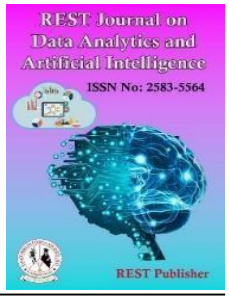
REST Journal on Data Analytics and Artificial Intelligence
Vol: 3(3), September 2024
REST Publisher; ISSN: 2583-5564
Website: http: //restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/3/3/16



Ensemble-Based Dynamic Phishing Domain Detection in Web Browsing Extensions

Ganesh R, *Johnvich Daniel Jacob N, Harisudhan S, S. Ananthi
St. Joseph's College of Engineering Chennai, India
*Corresponding author: johnvichdaniel@gmail.com

Abstract: *In the continually changing world of cybersecurity threats, this study aims to improve the effectiveness of dynamic phishing domain identification within online surfing extensions. Phishing attacks use deceptive tactics to exploit vulnerabilities introduced by popular browser extensions. Recognizing the importance of adaptive and robust defences, our study conducts a thorough review of several ensemble strategies to determine the most effective strategy. Ensemble approaches, well-known for their ability to combine several models, pro- vide a strategic response to the dynamic and ever-changing nature of phishing attempts. The primary goal of this study is to examine and compare several ensemble approaches, such as bagging, boosting, random forest, stacking, ensemble of ensembles, and gradient boosting. Accuracy, flexibility, computing efficiency, and interpretability are all carefully assessed to provide a complete picture of each technique's strengths and weaknesses. Our findings not only shed light on the complex intricacies of ensemble approaches, but also provide a practical guidance for selecting and implementing appropriate models for dynamic phishing domain detection. The study emphasizes the need of adaptive cybersecurity solutions in combating the constant evolution of cyber threats. We hope to harden web surfing extensions against phishing attackers' complex strategies, resulting in a more secure online experience. In conclusion, this study advances cybersecurity practices by providing actionable insights into ensemble-based techniques for dynamic phishing domain identification. As the digital landscape continues to present new difficulties, our work strives to provide cybersecurity practitioners with effective tools and techniques for building a resilient defence against the ever-changing threat landscape.*

1. INTRODUCTION

Phishing is a misleading cyber assault that uses humantrust to fool people into disclosing sensitive information such as usernames, passwords, or financial data. Phishing attacks, which are typically carried out via bogus emails, messages,or websites that impersonate legitimate companies, use social engineering techniques to trick consumers into disclosing sensitive information. The attackers frequently pose as trust- worthy entities to create a false sense of security, abusing human psychology and banking on the intrinsic desire to trust familiar sources. These initiatives evolve in response to technical improvements, making them difficult to detect and combat using traditional cybersecurity methods. The compli- cated dance between cybersecurity defenders and the covert criminals directing phishing assaults highlights the importanceof adaptive and robust defense methods. Traditional methods of detecting phishing domains fre- quently struggle to keep up with the dynamic nature of these attacks, whereas deep learning, while powerful, confronts difficulties in recognizing the subtle changes inside phishing URLs over time. This study aims to strengthen the digital de- fense line by offering an ensemble-based approach to dynamicphishing domain identification within web browser extensions.As we delve into the complex world of phishing attempts and their ever-changing methodologies, our methodology seeks to strategically use the collective intelligence of ensemblelearning methods. By doing so, we hope to not only identify and neutralizethe complex methods used by cyber adversaries, but also to seamlessly integrate these advanced protections into con- sumers' daily online experiences. The next sections explain thetechnical complexities of our research, providing a thorough examination of the symbiotic link between phishing attempts and novel cybersecurity measures.

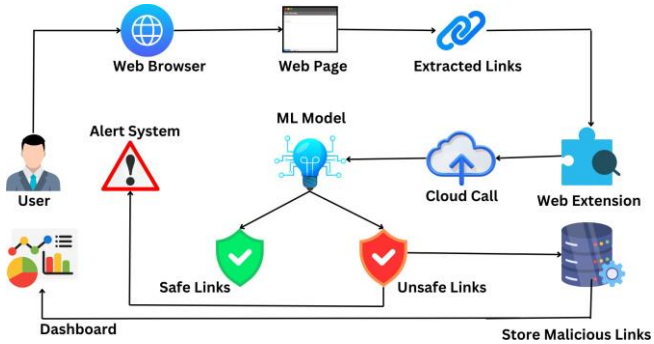


FIGURE 1. The Approach of Phishing Domain Detection using Web Extension

Furthermore, to strengthen the security against phishing as- saults, our suggested ensemble-based method for web browser extensions includes a robust alarm system. This technology runs in the background, continuously evaluating the URLs users interact with in real time. When the warning system detects a potentially harmful link, it instantly alerts the user via a user-friendly and non-intrusive interface. Visual indications, such as pop-up notifications or color-coded indicators in the browser, immediately indicate a potential hazard. These alerts are intended to be both informational and actionable, allowing users to make educated decisions regarding their online active-ties. When a user interacts with a URL, the extension launches a cloud call, which sends the URL's information to the cloud-based ensemble model. This data includes URL-derived features such as domain characteristics, historical context, and potential red flags found during the feature selection process. After receiving the cloud call, the ensemble model uses its collective intelligence to quickly assess the URL as benign or potentially harmful. The cloud-based execution enables the model to adjust in real time to emerging threats, harnessing the cloud server's computational power for fast and accurate evaluations.

