Intelligent Detection of Malicious Websites Using Machine Learning

[1]Dr. P. Harini,[2]Vema Pavithra,[3]Tatikonda Venkata Naga Hyma,[4]Polisetty Vamsi Lakshman

[1]Professor & HOD, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.

[2,3,4]B. Tech Student, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.

ABSTRACT

The rapid expansion of internet-based services has significantly increased the number of malicious websites involved in phishing, malware distribution, and online fraud. Traditional blacklist-based and rule-based detection techniques are ineffective against newly generated or zero-day malicious websites. This paper presents an intelligent detection system for malicious websites using machine learning techniques. The proposed approach analyzes website-related features such as URL characteristics and network traffic behavior to classify websites as malicious or legitimate. Supervised machine learning algorithms are trained on extracted features to achieve accurate and automated detection. Experimental results demonstrate that the proposed system provides reliable detection with improved accuracy and reduced false positives.

Keywords:- ***Malicious websites, Machine learning, Cyber security, URL analysis, Network traffic.***

INTRODUCTION

The internet plays a crucial role in modern communication, online transactions, and information sharing. However, the growing dependency on web-based services has also increased exposure to cyber threats. Malicious websites are designed to steal sensitive information, distribute malware, or mislead users through fraudulent activities. Conventional detection techniques rely on blacklists and predefined rules, which are ineffective against newly created malicious websites.Machine learning provides an adaptive solution by learning patterns from historical data and identifying malicious behavior automatically. By analyzing URL structures, traffic patterns, and behavioral features, machine learning models can accurately classify websites. This paper proposes a machine learning-based system for intelligent detection of malicious websites to enhance web security.

LITERATURE SURVEY

Several studies have explored machine learning approaches for malicious website

detection. Researchers have shown that lexical URL features and network traffic characteristics are effective indicators of malicious behavior. Supervised learning models such as Support Vector Machines, Random Forest, and Naïve Bayes have demonstrated high accuracy in classifying malicious websites.Despite promising results, many existing approaches suffer from limitations such as high false-positive rates, limited scalability, and inability to detect evolving attack patterns. These challenges highlight the need for an improved detection mechanism that combines feature analysis with intelligent learning models.

RELATED WORK

Traditional blacklist-based methods are ineffective against newly generated malicious websites. Machine learning techniques have been widely adopted to classify websites using URL and behavioral features. Studies show that combining network traffic analysis with supervised learning improves detection accuracy. Algorithms such as Random Forest and Support Vector Machine are commonly used due to their robustness. These approaches form the basis for intelligent malicious website detection systems.

EXISTING SYSTEM

Existing malicious website detection systems mainly depend on blacklist-based and rule-based approaches. While these systems can identify known threats, they fail to detect zero-day attacks and newly generated malicious URLs. Manual updates and static rules limit adaptability, making these systems inefficient against modern cyber threats.

PROPOSED SYSTEM

The proposed system utilizes supervised machine learning techniques to detect malicious websites intelligently. Website data is collected, and relevant features such as URL length, packet statistics, response time, and traffic behavior are extracted. These features are used to train machine learning classifiers that learn patterns associated with malicious activity.The trained model predicts whether a website is malicious or legitimate, enabling automated and real-time detection. This approach improves detection accuracy and enhances system scalability.

SYSTEM ARCHITECTURE

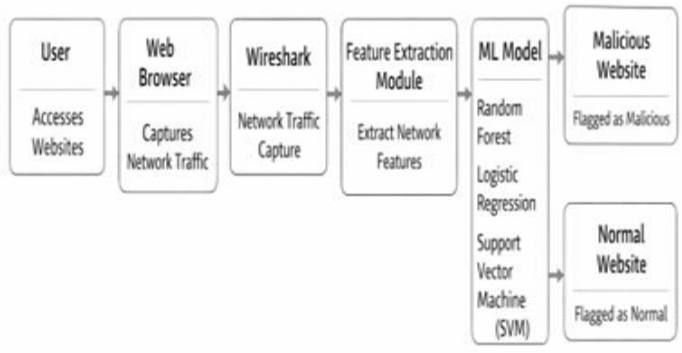

Fig 1: Architecture diagram

METHODOLOGY DESCRIPTION

The system detects malicious websites using machine learning based on network traffic behavior. When a user opens a website, network traffic is captured using the Wireshark tool. Important traffic features such as duration, packet flow, and response time are extracted from the captured data. These features are sent to a Flask-based web application for processing. The extracted values are analyzed using trained machine learning models, including Random Forest, Logistic Regression, and SVM. The model classifies the website as malicious or normal based on traffic patterns. Finally, the prediction result is displayed to the user through the web interface.

