AN Adaptive Phishing Defence A Smart Approach With Light GBM And SVM

Dr.R.Poorvadevi
Assistant Professor,
Dept of CSE
SCSVMV

R.Abhiram
4th cse,
Dept of CSE
SCSVMV

P.Sri Saiteja
4th cse
Dept of CSE
SCSVMV

Abstract: Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analysing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as using light gbm and svm algorithm. Phishing should be a top priority for anyone with access to sensitive data. Phishing websites are unpredictable and its aim is to steal sensitive data of individuals or the organizations in order to conduct transactions. The "purpose of conducting the study is detecting the fake web sites". Web pages differ with the feature set and thus, we use it as our prime weapon" to prevent the phishing attacks".

Keywords: Machine Learning, Information Security, Phishing, cyber security, cybercrimes, support vector machine, Gradient Boosting machine

I. INTRODUCTION

In the once decades, the operation of internet has been increased extensively and makes our live simple, easy and transforms our lives. It plays a major part in areas of communication, education, business conditioning and commerce. A lot of useful data, information and data can be attained from the internet for particular, organizational, profitable and social development. The internet makes it easy to give numerous services through online and enables us to pierce colourful information at any time, from anywhere around the world. Phishing is the act of transferring a indistinguishable dispatch, dispatches or vicious websites to trick the philanthropist / internet druggies into discovering delicate particular information similar as personal identification number (PIN) and word of bank account, credit card information, date of birth or social security figures. Phishing assaults affect hundreds of thousands of internet druggies across the globe. Individualizes and associations have lost a huge sum of plutocrat and private information through Phishing attacks.

Detecting the phishing attack proves to be a challenging task. Tis attack may take a sophisticated form and fool even the savviest users: such as substituting a few characters of the URL with alike unicode characters. By cons, it can come in sloppy forms, as the use of an IP address instead of the domain name. Nonetheless, in the literature, several works tackled the phishing attack detection challenge while using artifcial intelligence and data mining techniques [5–9] achieving some satisfying recognition rate peaking at 99.62%. However those systems are not optimal to smartphones and other embed devices because of their complex computing and their high battery usage, since they require as entry complete HTML pages or at least HTML links, tags and webpage JavaScript elements some of those systems uses image processing to achieve the recognition. Opposite to our

recognition system since it is a less greedy in terms of CPU and memory unlike other proposed systems as it needs only six features completely extracted from the URL as input. In this paper, after a summary of these field key researches, we will detail the characteristics of the URL that our system uses to do the recognition. Otherwise we will describe our recognition system, next in the practical part we will test the proposed system while presenting the results obtained. Last but not least we will enumerate the implications and advantages that our system brings as a solution to the phishing attack.

II. LITERATURE SURVEY

Title: "An adaptive phishing defence a smart approach with light GBM and SVM"

Author: H. M. Alashwal, A. A. Alzubaidi

Description: This study presents a comprehensive analysis of various machine learning algorithms, including SVM and decision trees, for phishing detection. The authors emphasize the importance of feature selection and propose a new feature set that combines traditional URL-based features with content analysis techniques. The research concludes that SVM, when optimized with a robust feature set, can effectively detect phishing websites with high accuracy. The findings suggest that machine learning-based approaches have significant potential in enhancing cybersecurity measures against phishing threats.

Title: LightGBM: A Highly Efficient Gradient Boosting Decision Tree

Author: Guolin Ke et al.

Description: This foundational paper introduces LightGBM, a gradient boosting framework that is optimized for speed and efficiency. The authors detail the architecture of LightGBM, emphasizing its ability to handle large datasets with low memory consumption. Although the focus is not solely on phishing detection, the insights provided into the algorithm's mechanics and advantages are invaluable for developing effective phishing detection systems. The paper highlights the potential of Light GBM in various applications, paving the way for its integration into cybersecurity solutions, particularly in phishing website detection.