2. ENSEMBLE APPROACH FOR PHISHING DOMAIN DETCTION

Dataset Overview and Pre-processing:

Dataset Information: The dataset used in this study includes 11,054 occurrences, each with 32 features. The dataset was gathered from Kaggle and is specifically curated for phishing detection, providing a diverse set of features toaid in robust model training and evaluation.

Feature Exploration: The features include a variety of URL-related properties, such as structural traits, security indicators, and behavioural patterns. This variability enables theensemble model to capture the subtle subtleties of phishing URLs and develop a thorough awareness of potential risks.

```
1   UsingIP                  11054 non-null  int64
2   LongURL                  11054 non-null  int64
3   ShortURL                 11054 non-null  int64
4   Symbol@                  11054 non-null  int64
5   Redirecting//            11054 non-null  int64
6   PrefixSuffix-            11054 non-null  int64
7   SubDomains               11054 non-null  int64
8   HTTPS                    11054 non-null  int64
9   DomainRegLen             11054 non-null  int64
10  Favicon                  11054 non-null  int64
11  NonStdPort               11054 non-null  int64
12  HTTPSDomainURL           11054 non-null  int64
13  RequestURL               11054 non-null  int64
14  AnchorURL                11054 non-null  int64
15  LinksInScriptTags        11054 non-null  int64
16  ServerFormHandler        11054 non-null  int64
17  InfoEmail                11054 non-null  int64
18  AbnormalURL              11054 non-null  int64
19  WebsiteForwarding        11054 non-null  int64
20  StatusBarCust            11054 non-null  int64
21  DisableRightClick        11054 non-null  int64
22  UsingPopupWindow         11054 non-null  int64
23  IframeRedirection        11054 non-null  int64
24  AgeofDomain              11054 non-null  int64
25  DNSRecording             11054 non-null  int64
26  WebsiteTraffic           11054 non-null  int64
27  PageRank                 11054 non-null  int64
28  GoogleIndex              11054 non-null  int64
29  LinksPointingToPage      11054 non-null  int64
30  StatsReport              11054 non-null  int64
```

FIGURE 2. Features to classify the domains

Pre-processing: Data pre-processing is a crucial step to ensure the dataset's readiness for machine learning models. Checking Missing Data, the dataset was examined for missing values, and fortunately, no missing data was found, facilitatingsmooth processing.

Exploratory Data Analysis (EDA): In statistical aspect, Exploratory data analysis (EDA) develops as an

important methodology for evaluating datasets and elucidating their key characteristics, frequently using visual techniques. While statistical models may be used, the essence of EDA is de- riving insights from data that go beyond the scope of formal modelling or hypothesis testing tasks

3. BUILDING MACHINE LEARNING MODELS

FIGURE 3. Overview of Classification Approach

Random Forest: Random Forest is a powerful and adaptable machine learning method that fits into the ensemble learning paradigm. This approach generates an ensemble of decision trees, each trained on a random portion of the dataset and making individual predictions. Random Forest improves accuracy and generalization performance by aggregating these predictions using a voting mechanism, which typically requires a majority vote for classification tasks. Random Forest is distinguished by its intrinsic capacity to handle complicated relationships within data, making it especiallyuseful for applications such as phishing URL identification. It introduces randomization by selecting a subset of features for each tree, hence increasing diversity and reducing overfitting. Furthermore, the use of bagging, which involves training each tree on a bootstrapped sample of data, improves robustness. Feature significance is another important part of Random Forest because it provides information about the relevanceof various attributes. Random Forest is still a popular