RESULTS AND DISCUSSION

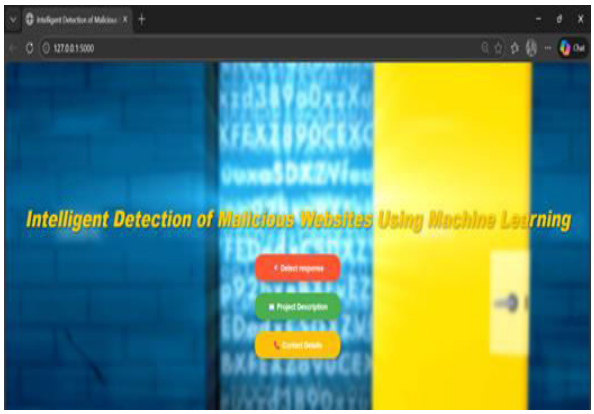

Fig 2: Home Page

This output shows the application's main interface, which enables users to access the contact, project description, and detection modules. It attests to the web-based system's successful deployment.

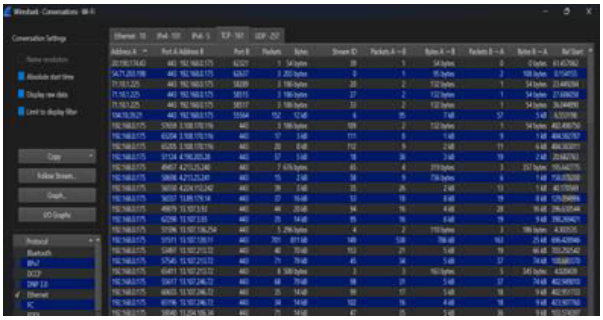

Fig 3: Wire Shark Traffic Screenshot

The network traffic that is recorded when a website is accessed using the Wireshark tool. The screenshot shows packet-level information that is used as input features for the detection of malicious websites, such as data flow, packet length, and response timing. This collected traffic is the main source of data for the machine learning model and aids in the analysis of website behavior.

Fig 4: Network Traffic Values

This screen shows the input form where users enter traffic parameters such as duration, flow bytes, packet length, and response time captured from Wireshark. It validates the interaction between the user and the backend system.

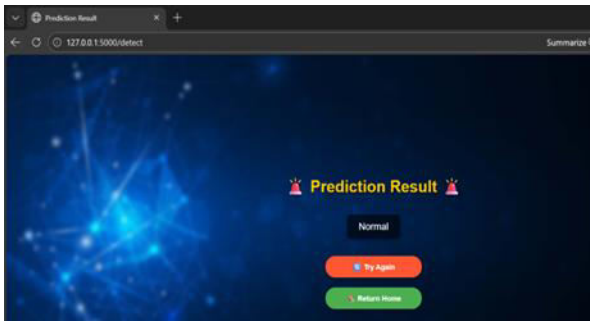

Fig 5: Result-Normal Website Detection

This screen shows the classification of a website as **"Normal"** when the traffic behavior matches legitimate browsing patterns. It demonstrates the model's ability to distinguish normal traffic from malicious activity.

