyperparameter grid CV are all included. - RF + MLP Stacking Classifier with Light GBM. Calculated algorithmic correctness.
- User signup & login: Using this module will result in user registration and login.
- User input: This module provides input for forecasting
- Prediction: final forecast is shown

As an extension, we used an ensemble approach that combined the forecasts from several different individual models to provide a final prediction that was more reliable and accurate.

However, we can further enhance the performance by exploring other ensemble techniques such as Stacking Classifier with RF + MLP with LightGBM which got 100% accuracy.

IV. Implementation

We are using the following algorithms in this project.

LR: Predictive analytics and categorization often use this type of statistical model, often known as a logit model. A dataset of independent variables is used in logistic regression to calculate the probability of an occurrence, like voting or not.

RF: Leo Breiman and Adele Cutler created the well-known machine learning method known as Random Forest, which combines the results of multiple decision trees to get a single conclusion. Its versatility and ease of use have led to its acceptance since it can handle problems related to both regression and classification.

DT: A non-parametric supervised learning approach that is used for both regression and classification applications is the decision tree. With a root node, branches, internal nodes, and leaf nodes, it has a hierarchical tree structure.

SVM: SVM is a powerful supervised technique that works better on smaller, more complicated datasets. Support vector machines, or SVMs for short, are helpful in regression applications as well as classification applications, but their effectiveness is usually highest in the former.

NB: "Naive Bayes" approaches to supervised learning are based on the "naive" assumption of conditional independence between each pair of features, given the value of the class variable. The Bayes theorem is used in these algorithms.

GB: Gradient boosting is a popular boosting method in machine learning for regression and classification issues. An example of an ensemble learning strategy is "boosting," where a model is trained iteratively, with each iteration aiming to outperform the previous one. Many ineffective learners become effective ones as a result of it.

Hybrid LSD Soft: This machine learning algorithm combines the advantages of two distinct algorithms, the soft algorithm and the Locally Sensitive Discriminant analysis (LSD). While the Soft algorithm does well when processing noisy data, the LSD algorithm excels at identifying patterns in data. As a result, the Hybrid LSD Soft algorithm is an effective tool for data analysis since it can identify patterns in noisy data.

IJSES
International Journal of Scientific Engineering and Science
Volume 8, Issue 2, pp. 32-37, 2024.
ISSN (Online): 2456-7361

Hybrid LSD Hard: Combining the advantages of the Hard and Locally Sensitive Discriminant (LSD) algorithms, the Hybrid LSD Hard algorithm is an optimization technique. While the Hard algorithm performs well when handling data that is devoid of noise, the LSD algorithm is good at finding patterns in data. The Hybrid LSD Hard method, which combines these two techniques, is an effective tool for data analysis since it can effectively identify patterns in noise-free data.

LSD with Hyperparameter grid cv: LSD with Hyperparameter grid cv is a model selection technique for Locally Sensitive Discriminant Analysis (LSD). It searches for the best combination of parameters by trying different values and evaluating performance through cross-validation. This optimizes LSD's effectiveness, leading to better model accuracy and generalizability.

Stacking Classifier: A stacking classifier is an ensemble learning technique that builds a single "super" model by combining several classification models. As a result, performance can frequently be enhanced because the merged model can benefit from each unique model's advantages.

V. Experimental Results

Dataset:

The "URL-based phishing dataset" was taken from the well-known Kaggle dataset source and utilized in the suggested solution. It is made up of vector-formatted phishing and genuine URLs that were gathered from more than 11,000 websites.

```
data = pd.read_csv("archive/phishing.csv")
data.head()
```

	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedirection
0	0	1	1	1	1	1	-1	0	1	-1	...	1	1
1	1	1	0	1	1	1	-1	-1	-1	-1	...	1	1
2	2	1	0	1	1	1	-1	-1	-1	1	...	1	1
3	3	1	0	-1	1	1	-1	1	1	-1	...	-1	1
4	4	-1	0	-1	1	-1	-1	1	1	-1	...	1	1