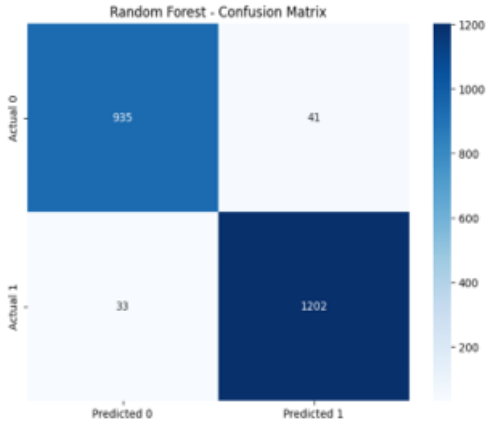


FIGURE 4. Confusion Metrics for Random Forest

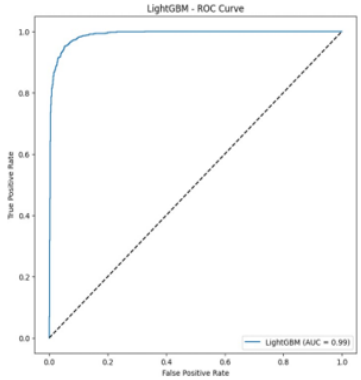


FIGURE 5. ROC Curve for Random Forest

tree on a bootstrapped sample of data, improves robustness. Feature significance is another important part of Random Forest because it provides information about the relevance of various attributes. Random Forest is still a popular choice in many machine learning applications due to its simplicity, ability to handle categorical and numerical input, and resistance to overfitting. Its Python version, which makes use of tools such as scikit-learn, provides an easy-to-use but powerful tool for developing predictive models.

XG Boost (Extreme Gradient Boosting): XG Boost is a versatile and strong machine learning technique that has shown effective in a variety of applications, including phishing URL identification. As an implementation of the gradient boosting framework, XG Boost successively constructs an ensemble of weak learners, often decision trees, to repeatedly repair errors produced by previous models.

XG Boost stands out for its ability to handle complicated data relationships, deliver high predicted accuracy, and quickly manage both numerical and categorical features. It reduces overfitting thanks to a strong regularization framework, and its hyper parameter tweaking flexibility enables practitioners to fine-tune models for specific purposes. XG Boost's parallel processing capabilities, distributed computing support, and optimized speed make it ideal for huge datasets.

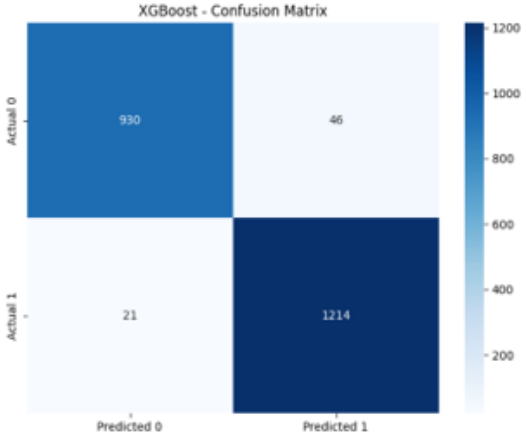


FIGURE 6. Confusion Metrics for XG Boost

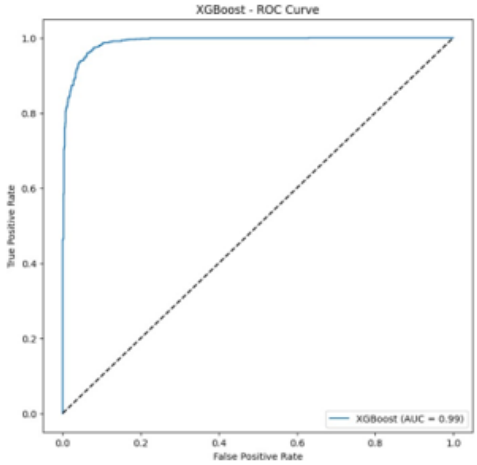


FIGURE 7. ROC curve for XG Boost

Light GBM: Light GBM, or Light Gradient Boosting Machine, is a fast and efficient open-source gradient boosting framework that can handle large-scale machine learning applications with lightning speed and performance. It employs the gradient boosting framework, in which weak learners, often decision trees, are

incrementally added to employs the gradient boosting framework, in which weak learners, often decision trees, are incrementally added to