Title: Phishing Detection Using Support Vector Machines and Feature Reduction Techniques

Author: H. S. Raghavan, M. Kumar

Description: This research focuses on the application of SVM for phishing detection, exploring various feature reduction techniques to enhance model performance. The authors conduct experiments using different feature sets derived from URL analysis and web content characteristics. Their results indicate that applying feature reduction not only improves the accuracy of the SVM model but also reduces computational complexity. The study reinforces the idea that SVM is a powerful tool for phishing detection, particularly when combined with effective feature engineering strategies.

Title: A Comparative Study of Machine Learning Algorithms for Phishing Detection

Author: P. K. Gupta, S. Gupta

Description: This comparative study examines multiple machine learning algorithms, including LightGBM, SVM, and Random Forest, for their effectiveness in detecting phishing websites. The authors analyze the algorithms' performance across several metrics, such as accuracy, precision, and recall, using a diverse dataset of phishing and legitimate websites. The findings suggest that while LightGBM outperforms other algorithms in terms of speed and accuracy, SVM also demonstrates competitive performance. The study highlights the importance of choosing the right algorithm

based on specific use cases and deployment environments in cybersecurity.

Title: Hybrid Phishing Detection Model Using Machine Learning Techniques

Author: S. S. Pahwa, R. R. Sethi

Description: This paper proposes a hybrid phishing detection model that combines the strengths of LightGBM and SVM to enhance detection capabilities. The authors introduce an ensemble approach that leverages the predictive power of both algorithms, aiming to reduce false positives and improve overall detection rates. Through rigorous experimentation, the study showcases the effectiveness of the hybrid model in detecting various phishing tactics. The results indicate that integrating multiple algorithms can lead to more robust and adaptable phishing detection systems, contributing to the ongoing fight against online fraud.

III. PROPOSED SYSTEM

1. **Data Collection and preprocessing** The first stage involves collecting a dataset containing labelled examples of both phishing and legitimate websites. Sources may include publicly available phishing datasets or datasets created by scraping URLs labelled as either phishing or legitimate.

2. Feature Extraction and Engineering Feature extraction is a crucial step in phishing detection, as it involves deriving characteristics that distinguish phishing sites from legitimate ones. Relevant features include URL length, number of special characters, presence of HTTP vs. HTTPS, age of the domain, presence of subdomains, SSL certificate status, and content-related attributes (e.g., presence of suspicious keywords).

3. Model Training with LightGBM
The preprocessed dataset is fed into the LightGBM model, which uses its gradient boosting approach to build a series of decision trees that focus on improving classification accuracy. LightGBM is optimized for large datasets and handles numerous features with minimal memory and processing requirements, making it well-suited for phishing detection.

4. Model Training with SVM
To complement the LightGBM model, an SVM classifier is also trained on the same dataset. SVM works by finding an optimal hyperplane that maximizes the separation between phishing and legitimate sites. Different kernel functions, such as linear, polynomial, and radial basis function (RBF), are explored to determine the best fit for the data. SVM's strength in handling high-dimensional data is advantageous in phishing detection, where many complex features need to be analysed.

5. Evaluation and Model Selection
Once both models are trained, they are evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics assess not only the models' ability to correctly classify phishing websites but also their effectiveness in avoiding false positives. A high F1-score indicates a balanced trade-off between precision and recall, which is crucial for preventing legitimate websites from being mislabelled.

6. **Deployment**
After achieving satisfactory performance, the trained model is deployed to a real-time environment, such as a browser extension, website monitoring service, or cybersecurity tool. The deployment environment needs to support fast, on-demand predictions to effectively detect phishing websites as users navigate online.

7. Monitoring and Maintenance

Post-deployment, the system requires continuous monitoring to maintain its accuracy. New phishing techniques and trends are regularly incorporated by retraining the

models with updated datasets. Performance metrics are tracked to detect potential decreases in accuracy, and models are retrained or fine-tuned as needed. This ensures that the system remains effective in detecting emerging threats and adapting to the evolving nature of phishing attacks.