5 rows × 32 columns

Fig. 2. Dataset

	ML Model	Accuracy	f1_score	Recall	Precision	Specificity
0	Linear Regression	0.934	0.941	0.943	0.927	0.909
1	Support Vector Machine	0.951	0.957	0.969	0.947	0.909
2	Naive Bayes Classifier	0.605	0.454	0.292	0.957	0.909
3	Decision Tree	0.957	0.962	0.991	0.960	0.909
4	Random Forest	0.969	0.972	0.993	0.990	0.909
5	Gradient Boosting Classifier	0.974	0.977	0.994	0.986	0.909
6	Hybrid LSD - SOFT	0.959	0.964	0.977	0.965	0.909
7	Hybrid LSD - HARD	0.950	0.956	0.967	0.946	0.909
8	Hybrid LSD	1.000	1.000	1.000	1.000	0.426
9	Stacking Classifier	1.000	1.000	1.000	1.000	0.426

Fig. 3. Performance Evaluation

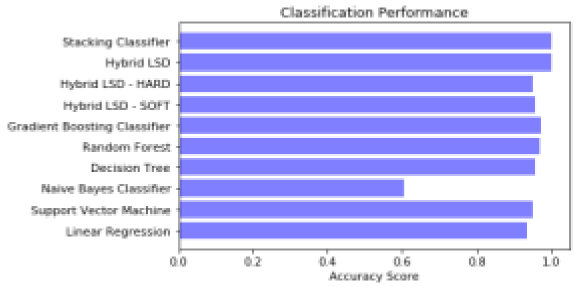

Fig. 4. A comparison graph of all algorithms' accuracy

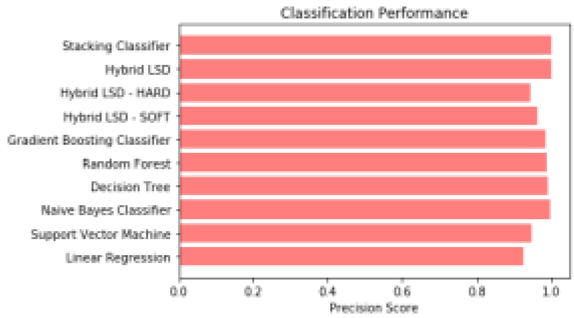

Fig 5. Accurate comparison chart for every method

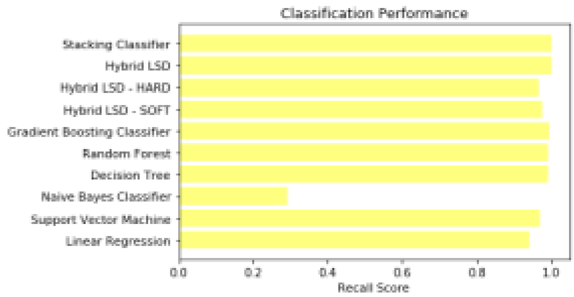

Fig. 6. All algorithms' recall comparison graph

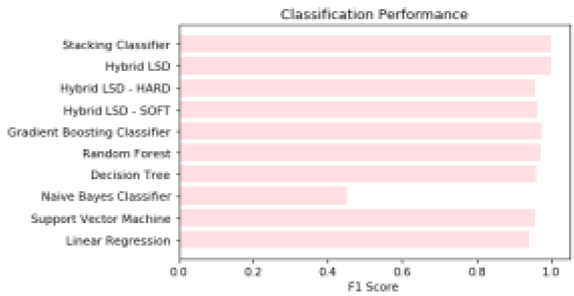

Fig. 7. All algorithms' F1-Score comparison graph

Fig. 8. Home page

International Journal of Scientific Engineering and Science
Volume 8, Issue 2, pp. 32-37, 2024. ISSN (Online): 2456-7361

Fig. 9. Signup page

Fig. 10. Signin page

Fig. 11. Main page

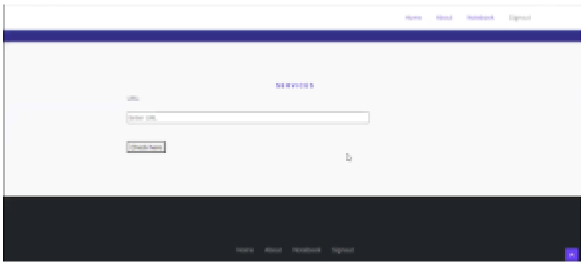

Fig. 12. Upload URL

Fig. 13. Prediction result

Fig. 14. search

Fig. 15. Prediction result for another URL

Similarly, we can check another URLs.