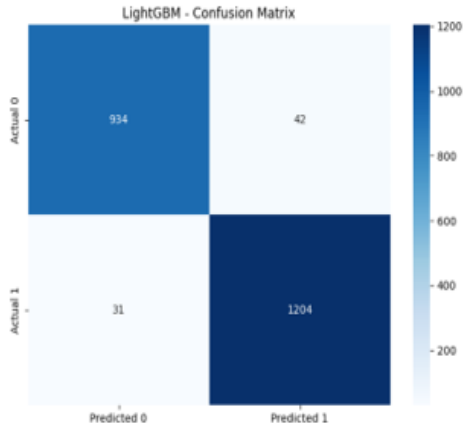


FIGURE 8. Confusion Metrics for Light GBM

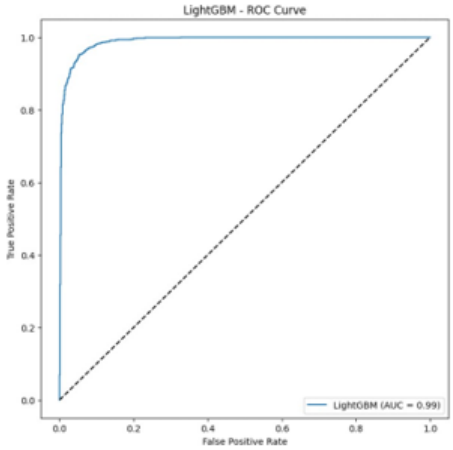


FIGURE 9. ROC for Light GBM

the ensemble to repair errors committed by prior models. What distinguishes LightGBM is its novel leaf-wise tree development technique, which constructs trees by selecting the leaf with the highest delta loss, resulting in faster convergence and lower computational complexity. This framework excels in handling categorical features thanks to its" Gradient-Based One-Side Sampling" technique, and it uses a histogram-based learning approach to ensure efficient computation. LightGBM is designed for speed and scalability, making it lightweight and ideal for huge datasets and distributed computing. It uses regularization approaches to avoid overfitting and provides a variety of hyper parameters for fine-tuning. Its Python implementation is basic, and its ability to balance speed, accuracy, and scalability makes it a popular choice for a wide range of machine learning applications, including phishing URL detection.

Cat Boost: Cata boost is a versatile machine learning method that excels at handling categorical information, making it ideal for applications like as phishing domain detection. Cat Boost's capacity to process category information, such as domain names and top-level domains (TLDs), is useful in cybersecurity, where identifying fraudulent domains is crucial.The method works within the gradient boosting framework, building decision trees successively to refine predictions and capture intricate patterns in phishing datasets. Cat Boost's skillsstand out in the context of phishing domain detection, where imbalanced datasets are widespread due to a lack of positive occurrences. The method includes techniques for addressing class imbalances, such as the ability to customize the weights allocated to positive and negative instances during training. This is critical for maintaining the model's sensitivity to smallpatterns suggestive of phishing activities. The hyper parameter tuning capabilities of Cat Boost are critical for optimizing its performance. With regularization strategies in place,the algorithm provides a means to avoid overfitting and manage model complexity. Fine-tuning parameters such

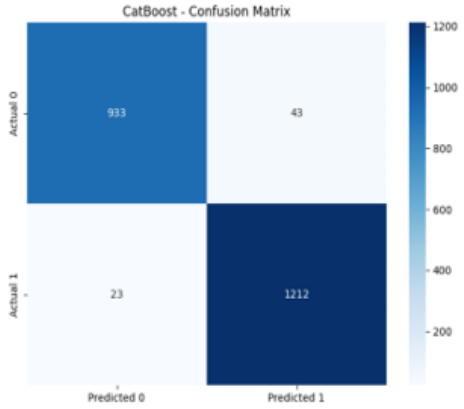


FIGURE 10. Confusion Metrics for Cat Boost

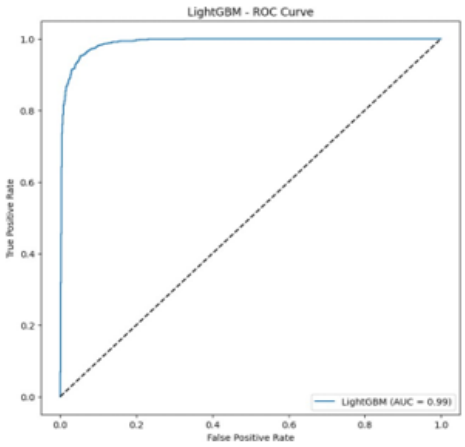


FIGURE 11. ROC for Cat Boost

as tree depth, learning rate, and regularization strength allow the model to be tailored to the complexities of phishing domain detection, while balancing precision and generalization. Cat Boost stands out not only for its predictive strength, but also for its interpretability. Cat Boost's feature significance ratings provide insight into how different features influence the model's judgments. In the case of phishing domain detection, this translates to a better understanding of the differentiating characteristics of malicious sites, which helps cybersecurity experts with threat analysis and decision- making. The success of Cat Boost is dependent on the quality of the dataset. In the dynamic field of cybersecurity, where phishing strategies very quickly, model adaption and upgrades are critical. Cat Boost can be effectively integrated into a comprehensive cybersecurity strategy when combined with other detection mechanisms and threat intelligence sources, resulting in a strong defense against evolving phishing threats.

Ada Boost: Ada Boost (Adaptive Boosting) is an ensemble learning technique that improves the performance of weak learners by giving alternative weights to instances during training iterations. In the area of phishing domain identifi- cation, Ada Boost can help build a strong model capable of reliably discriminating between legitimate and malicious sites. The technique works in a sequential manner, training weak learners, which are often simple decision trees, and adding weights to instances depending on classification failures. In the context of phishing domain detection, significant features may include domain length, the use of numerals or special characters, and the use of subdomains. Furthermore, website content elements and historical data, such as registration dates and hosting information, can be included in the feature set. The initial weak learner is trained on the full dataset, and misclassified cases are then given greater weights. In each iteration, Ada Boost gives more emphasis to the misclassified instances, prompting the subsequent weak learners to focus on rectifying these errors. This iterative process continues until a predefined number of weak learners are trained or until a specified level of accuracy is achieved. The final model is an ensemble of these weak learners, each contributing to the overall classification

decision based on their individual strengths. AdaBoost's adaptability to diverse weak learners and ability to handle skewed datasets make it ideal for the com- plicate challenge of detecting phishing domains. The method performs well in circumstances where the target classes are imbalanced, as it dynamically adjusts the weights to stress the significance of correctly categorizing instances from the minority class, which is frequently the case with phishing sites. Furthermore, Ada Boost is not constrained to a specific base learner, offering flexibility in incorporating diverse weak learners into the ensemble. This adaptability makes it well- suited for capturing intricate patterns and subtle characteristics indicative of phishing domains. Overall, Ada Boost stands as a powerful tool in the arsenal of techniques for building accurate and robust models for phishing domain detection, leveraging its ability to adapt and focus on challenging instances in the dataset.

