

Enhanced Multi-Threat Detection Classification using Light GBM, Cat Boost and Random Forest Models

Mr.Saggurthi Ramesh¹, Kurmala Pradeep Gupta², Challa Nidesh³ and Kalidindi Yoga Narasimha Gupta⁴

¹Assistant Professor, Department of Artificial Intelligence and Data Science, Lakireddy Balireddy College of Engineering, Mylavaram, Andhra Pradesh, India

²⁻⁴Students, Department of Artificial Intelligence and Data Science, Lakireddy Balireddy College of Engineering, Mylavaram, Andhra Pradesh, India

Email: rameshsaggurthi1245@gmail.com, kpgupta240@gmail.com, nideshchalla2003@gmail.com, yogikalidindi2003@gmail.com

Abstract—Effectively identifying and categorizing dangerous URLs has become essential for cyber security defense due to the rise in malicious online activity. In order to classify multi-threat URLs as safe, phishing, or malware-related, this study suggests a machine learning-based method. We created a classifier with a strong 96% accuracy rate using ensemble learning models, namely Light GBM, Cat Boost, and Random Forest. Each model's capacity to handle unbalanced data, adjust to various threat types, and provide reliable detection accuracy was assessed. Our method highlights the advantages of ensemble approaches in managing intricate threat classifications by optimizing feature selection and hyper parameters. The viability of implementing these effective models in actual cyber security systems to successfully counter URL-based threats is demonstrated by this study.

Index Terms— Url-Detection, Machine Learning Model, Light GBM, Cat Boost, Random Forest, Cyber Security, Deep Learning, Deep neural networks.

I. INTRODUCTION

The swift development of web technology has led to an unparalleled surge in cyber dangers. Because malicious URLs make it easier for assaults like phishing, virus distribution, and data theft to occur, they present a serious threat to internet security. These attacks cause significant monetary and personal losses by taking advantage of flaws in human behavior and web security systems. Because of their inability to keep up with the complexity of contemporary cyberthreats, traditional rule-based methods for URL threat identification frequently result in high false-positive rates and little flexibility.

Machine learning (ML) techniques have become popular as a more flexible and successful threat detection method in response to these issues. ML models are better equipped to identify new threats by identifying intricate patterns and differentiating between safe and dangerous URLs..

Because of their capacity to manage big, heterogeneous datasets and generate reliable predictions, ensemble models such as LightGBM, CatBoost, and Random Forest stand out among machine learning algorithms. Ensemble learning techniques strengthen generalization capabilities, decrease overfitting, and improve classification accuracy by combining the advantages of several algorithms.

Using the LightGBM, CatBoost, and Random Forest algorithms, this study offers a multi-threat classification

model for identifying secure URLs, phishing efforts, and malware distribution websites.

These models were chosen due to their accuracy, speed, and ability to handle unbalanced data in challenging classification tasks. To guarantee ideal model performance, our approach comprised intensive feature engineering, hyper parameter optimization, and performance assessment. With a 96% accuracy rate, the model has a great chance of being used in the actual world.

II. LITERATURE SURVEY

ZheMing Zhang[1] These methods have the benefit of being able to learn from big datasets and spot intricate patterns that conventional approaches would over look. The use of machine learning, more especially LightGBM (Light Gradient Boosting Machine), for Windows malware detection is investigated by Zhang (2023). In contrast to conventional techniques that depend on preset rules or signatures, LightGBM uses feature engineering to enhance model performance and process data effectively. The gradient boosting framework LightGBM is well-known for its effectiveness and speed, which makes it appropriate for extensive virus detection jobs. The main statistic Zhang uses to assess the LightGBM model's performance is AUC-ROC (Area Under the Receiver Operating Characteristic Curve), a common way to gauge the effectiveness of classification models. LightGBM outperformed other widely used algorithms with the greatest AUC-ROC score of 0.684, according to the trial data. According to the experimental data, LightGBM outperformed other well-known models including CatBoost (0.638) and XGBoost (0.677), achieving the highest AUC-ROC score of 0.684.

Gradient boosting techniques like XGBoost and CatBoost, which are renowned for their effectiveness in handling big datasets and complicated features, have also been investigated in earlier malware detection research, such as those conducted by Sun et al. (2022) and Zhang et al. (2021). Zhang's work advances this field by showing that LightGBM can perform better than conventional models in terms of AUC-ROC score, underscoring its promise for quicker and more precise malware detection.