VI. Conclusion

The study comes to the conclusion that phishing assaults are a serious and hazardous cybercrime on the internet, and that there isn't a perfect solution available right now to stop them. When it comes to protecting against cybercrimes involving phishing assaults, machine learning is essential. Thus, Decision trees, linear regression, random forests, naive Bayes, gradient boosting classifiers, K-neighbors classifiers, support vector classifiers, and a suggested hybrid LSD model are just a few of the machine learning models that are used in this suggested method. Combining decision trees, logistic regression, and support vector machines with both soft and hard voting results in the LSD model. Therefore, the best results are obtained by the hybrid machine learning technique based phishing detection system that is proposed. It performs better than other models. Future plans call for creating a phishing detection system that operates in real time, researching the application of deep learning techniques to the task, investigating the use of additional features for the purpose, testing the suggested methodology on a larger and more varied dataset, and creating an intuitive user interface for the system.

References

[1] N. Z. Harun, N. Jaffar, and P. S. J. Kassim, *''Physical attributes significant in preserving the social sustainability of the traditional malay settlement,''* in Reframing the Vernacular: Politics, Semiotics, and Representation. Springer, 2020, pp. 225–238.

[2] D. M. Divakaran and A. Oest, ''Phishing detection leveraging machine learning and deep learning: A review,'' 2022, arXiv:2205.07411.

[3] A. Akanchha, ''Exploring a robust machine learning classifier for detecting phishing domains using SSL certificates,'' Fac. Comput. Sci., Dalhousie Univ., Halifax, NS, Canada, Tech. Rep. 10222/78875, 2020.

[4] H. Shahriar and S. Nimmagadda, ''Network intrusion detection for TCP/IP packets with machine learning techniques,'' in Machine Intelligence and Big Data Analytics for Cybersecurity Applications. Cham, Switzerland: Springer, 2020, pp. 231–247.

[5] J. Kline, E. Oakes, and P. Barford, ''A URL-based analysis of WWW structure and dynamics,'' in Proc. Netw. Traffic Meas. Anal. Conf. (TMA), Jun. 2019, p. 800.

[6] A. K. Murthy and Suresha, ''XML URL classification based on their semantic structure orientation for web mining applications,'' Proc. Comput. Sci., vol. 46, pp. 143–150, Jan. 2015.

[7] A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam, ''Phishing website detection: An improved accuracy through feature selection and ensemble learning,'' Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 252–257, 2019.

[8] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, *''PhishAri: Automatic realtime phishing detection on Twitter,''* in Proc. eCrime Res. Summit, Oct. 2012, pp. 1–12.

[9] S. N. Foley, D. Gollmann, and E. Snekkenes, Computer Security—ESORICS 2017, vol. 10492. Oslo, Norway: Springer, Sep. 2017.

[10] P. George and P. Vinod, ''Composite email features for spam identification,'' in Cyber Security. Singapore: Springer, 2018, pp. 281–289.
[11] H. S. Hota, A. K. Shrivas, and R. Hota, ''An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique,'' Proc. Comput. Sci., vol. 132, pp. 900–907, Jan. 2018.
[12] G. Sonowal and K. S. Kuppusamy, ''PhiDMA—A phishing detection model with multi-filter approach,'' J. King Saud Univ., Comput. Inf. Sci., vol. 32, no. 1, pp. 99–112, Jan. 2020.
[13] M. Zouina and B. Outtaj, ''A novel lightweight URL phishing detection system using SVM and similarity index,'' Hum.-Centric Comput. Inf. Sci., vol. 7, no. 1, p. 17, Jun. 2017.