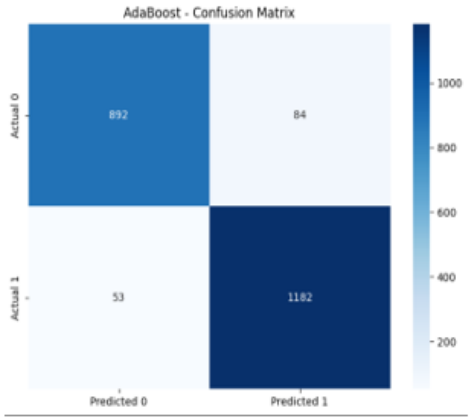


FIGURE 12. Confusion Metrics for Ada Boost

Gradient Boosting: Gradient Boosting is a powerful machine learning technique that uses an ensemble of weak learners, typically decision trees, to build a strong prediction model. This approach is very useful for detecting phishing domains since it can handle complex, non-linear correlations in the data. The algorithm iteratively constructs a sequence of decision trees, with each tree attempting to repair the flaws of its predecessors. During each iteration, the algorithm assigns higher weights to instances that were incorrectly classified in previous rounds, allowing it to focus on improving accuracy for difficult cases, which is critical in detecting phishing domains where subtle patterns may indicate malicious intent. In the case of phishing domain detection, the features utilized

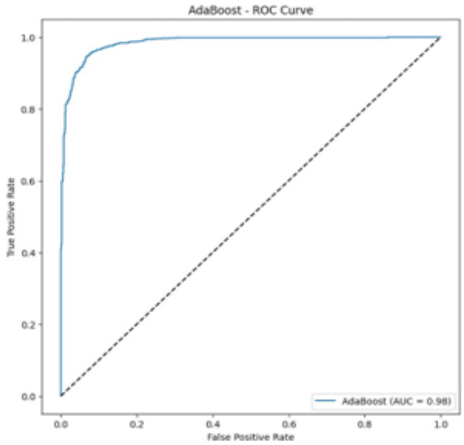


FIGURE 13. ROC for Ada Boost

to train the model could include domain name length, the inclusion of special characters, closeness to valid domains, and the use of specific keywords. The decision trees in the ensemble work together to capture the subtle patterns associated with phishing behavior. Gradient Boosting's sequential structure enables it to adapt to the complexities of the data and gradually enhance its predictions. Regularization techniques are commonly used to improve model performance and re- duce overfitting. These strategies involve limiting tree depth, introducing shrinkage (cutting each tree's contribution), and imposing weight limitations on instances. Gradient Boosting's

predictive power in phishing domain identification is enhanced by its capacity to handle imbalanced datasets, in which the number of valid domains considerably outnumbers the number

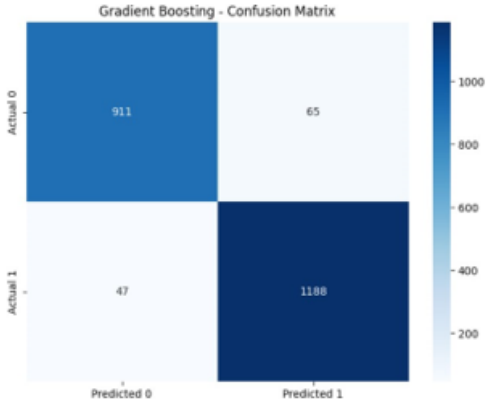


FIGURE 14. Confusion Metrics for Gradient Boosting

of phishing domains. Furthermore, the method handles missing data and categorical characteristics well, allowing for greater feature representation flexibility. Gradient Boosting's strength is its versatility, sequential learning, and ability to capture complex correlations within data. This makes it a powerful tool for creating accurate and trustworthy models for detect- ing phishing domains, where identifying subtle and dynamic trends is critical for effective cybersecurity.