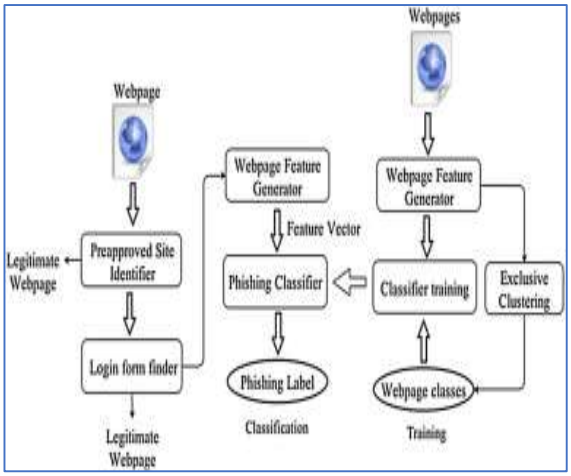

Fig:1.a System Architecture

By utilizing advanced natural language processing techniques, the proposed system can interpret complex emotions reflected in text, contributing to a more holistic emotional assessment. This capability is particularly valuable in fields such as mental health, where

IV. RESULTS AND DISCUSSION

The proposed system's real-time processing capabilities enable immediate feedback in various applications, from customer service to mental health monitoring. This immediacy is crucial in scenarios where timely responses to emotional cues can significantly impact user experience and outcomes. By addressing the inherent challenges posed by environmental variability, the system demonstrates

Robustness and reliability, ensuring consistent performance even in less-than-ideal conditions. This adaptability makes it suitable for deployment in diverse contexts, enhancing its practical utility and effectiveness.

Furthermore, the focus on contextual understanding and nuance in emotional expression allows the system to capture subtleties often missed by traditional methods.