Amanuel Assefa and Rahul Katarya[2] This paper primarily focuses on deep learning techniques to detect phishing websites. Specifically, it utilizes feature extraction to analyze the characteristics of URLs and a neural network-based approach for classification. The paper also compares different classifiers such as Decision trees, Support Vector Machines (SVM) and Auto Encoders. Among these classifiers, the Auto Encoders gives the high accuracy of 91.24%.

Mohammad Nazmul Alam, Dhiman Sarma, Farzana Firoz Lima, Ishita Saha, Rubaiath-E-Ulfath, Sohrab Hossain[3] According to studies, models can "learn" common phishing traits by leveraging datasets with phishing patterns, like those found on websites like Kaggle. In instance, RF models have shown great efficacy, detecting phishing assaults with an accuracy of up to 97%. Building strong cybersecurity defenses will require integrating machine learning (ML) with threat intelligence and behavioural analytics as phishing methods change. In order to stay up with new phishing tactics and create a safer online environment for both consumers and businesses, future research attempts to improve these models' adaptability.

III. PROPOSED MODEL

The proposed model for multi-threat URL classification is designed to detect and categorize URLs into four distinct categories: Safe URLs, Phishing URLs, and Malware URLs, Defacement URLs. The model uses state-of-the-art ensemble machine learning techniques to improve the accuracy and reliability of URL classification.

In this study, we present a machine learning-based approach to the classification of multi-threat URLs. The model's purpose is to categorize URLs into four groups: malicious URLs (including those that spread malware), phishing URLs, and safe URLs, defacement URLs. To increase the model's robustness and classification accuracy, we make use of ensemble learning strategies, including LightGBM, CatBoost, and Random Forest.

A. Input Data: Raw URLs

Raw URLs—web addresses that require classification—are first fed into the model. These URLs can be found in a number of places, such as network traffic records, email links, user input, and web crawlers. Effectively classifying URLs is essential to preventing malware infections and phishing attempts, as they are frequently the initial point of attack in cyber threats.

B. Data Preprocessing and Feature Extraction

To extract useful features, URLs must be analyzed before machine learning models are used. A number of components seen in raw URLs may be signs of malevolent intent. We frequently extract the following characteristics from URLs:

- *URL Length*: The URL's total character count. Given that phishing efforts sometimes employ lengthy, obfuscated strings to deceive consumers, longer URLs may be a sign of an attempt.
- *Number of Subdomains*: In order to look authentic, malicious URLs frequently employ a number of subdomains. Phishing websites, for example, could trick customers by using subdomains.
- *IP Address Presence*: Because hackers frequently employ IP addresses for malicious sites to evade detection by conventional domain-based filtering systems, URLs that contain IP addresses instead of domain names may be harmful. May suggest efforts to conceal the website's actual purpose, a tactic frequently employed in malware and phishing URLs.
- *Domain Name*: A URL's authenticity is frequently strongly suggested by the domain name itself. Similar-looking domains, such as might be used by malicious URLs to trick consumers. Regular expressions, string parsing, and domain analysis approaches are used to extract these properties. After that, the URL is converted into a structured format that machine learning algorithms may utilize.

C. Experimental Setup

This project's experimental setup compares and assesses the performance of three machine learning models for multi-threat URL classification: LightGBM, CatBoost, and Random Forest. This section describes the experimental design, the data collected, the tools and libraries used, and the assessment measures.

LightGBM, CatBoost, and Random Forest are the three machine learning models used for this project. These models were selected due to their resilience in classification tasks, their shown effectiveness in managing high-dimensional features, and their capacity to handle imbalanced data.

D. Light Gradient Boosting Machine (LightGBM)

One gradient boosting model that is well-known for its effectiveness with big datasets is LightGBM. By concentrating on the leaves that have the most potential to lessen the loss, it constructs decision trees in a leaf-wise fashion. This enables LightGBM to outperform conventional gradient boosting models in terms of speed while capturing intricate patterns in the data. Additionally, LightGBM has native support for categorical features, which is useful for managing the categorical data included in URLs, like domain names.

E. CatBoost

Another gradient boosting approach that works well with categorical features is called CatBoost. It works especially well with datasets that contain categorical variables, such as URL domains, when one-hot encoding might not be as effective. Even with tiny or noisy datasets, CatBoost's use of ordered boosting minimizes overfitting and increases the model's resilience. In cybersecurity, where enemies could try to alter attributes to avoid detection, this is very crucial.

