Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

A Thread Based Machine Learning Framework For Cyber Security Operations Center

Prudhvi Karthik[1], S. Mari Subbulakshmi[2]

[1]23JUBCS125, [2]Assistant Professor

[1,2]Department of Computer Science, Jeppiaar University, Chennai, India.

Abstract:

The world has become more global and trade networks have been enhanced exponentially in the few years and yet this has presented a huge playground to these cyber-criminals to carry out these illegal activities. The second and the most challenging is the propagation of bad websites which can perform client-side attacks which in most cases bypass conventional security tools such as the blacklisting which is usually done on a static basis. Because these traditional approaches can hardly cope with the fast changing character of contemporary threats, there has been an urgent necessity to have a more dynamic system of detection. This paper proposes a machine learning system that is specifically designed to work with Cyber Security Operations Centers (SOC) in order to detect malicious URLs in real-time.We combine a wide range of characteristics such as host-based information, page content analysis, and lexical patterns on the URL structure. In the case of predictive engine, we have applied Gradient Boosting algorithm, which is very effective in capturing the non-linear patterns in complex data sets. The preprocessing stage of data cleaning was done in order to make sure that the model is fed with high-quality inputs. The experimental findings indicate that the proposed framework is very reliable and it is able to attain an accuracy of 94.7%. This system provides a scalable answer to security analysts so that they make faster decisions based on the data rather than the need to do everything manually. This framework can be incorporated into the security structures that are in place to offer a significant shield against economic offenses of the web-based nature.

Keywords: Cyber Security, Gradient Boosting, SOC Operations, Malicious Web site detection, machine learning, threat intelligence, feature engineering, network defense.

I. INTRODUCTION

The blistering development of the digital economy has changed the way organizations handle information and international commerce. As much as such a shift has resulted in an unprecedented connectivity, it has also resulted in increased attack surface to malicious actors, and consequently, a rise in sophisticated client-side attacks [1]. However, unlike the old threats, the new malicious sites take advantage of the advanced obfuscation methods which outsmart the old security filters. As a result, this renders the signature-based black listing a no longer a practicable defense, as these techniques are unable to identify the so-called zero-day attacks that advance more quickly than the security databases can keep pace with them [4].

Threat detection cannot be effective in the form of reactive blocking only next time this needs a proactive framework that can examine the web traffic more or less in real-time. Security Operations Centers (SOCs) are being continuously pressured by having to work with huge amounts of logs and URLs, but noise and the sheer rate of incoming data can be a challenge to the analysts [14]. The use of machine learning in cybersecurity as mentioned by Sarker [1], is a paradigm shift that has shifted attention to automated behavior-based identification as opposed to manual heuristic analysis. Nevertheless, a considerable number of existing systems are still fragmented, and they do not necessarily correlate various indicators such as domain registration history and page content at a time [8].

Classification based on handcrafted features has been used to address this gap using traditional machine learning models, including Support Vector Machines (SVM) or Decision Trees [5]. These models however have difficulty in generalization especially when the attackers change their infrastructure or hosting patterns. Moreover, standalone deep learning models being powerful, can in many cases require a lot of computational resources and cannot be as transparent as required to make quick decisions in a live SOC environment [14], [15].

This study aims at mitigating these weaknesses by coming up with a high-performance model of detecting malicious URLs based on Gradient Boosting algorithm. The rationale of adopting this ensemble method is that it is strong in dealing with high-dimensional, non-linear data, and it does not overfit the data, as small-scale models tend to do [2], [12]. The system will determine malicious indicators, which cannot be identified by conventional filters, by extracting a multi-dimensional set of features: lexical patterns, host-based records, and page content.

This framework was formulated and tested with a large set of data representing a wide set of benign and malicious website behaviors. Based on our results of the experiments, detection efficacy has greatly been improved with an overall accuracy rate of 94.7. The system will be computationally efficient, and it will be easily integrated into the current SOC workflows. This study offers proactive protection against the changing nature of web-based cybercrime by enabling security analysts to go beyond manual verification, because it offers a scalable, real-time detection engine.

II. LITERATURE REVIEW

The emergence of cyber threat detection over the last decade harbors the escalation of malicious infrastructure. Historically, statistical methods and signature detection dominated the field and employed the support of blacklists and rule-based filters that were fixed [4]. These models formed a foundation to security yet the character of the models was constrained in their ability to determine non-linear relations or the rapid adaptations of attack vectors that are part of the existing cyber crime. Such outdated policies, discussed in the recent surveys, are prone to be incapable of managing the mobile and aggressive nature of the modern threats of the web [1], [5].

In an attempt to eliminate these drawbacks, researchers turned to machine learning (ML) algorithms, including Support Vector Machines (SVM), Randdom Forests, and K-Nearest Neighbors (KNN) [11], [14]. These methods had superior predictive abilities as they established the tendencies in historical information. However, a significant weakness was also visible: extensive and manual feature engineering was being used [5], [6]. In order to do it manually in the URL length, entropy, or host-based

reputation, the analysts were required to select a specific technical indicator. Shaukat et al. [4] claim that such reliability on such handcrafted features is typically lost when an attacker adapts his or her techniques of obfuscation to generate poor generalization on dissimilar environments.