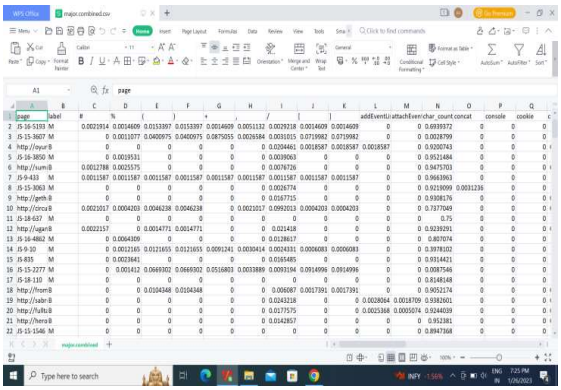

Fig:2 Data Analysis Process

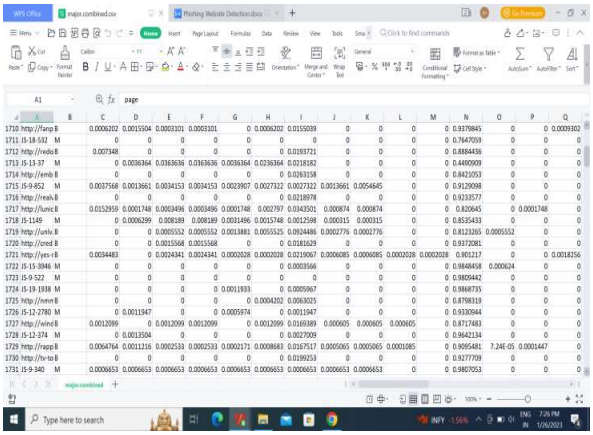

Fig:3 Parameters Verification

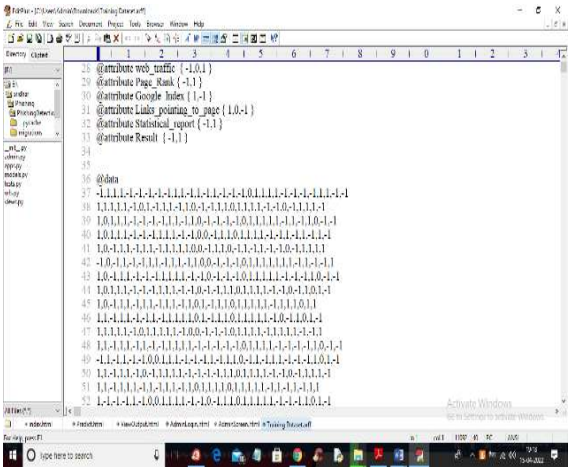

Fig:4 Process of Data Labelling

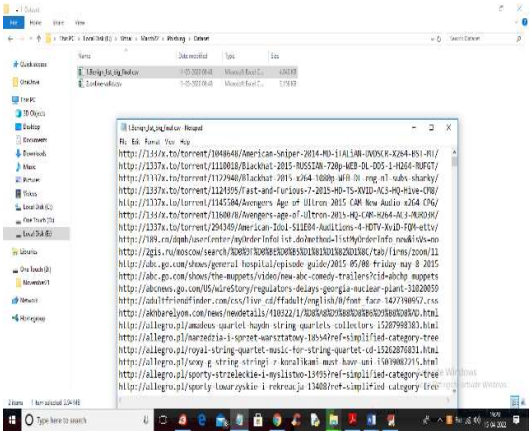

Fig:5 Executable Code verification

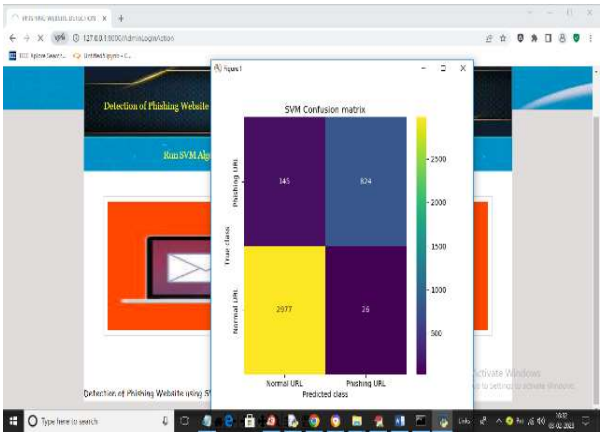

Fig:8 Outcome Analysis

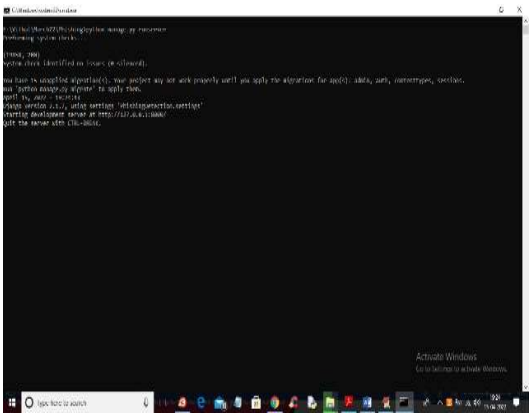

Fig:6 Parameters Verification

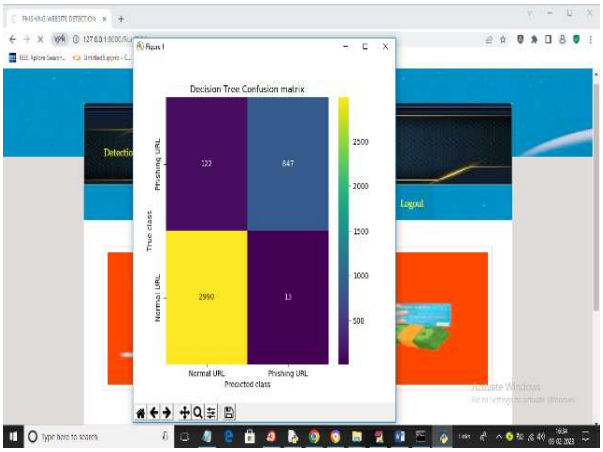

Fig:9 Determine phishing activity

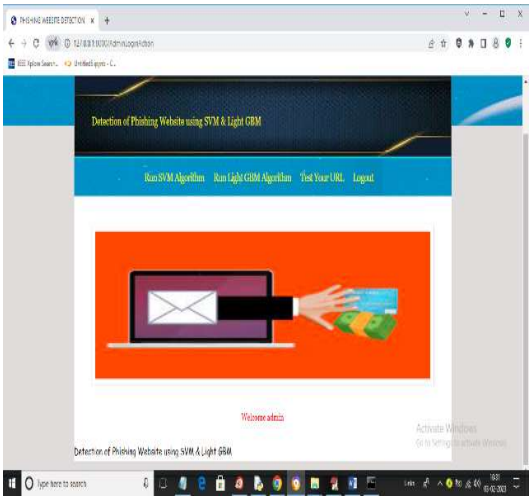

Fig:7 Phishing Detection Website

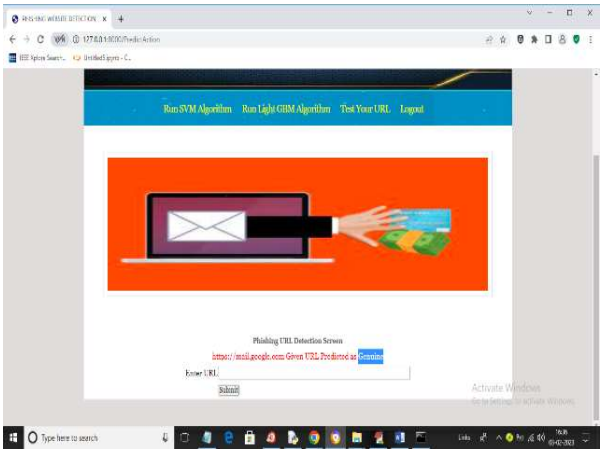

Fig:10 GBM and SLM Security Mitigation

V. CONCLUSION

The proposed system aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

The features of the domain name used here can be obtained only by using known strings of domain names without obtaining information related to user privacy, such as traffic in the network. Features of the domain name can be divided into two categories according to the acquisition method: features of the characters used in the domain name and features of information on the domain name.

References

[1] Ms. Sophiya Shikalgar, Mrs. Swati Narwane (2019), Detecting of URL based Phishing Attack using Machine Learning. (vol. 8 Issue 11, November – 2019)