F. Random Forest

Several decision trees are built using random subsets of the data in Random Forest, an ensemble learning technique based on bagging. The forecasts from each individual tree are averaged to determine the model's final prediction.

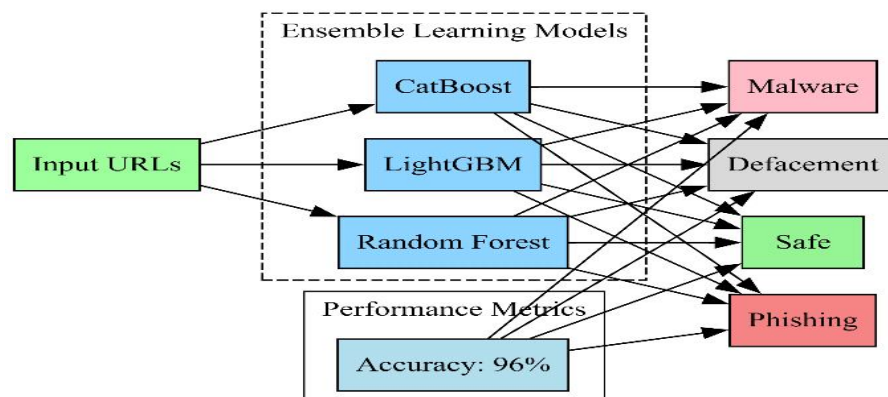


Fig 1: Ensemble Models for Classifying URL Threats

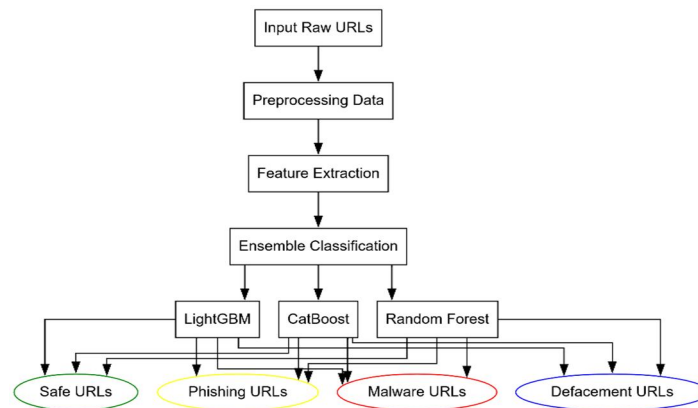


Fig 2: Ensemble Models For Classifying URL Threats

G. Model Architecture Diagram

Figure 3.2: Architecture of the Ensemble models for classifying URL Threats

H. Utilizing ML flow Model Integration

We were able to deliver the precise versions of the models that outperformed the others in our tests by integrating our models from ML flow's Model Registry into Streamlit.

Direct Model Access: Our Streamlit app pulls the chosen LightGBM model for predictions via ML flow's serving capabilities, guaranteeing dependable and consistent outcomes that meet the high-performance criterion of our project.

Logging and Feedback: We may log user input data or improve model predictions over time with ML flow's tracking, creating a feedback loop for even more model enhancement.

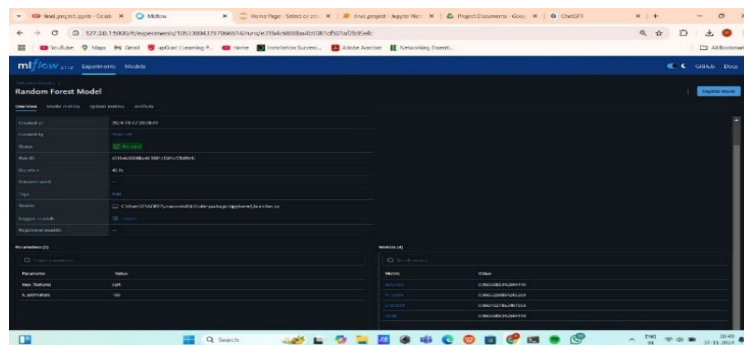


Fig 3. ML Flows Webpage For Model Enhancement

IV. RESULTS

Our project's main goal was to develop a machine learning model that could correctly identify URLs as malware, phishing, or safe. By means of meticulous model selection, hyper parameter adjustment, and performance assessment, we were able to attain noteworthy outcomes concerning classification accuracy, precision, recall, and overall resilience. Here, we highlight our main conclusions and results.