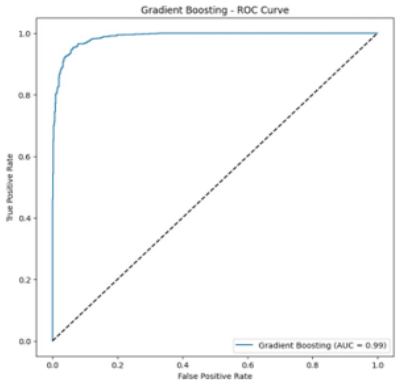


FIGURE 15. ROC for Gradient Boosting

A. MLP Classifier

Deep learning is an area of machine learning that focuseson neural networks with numerous layers, commonly knownas deep neural networks (DNNs). These networks are intendedto automatically learn and represent hierarchical characteristics from raw data, allowing them to detect complicated patterns and relationships. Deep learning models can be trained toexamine numerous features connected with domain names, URLs, and web content in order to detect phishing sites. However, while deep neural networks have exhibited great performance in a variety of sectors, they confront hurdles whenused to detect phishing. One of the fundamental constraintsof deep learning in phishing domain detection is a lack of labeled training data. Deep neural networks thrive when feda large amount of labeled data for training, allowing themto learn subtle patterns and generalizations. In the case of phishing domains, getting a sufficiently large and diverse dataset with precisely classified examples is a difficult issue. Phishing attacks develop frequently, and attackers constantly adapt their techniques, making it challenging to keep an up- to-date and comprehensive dataset. As a result, deep learning models may struggle to generalize effectively to new and previously unknown phishing behaviors. Another problem is the interpretability of deep neural networks. To increase trust in the model's predictions, phishing detection requires opennessand explain ability.

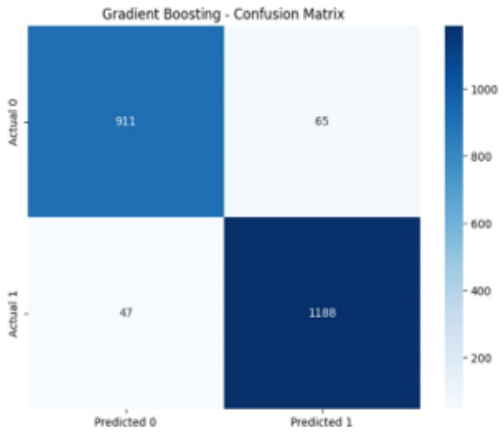


FIGURE 16. Confusion Metrics for MLP Classifier

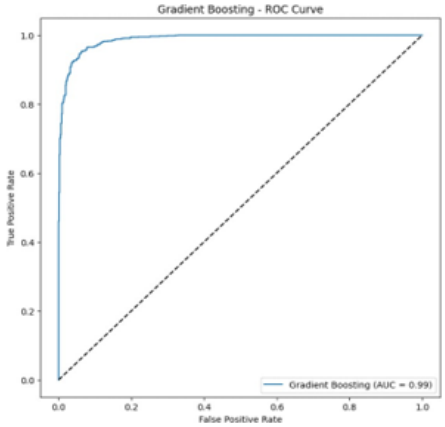


FIGURE 17. ROC for MLP Classifier

Deep learning models, particularly when deep and compli- cated, are frequently viewed as" black boxes" with poor inter- pretability. Understanding why a certain domain is tagged as phishing gets more difficult, making it harder for cybersecurity professionals to confirm and trust the model's judgments. In contrast, ensemble learning provides a convincing alternative for detecting phishing domains. Ensemble approaches, such as Random Forest, LightGBM, and XG Boost, use the combined strength of numerous weak learners to build a robust and accurate model. In the context of phishing detection, ensemble learning can include decision trees that excel at capturing specific characteristics associated with phishing behavior. Ensemble models, unlike deep neural networks, are more inter- pretable, providing insights into the features that contribute to a certain choice. This transparency is critical for cybersecurity specialists who want to comprehend and confirm the model's results. Ensemble learning also reduces the likelihood of over- fitting, a typical issue with deep neural networks. Overfitting happens when a model learns noise or peculiarities in training data that do not translate well to fresh, untested data. Ensem- ble approaches, which aggregate predictions from numerous models, are less prone to overfitting, resulting in improved generalization performance. Furthermore, ensemble learning approaches can efficiently handle imbalanced datasets, which are a prevalent feature of phishing domain detection when the number of real domains vastly outnumbers the number of phishing domains. Algorithms like as XGB Boost and Random Forest provide mechanisms for assigning appropriate weightage to minority classes, enhancing the model's capacity to detect infrequent phishing attempts. While deep learning has demonstrated enormous potential in a variety of disciplines, its application to phishing detection faces problems due to data scarcity, interpretability, and generalization. Ensemble learning, with its transparency, interpretability, and ability to deal with imbalanced datasets, appears as a more practical and effective method for developing robust phishing detection models in real-world cybersecurity settings.