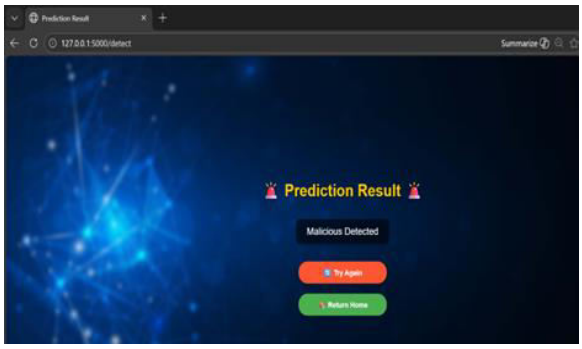

Fig6: Result-Malicious Website Detection

CONCLUSION

This project offers a clever method that uses network traffic analysis and machine learning to identify malicious websites. The system successfully distinguishes between malicious and legitimate websites by gathering real-time traffic data using Wireshark and examining behavioral characteristics. The Random Forest algorithm outperformed other models in terms of accuracy and malicious traffic detection, according to experimental results, demonstrating the dependability and efficacy of the suggested approach.

FUTURE SCOPE

By incorporating real-time traffic capture straight from Wireshark to automate data extraction, the system can be improved. CNNs and other sophisticated deep learning models can be used to increase the detection accuracy of intricate attack patterns. Furthermore, real-time security and scalability can be achieved by implementing the system as a cloud-based service or browser extension.

REFERENCES

[1] Harini, D. P. (2013). Two Level Intrusion Detection For Detecting Intruders in Multitier Web Applications. *International Journal of Engineering & Science Research*, *3*, 472-478.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[3] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in *Proc. ACM SIGKDD*, pp. 1245–1254, 2009.

[4] M. Aburrous, T. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing website detection using random forest classifier," *Int. J. Advanced Computer Science and*

Applications, vol. 8, no. 1, pp. 180–186, 2017.

[5] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler: A fast filter for the large-scale detection of malicious web pages," in *Proc. World Wide Web Conf.*, pp. 197–206, 2014.

[6] S. Singh and R. Kaur, "Network traffic based malicious website detection using machine learning," *Journal of Information Security*, vol. 11, no. 3, pp. 150–162, 2020.

[7] H. Zhang and G. Liu, "A survey of malicious websites detection techniques," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2094–2121, 2018.

[8] M. Al-Fawa'reh, S. A. Rashaideh, and A. Aljawarneh, "Detection of malicious websites using machine learning techniques," *Security and Communication Networks*, vol. 2019, pp. 1–10, 2019.

[9] R. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Computing and Applications*, vol. 25, no. 2, pp. 443–458, 2018.

[10] D. Sahoo, C. Liu, and S. C. Hoi, "Malicious URL detection using machine learning: A survey," *ACM Computing Surveys*, vol. 52, no. 3, pp. 1–36, 2019.

[11] M. Shafiq, X. Yu, A. A. Laghari, and L. Yao, "Network traffic classification using machine learning techniques," *Journal of Network and Computer Applications*, vol. 90, pp. 1–16, 2016.

[12] Y. Zhao, Y. Chen, and Z. Wang, "Detection of malicious web traffic using machine learning," *IEEE Access*, vol. 8, pp. 104199–104210, 2020.

[13] J. Kim and S. Lee, "Cyber threat detection based on network traffic behavior," *Computers & Security*, vol. 92, pp. 101–112, 2020.

[14] K. Srinivasan and S. Ramachandran, "Behavior-based malicious website detection using traffic features," *International Journal of Cyber Security*, vol. 3, no. 2, pp. 45–53, 2021.

[15] A. Buczak and E. Guven, "A survey of data mining and machine learning methods for cybersecurity intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[16] R. Verma and A. Das, "What works and what does not: A study of phishing detection," in *Proc. IEEE Int. Conf. Data Mining*, pp. 1–10, 2017.

[17] T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification," *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, pp. 56–76, 2008.

[18] J. Saxe and K. Berlin, "Deep neural network based malware detection using two dimensional binary program features," in *Proc. IEEE Malware Conf.*, pp. 1–8, 2015.

[19] Y. Zhang, X. Luo, and S. Yu, "Detection of malicious websites based on traffic analysis," *Future Generation*

Computer Systems, vol. 102, pp. 481–489, 2019.

[20] A. Kumar and T. Lim, "Early detection of malicious websites using machine learning," *International Journal of Computer Networks & Communications*, vol. 10, no. 5, pp. 25–36, 2018.