A. Comparison of Model Performance

The performance of three machine learning algorithms Light GBM, CatBoost, and Random Forest in classifying multi-threat URLs was assessed. The performance of each model is summed up by the following metrics.

LightGBM demonstrated strong precision and recall scores for both malware and phishing detection, with the best accuracy of 96.0%. It was the best model for this assignment because of its leaf-wise growth, which made it easier to handle intricate patterns in the data.

With 95% accuracy, CatBoost showed balanced performance, specializing in categorical characteristics that

helped detect phishing attempts by identifying patterns in URLs based on keywords. With an accuracy of 96%, Random Forest consistently performed, albeit marginally worse. Although its precision and recall scores were marginally lower than those of the boosting models, it demonstrated resilience when handling noisy data. Real-Time Predictions: The models' quick and accurate classifications made them appropriate for threat analysis in real time.

Model	Accuracy	Precision (Phishing)	Recall (Phishing)	Precision (Malware)	Recall (Malware)	F1 Score (Overall)
LightGBM	96.0%	95.5%	96.2%	95.8%	96.0%	96.0%
Random Forest	96.0%	95.2%	95.9%	95.5%	95.8%	96.0%
CatBoost	95.0%	94.7%	95.0%	94.6%	94.9%	95.0%

Fig. 4: Overall Accuracy For Enhanced Model

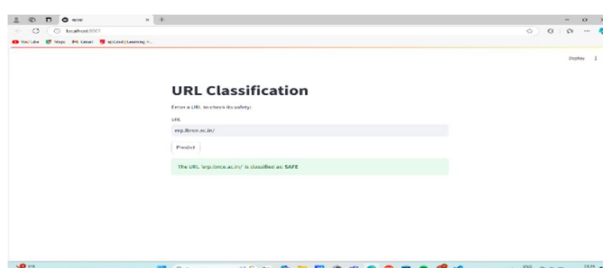


Fig. 5: URL classification With Safe URLs

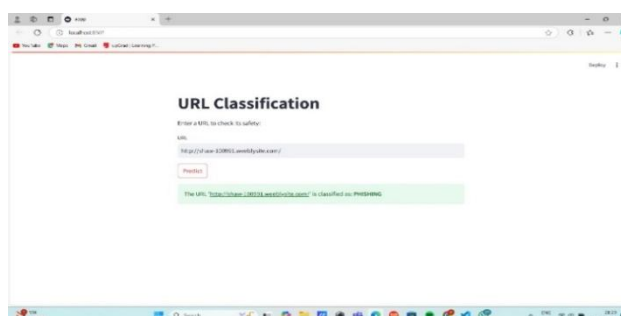


Fig. 6: URL classification With Phishing URLs

It has been determined that this URL is a phishing effort. To steal your information, it can be attempting to mimic a trust Deceptive aspects used in phishing URLs are frequently intended to get private information, such as account information or passwords. We advise against clicking on this link.

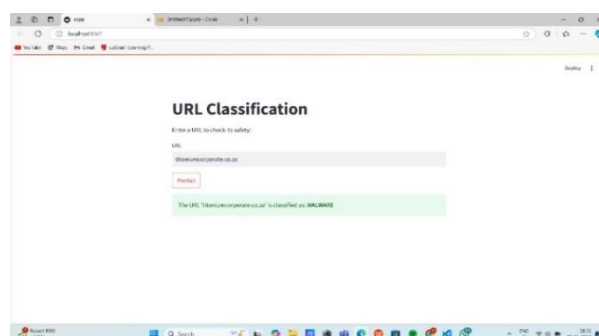


Fig 7: URL classification with Malware and Defacement URLs

Users are strongly warned about the high level of risk by this image. By encouraging quick avoidance of the URL, the image shields viewers from possible danger. This URL has been flagged as possibly containing malware. Clicking on this link may jeopardize the security of your device. Malware URLs frequently result in dangerous downloads or hacked websites that have the potential to infect your system. Stay away from this URL to protect your data and device.

V. CONCLUSION

Our research offers a strong and intuitive method for detecting URL threats in real time by utilizing cutting-edge machine learning algorithms. It efficiently detects phishing, malware, and defacement URLs with a high degree of accuracy and an interactive platform, making people's online experiences safer. It will become much more effective with future developments like ensemble learning, real-time data integration, and wider cybersecurity applications. This technology can be extremely helpful in bolstering cybersecurity defences and increasing awareness of online risks by increasing its functionality and accessibility.