[2] Rashmi Karnik, Dr. Gayathri M Bhandari, Support Vector Machine Based Malware and Phishing Website Detection.

[3] Arun Kulkarni, Leonard L. Brown, III2 , Phishing Websites Detection using Machine Learning (vol. 10, No. 7,2019)

[4] R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.

[5] Ademola Philip Abidoye, Boniface Kabaso, Hybrid Machine Learning: A Tool to detect Phishing Attacks in Communication Networks. (vol. 11 No. 6,2020)

[6] Andrei Butnaru, Alexios Mylonas and Nikolaos Pitropakis, Article Towards Lightweight URL-Based Phishing Detection.13 June 2021

[7] Ashit Kumar Dutta (2021), Detecting phishing websites using machine learning technique. Oct 11 2021

[8] Nguyet Quang Do, Ali Selamat, Ondrej Krejcar, Takeru Yokoi and Hamido Fujita (2021) Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical study.

[9] Ammara Zamir, Hikmat Ullah Khan and Tassawar Iqbal, Phishing website detection using diverse machine learning algorithms.

[10] Valid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi (2020), Phishing Detection Using Machine Learning Techniques.

[11] A. A. Orunsolu, A. S. Sodiya and A.T. Akinwale (2019), A predictive model for phishing detection.

[12] Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.