4. IMPLEMENTATION AND INFERENCE FROM THE ALGORITHMS

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.9665	0.9665	0.9665	0.9665
XGBoost	0.9697	0.9699	0.9697	0.9697
LightGBM	0.967	0.967	0.967	0.967
CatBoost	0.9701	0.9702	0.9701	0.9701
AdaBoost	0.938	0.9382	0.938	0.9379
Gradient Boosting	0.9493	0.9494	0.9493	0.9493
Deep Learning	0.9674	0.9675	0.9674	0.9674

FIGURE 18. Performance of individual Models

5. CONFUSION METRICS

A confusion matrix is an important tool for evaluating the efficacy of a classification model since it allows for a detailed study of its predictions in comparison to the ground truth. The confusion matrix is especially useful in binary classification, where the outcome can be classified as either positive or negative.

6. ROC CURVE

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a binary classification model's performance at different decision thresholds. It compares the genuine positive rate (sensitivity) to the false positive rate (1 - specificity), demonstrating the trade-off between accurately

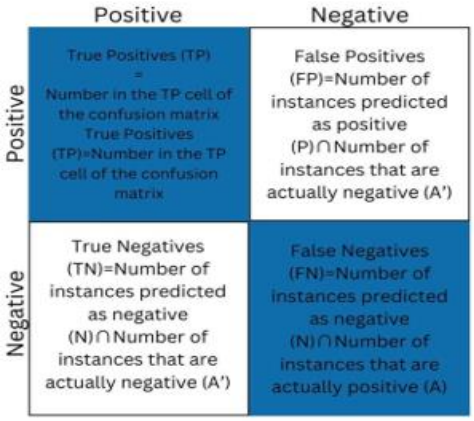


FIGURE 19. Confusion Matrix

detecting positive and wrongly classifying negative occur- rences. The curve assesses the model's discriminatory power, with an ideal curve in the upper-left corner suggesting high sensitivity and a low false positive rate. The Area Under the ROC Curve (AUC-ROC) measures overall performance and ranges from 0 to 1, with higher values indicating superior discrimination. The ROC curve and AUC-ROC are useful tools for evaluating and comparing the performance of binary classification models.

7. F1 SCORE

The F1 score is a complete metric for binary classification that incorporates precision and recall. Precision quantifies the accuracy of positive predictions, determining how many projected positives are actually positive. It is determined as the ratio of genuine positives to the sum of true positives plus false positives. Recall evaluates the model's capacity to capture all relevant positive events and determines how many actual positives were accurately predicted. The formula uses the ratio of true positives to the sum of true positives and false negatives. The F1 score, calculated as the harmonic mean of precision and recall, gives a balanced evaluation of a model's performance by accounting for both false positives and false negatives. This makes it especially useful in

situations where striking a balance between precision and recall is critical for determining the overall success of a binary classification model. The F1 score runs from 0 to 1, with 1 being an ideal mix of precision and recall.

8. PRECISION

Precision is a statistical indicator used widely in statistics and machine learning, particularly in classification settings.

$$F1=2 \text{ Precision x Recall/ Precision + Recall}$$

In binary classification, where there are two possible out- comes, precision measures the accuracy of a model's positive predictions. It is determined by dividing the number of true positive predictions by the sum of true positives and false positives. Simply said, precision measures the proportion of accurately predicted positive instances among all instances projected as positive by the model. This statistic is especially useful when the goal is to reduce false positives, as in instances where errors in positive predictions have serious effects. A high precision score indicates that when the model predicts a favorable outcome, it is likely to be correct; but, it may be more careful in generating positive forecasts, potentially excluding some positive examples.

$$\text{Precision=True Positive/True Positive + False Positive}$$

9. RECALL

Recall, an important metric in statistics and machine learning, measures a model's ability to properly identify and include all instances of a certain class inside a dataset. Recall is calculated as the ratio of true positive predictions to the sum of true positives and false negatives. It offers information on the model's sensitivity to actual positive instances. This statistic is especially important in situations when neglecting good events can have serious effects. A high recall score indicates that the model accurately catches the majority of positive cases, albeit this may be accompanied by an increase in false positives. Achieving a balance between recall and precision is frequently required for optimizing the performance of a classification model.

$$\text{Recall=True Positive/ True Positive + False Negative}$$

10. FUTURE WORK

Future work on phishing domain identification could include seamlessly integrating the suggested detection method with the WHOIS database, which would improve the model's accuracy by incorporating valuable domain registration and ownership information. This integration would add to the feature set, offering more context for discriminating between genuine and potentially malicious domains. Simultaneously, the creation of a real-time monitoring system for continuously tracking newly indexed domains is seen as a critical breakthrough. By developing a pipeline that retrieves and analyzes recently indexed domains in real time, the model can quickly identify domains as phishing or authentic. This dynamic monitoring strategy enables proactive detection and rapid reaction to developing phishing attempts, resulting in a more robust protection against growing cyber threats. Furthermore, implementing a system for dynamic feature updating that responds to changes in phishing tactics, content similarity, and URL structures would improve the model's adaptability over time. Furthermore, establishing engagement with the larger cybersecurity community and participating in knowledge-sharing initiatives could improve the model's capabilities by combining insights, datasets, and approaches. These proposed future steps aim to strengthen phishing domain detection, offering a complete and proactive defense against the ever-changing world of cyber-attacks.