In addition to enhancing technology, this effort is essential in increasing public awareness of the dangers associated with online security. It helps create a safer online environment by teaching users safe browsing techniques and offering real-time security. All things considered, this project represents a major advancement in AI-powered cybersecurity solutions that provide proactive protection against the increasing number of online threats.

FUTURE SCOPE

It will have a greater impact if the technology is extended to identify other cybersecurity threats, like malware-infected files, social media frauds, and email phishing. Additionally, scalability and accessibility for both individual users and companies will be guaranteed by implementing the concept as a cloud-based service. The implementation of explainable AI can improve the system's decision-making transparency and trustworthiness, and adaptive learning strategies can assist the model in adapting to shifting assault patterns. Additional tools to promote safer online behavior include a browser plugin for real-time URL inspection and an integrated cybersecurity awareness education module.

REFERENCES

- [1] Srushti Patil, and Sudhir Dhage, "A Methodical Overview On Phishing Detection Along With An Organized Way To Construct an Anti-Phishing Framework", 2019 5th International Conference On Advanced Computing & Communication System(ICACCS), pp. 1-6.
- [2] Huaping Yuan, Xu Chen, Yukun Li, Zhenguo Yang and Wenyin Liu, "Detecting Phishing Websites and Targets Based On URLs and Webpage Links", 2018 24th International Conference on Pattern Recognition(ICPR) Beijing, China, August 20-24, 2018.
- [3] Vaibhav Patil, Pritesh Thakkar, Cjirag Shah, Tushar Bhat, Prof. S. P.Godse, "Detection and Prevention of Phishing Websites using Machine Learning Approach", 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [4] M. Sameen, K. Han, and S. O. Hwang, "PhishHaven—An efficient real-time AI phishing URLs detection system," IEEE Access, vol. 8, pp. 83425–83443, 2020.
- [5] G. Zhang, Y. Liu, and X. Jin, "A survey of autoencoderbased recommender systems," Frontiers of Computer Science, vol. 14, no. 2. Higher Education Press, pp. 430–450, Apr. 01, 2020. doi: 10.1007/s11704-018-8052-6.
- [6] Asha S Manek, D K Shamini, Veena H Bhat, P Deepa Shenoy, M. Chandra Mohan, K R Venugopal, L M Patnaik. "ReP-ETD: A Repetitive Preprocessing technique for Embedded Text Detection from images in spam emails", 2014 IEEE International Advance Computing Conference (IACC), 2014
- [7] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions," in Proceedings of the 28th international conference on Human factors in computing systems, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 373–382.
- [8] Niakanlahiji, A., Chu, B.-T., & Al-Shaer, E. (2018). PhishMon: A Machine Learning Framework for Detecting Phishing Webpages. 2018 IEEE International Conference on Intelligence and Security Informatics (ISI).
- [9] Gutierrez, C., Kim, T., Della Corte, R., Avery, J., Cinque, M., Goldwasser, D., & Bagchi, S. (2018). Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks. IEEE Transactions on Dependable and Secure Computing.
- [10] Eduardo Benavides, Walter Fuertes, Sandra Sanchez & Manuel Sanchez(2019). "Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review". Smart Innovation, Systems and Technologies, vol 152. Springer, Singapore

- [11] M. A. U. H. Tahir, S. Asghar, A. Zafar, and S. Gillani, "A Hybrid Model to Detect 76 Phishing-Sites Using Supervised Learning Algorithms," in 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 2016, pp. 1126-1133.
- [12] M. Almseidin, A. M. Abu Zuraiq, M. Al-kasassbeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," International Journal of Interactive MobileTechnologies, vol. 13, no. 12, pp. 71–183, 2019,doi: 10.3991/ijim.v13i12.11411.
- [13] T. Shibahara, K. Yamanishi, Y. Takata, D. Chiba, M. Akiyama, T. Yagi, Y. Ohsita, and M. Murata, "Malicious URL sequence detection using event de-noising convolutional neural network," in Proc. IEEE Int. Conf. Commun. (ICC), May 2017, pp. 1-7.
- [14] Fette, I., Sadeh, N.M., Tomasic, A. "Learning to detect phishing emails." In Proceedings of the 16th International Conference on World Wide Web (WWW'07), May 2017.