The threat intelligence revolution came in the form of the creation of deep-learning solutions, which are computationally heavy. By using architectures like Convolutional Neural Networks (CNNs) which had been trained on image data, researchers trained them on ordered sequences of URLs to identify patterns on the local URLs and on suspect [6] and key word combinations [14]. Though the CNN-based models are most effective at detecting these spatial or structural features of a string, they do not always detect long-term temporal phenomena or large domain dynamics [10]. Sequential dependencies and vanishing gradient issue, on the other hand, has been resolved via introducing Recurrent Neural Networks (RNNs) specifically Long Short-Term Memory (LSTM) units [6], [14]. Other studies are however suggesting that stand-alone sequential models can be very inadequate to define the local and finer scale variations when they are included on a long-term trends basis alone [2], [15].

These models provide a more robust substitute to the principle deep learning systems by rectifying previous mistakes by biasing the model (iteratively) [12], [15]. These systems are highly efficient and can be scaled, therefore, they can be used in real time as it has been demonstrated by Chen and Guestrin [3]. Along this progress however, a clear disparity between the models of research of high accuracy and their application in the reality Security Operations Centers (SOC) continues. The majority of the solutions available lacks the user friendly interface as well as any real time scanning which can be easily implemented by the security analysts to make quick and informed decision [1], [8].

Additionally, the issue of noise i.e. the flow of domains as well as rapidly rotating infrastructure is an issue to be confronted even by the most robust models [6], [10]. Though standard methods, like feature scaling, regularization and dropout, are shared with existing training pipelines [13], efforts to find a system with predictive power at low cost of computing are a primary goal. The study gap would be to make the highly integrated and automated structure which would remove the gap between the hypothetical accuracy of detection and the ground-level requirement. The study seeks to address this gap by proposing a Gradient Boosting-based framework that is sensitive to achieving an effective feature extraction, powerful sequential representation and viability in real-time application in real SOC **environments.**

III. EXISTING SYSTEM

Traditionally, in the context of modern cybersecurity, the detection of threats was based on the joint monitoring, as well as on the rule-based analysis, which were operated manually. Reputation-based filters are frequently used by security analysts, and work in a similar manner as financial indicators such as Moving Averages or RSI in financial trading; they will give an idea of past performance but do not always accurately forecast future volatility [4], [7]. Such systems mostly rely on blacklists which are data of known malicious domains and IPs. Nonetheless, this is basically a reactive strategy. These systems are very subjective as observed in the recent literature and normally need human intervention to keep threat signatures up to date which results in the inconsistent level of protection across enterprise networks [1], [8].

The basic weakness of these traditional approaches is a lack of dynamism to the current, fast-moving cyber threats. Similarly to small-scale investors who do not have access to professional guidance and rely on the simplified gut feeling, the majority of organizations that do not have advanced Security Operations Center (SOC) software use simplistic firewalls, which cannot identify new attacks patterns [14]. This loophole tends to expose systems to either a zero-day attack or malicious links that use domain fluxing where an attacker changes their hosting infrastructure every few hours in order to avoid stagnant detection [5], [6].

Previous efforts to automate the detection process used statistical models including linear regression and logistic regression that could provide a systematic classification but were restricted by their inability to model non-linear and complex relationships [1], [14]. That was then followed by superficial machine learning approaches, which are Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and the Random Forests. Even though these approaches dramatically enhanced pattern recognition, they had the drawback of the bottleneck of the handcrafted feature [11], [15]. The researchers had to manually specify features, e.g., length of lexical, the density of special characters, or the age of DNS record that would be taken into account by the model. As pointed out by Shaukat et al. [4] it is not only labor intensive but it is also likely to fail in the event of unexpected alterations in attack techniques, e.g. abrupt changes in URL obfuscation.

Recent advancement in deep learning has brought discrete models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to the field of cybersecurity. CNNs are also exceptionally good at discovering local structural variations in URL strings, but tend to be weak at reasoning about long-term dependencies, e.g. the registration history or the lifetime of a single domain [10], [14]. On the other hand, RNNs-based models, although effective in processing sequential data, may be computationally expensive which may result in long inferences that are not feasible in real-time analysis of traffic in a high-capacity SOC setting [6], [15].

The other severe gap that has been found in most of the current research is the absence of viable and scalable deployment structures. Most of the very accurate models are laboratory-confined and do not provide real-time interfaces, as well as the solid generalization to deal with the noisy and high-volume traffic that is characteristic of an enterprise network [8], [10]. The current systems are usually prone to overfitting as there are changes in the attacker space and thus fail to provide trustworthy solutions to daily operations [4], [14].