11. CONCLUSION

Finally, this study investigated ensemble approaches for detecting dynamic phishing domains within online surfing extensions. The pervasive threat of phishing attempts needs adaptive and sophisticated defenses, and ensemble approaches prove useful in addressing the issues given by the dynamic nature of such threats. We discovered patterns of performance, adaptability, and efficiency after meticulously evaluating several ensemble approaches such as Bagging, Boosting, Random Forest, Stacking, Ensemble of Ensembles, and Gradient Boosting. Notably, our findings show that Cat Boost, a variation of the Boosting technique, consistently produces better results for the used dataset. Its ability to handle dynamic patterns and innate adaptability make it an appealing candidate for strong phishing domain identification within web surfing extensions. Furthermore, the feature selection process presented in this study refines the ensemble approach. By recognizing and prioritizing critical features, ensemble models, particularly Cat Boost, improve performance and efficiency. This emphasizes the need of careful feature selection in maximizing the effectiveness of ensemble-based systems for dynamic cybersecurity concerns.

REFERENCES

[1]. Detecting Phishing Websites Using Machine Learning: Aniket Garje1, Namrata Tanwani1, Sammed Kandale1, Twinkle Zope1, Prof. Sandeep Gore2 1 UG Students,2Assistant Professors, Computer Engineering Department, G H Raisoni College of Engineering and Management, Pune

[2]. Detection of Phishing Websites by Using Machine Learning-Based URL Analysis: Mehmet Korkmaz Yildiz Technical University Computer Engineering Department Istanbul/Turkey Ozgur Koray Sahingoz Istanbul Kultur University Computer Engineering Department Istanbul/Turkey Banu Diri Yildiz Technical University Computer Engineering Department Istanbul/Turkey

[3]. Model of detection of phishing URLs based on machine learning: Kateryna Burbela Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

[4]. Phishing Website Detection using Machine Learning Algorithms Rishikesh Mahajan MTECH Information Technology K.J. Somaiya College of Engineering, Mumbai, Irfan Siddavatam Professor, Dept. Information Technology K.J. Somaiya College of Engineering, Mumbai

[5]. Phishing Detection using Machine Learning based URL Analysis: A Survey Arathi Krishna V* , Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee Department of Computer Science and Engineering, College of Engineering Kidangoor Kottayam, India

[6]. Phishing Websites Detection Using Machine Learning P. Amba Bhavani ASST.PROFESSOR, Department of Information Technology, Maturi Venkat Subba Rao (MVSR) Engineering College Chalamala Madhumitha, Department of Information technology, Maturi Venkata Subba Rao (MVSR) Engineering, Hyderabad, India. Pinnam Sree Likhitha, Department of Information technology, Maturi Venkata Subba Rao (MVSR) Engineering, Hyderabad, India. Chanda Pranav Sai, Department of Information technology, Maturi Venkata Subba Rao (MVSR) Engineering, Hyderabad, India

[7]. Phishing URL detection using machine learning methods SK Hasane Ahammad a , Sunil D. Kale b , Gopal D. Upadhye b , Sandeep Dwarkanath Pande c,* , E Venkatesh Babu a , Amol V. Dhumane b , Mr. Dilip Kumar Jang Bahadur d a Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur 522502, India b Pimpri Chinchwad College of Engineering, Pune 411044 c MIT, Academy of Engineering, Alandi, Pune, India d Department of Computer and Information Sciences, Himalayan School of Science and Technology, Swami Rama Himalayan University, Dehradun, Uttarakhand, India

[8]. Detecting Phishing Domains Using Machine Learning Shouq Alnemari * and Majid Alshammari * Collage of Computer and Information Technology, Taif University, Taif 26571, Saudi Arabia

[9]. Phishing URL Detection using Machine Learning Nematullah Noori , Vyenkatash Bawanthad , Mayur Pakhare , Ramashray Agrawal , Vinod Kimbahune5 ,Department of Computer Engineering Dr D. Y. Patil Institute of Technology, Pimpri, Pune

[10]. Robust Ensemble Machine Learning Model for Filtering Phishing URLs: Expandable Random Gradient Stacked Voting Classifier (ERG-SVC) PUBUDU L. INDRASIRI 1 , MALKA N. HALGAMUGE 2 , (Senior Member, IEEE), AND AZEEM MOHAMMAD1 1School of Computing and Mathematics, Charles Sturt University, Melbourne, VIC 3000, Australia 2Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, VIC 3010, Australia Corresponding author: Malka N. Halgamuge (malka.nisha@unimelb.edu.au) This work was supported by Charles Sturt University (CSU).

[11]. An Adversarial Attack Analysis on Malicious Advertisement URL Detection Framework Ehsan Nowruz , Senior Member, IEEE, Abhishek , Member, IEEE, Mohammadreza Mohammadi , Member, IEEE, and Mauro Conti , Fellow, IEEE