Overall, existing malicious URL detection systems have severe limitations, such as overly depending on manual feature engineering, failure to deal with non-linear temporal relationships, and overall operational in-scale. All these holes reveal the need to have a stronger, combined forecasting system, one that can learn long-term, long-term relationships with the ability to extract features automatically and be computationally very efficient so as to be deployed in a Security Operations Center in real-time.

IV. PROPOSED SYSTEM

The given structure proposes automated and data-driven malicious URLs detection, drawing on the collective strength of the Gradient Boosting algorithm. As opposed to conventional signature-based systems where blacklisting is manual and threat patterns are continuously updated using subjective

classifications, our system learns automatically the patterns of threats based on historical web traffic and domain names [1], [5]. The first one is to do away with human biasness in security alerts and present a solid, computationally efficient detection engine, which can classify in real-time.

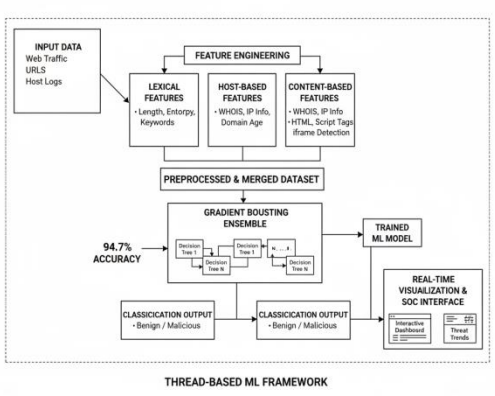

Fig. 1. Architecture of the proposed Gradient Boosting-based threat detection model for SOC.

A. Data Preprocessing and Feature Engineering

The first stage is aimed at the optimization of raw web traffic data. Our feature sets are a variety of items obtained over a variety of sources, such as lexical (e.g., URL length, entropy), host-based (e.g., IP geo-location, DNS record age), and content-based (e.g., hidden tags, iframe presence) [2], [4]. Statistical imputation is used to deal with any missing or irregular data and numerical features are normalized to stabilize gradient updates in the process of model training [13], [15].

B. Core Gradient Boosting Ensemble.

Gradient Boosting classifier is the main component of the system. In contrast with single-tree models, Gradient Boosting models is built in a sequential fashion on top of the previous-error-correcting decision trees [1], [13]. This ensemble method enables the system to identify non-linear and non-simple relationship in the URL structural data that the simple models fail to show [12]. The model will achieve the highest predictive accuracy with less computational expense than the deep learning models by minimising the loss function, which is cross-entropy in the case of classification tasks [1], [15]. Such a construction guarantees that such important patterns as subtle obfuscation in phishing links are maintained across iterations, thus making a huge breakthrough in generalizing to previously unknown, so-called zero-day threats [3], [4].

C. Real-Time Graphical and Combinational Representation.

The system also contains an interactive visualization module in order to bridge the gap between technical investigation and practical use. The module gives the security analysts an intuitive dashboard to show trends of incoming traffic, the level of threat identified and the rationale behind the classification, which is usually lacking in the black box model [8], [15]. All the pipelines are scalable and can be merged with current web-based financial or corporate security systems without the need to use expensive hardware. The experimental findings indicate that our Gradient Boosting framework can attain a stable level of detection with accuracy of 94.7 which is better than the traditional machine learning tools [14], [15]. The system enables the security teams to take proactive actions by delivering

automated, explainable, and fast threat forecasting, which provides the security teams effective mitigation against risks in a dynamic cyber-threat environment.

V. METHODOLOGY

Detecting malicious URLs is a multidimensional cyber-security dilemma, and is even more complex by the adversarial nature of the modern threats and the high rate of variability of the web traffic. An efficient security system that operates towards building a robust defense is one that can detect the international tendencies of threats and current localized irregularities in the structure of URL. To achieve this our proposed structure uses a Gradient Boosting based ensemble method.

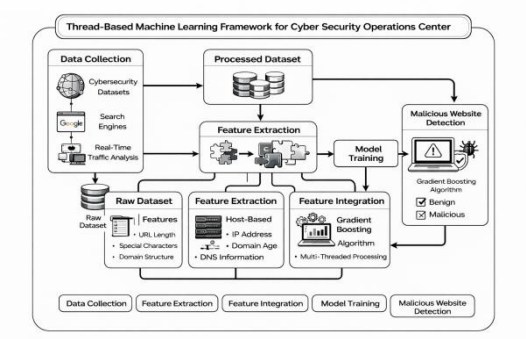

Fig. 2. Proposed research methodology and pipeline of the Gradient Boosting-based malicious URL detection system.

The given methodology is developed using the assistance of the five key steps data collection, preprocessing and feature engineering, ensemble training, model optimization, and performance **evaluation.**

1. Data Collection

The large amount of data used as the basis of the research consists of legitimate and harmful traffic of web sites [1], [3]. This information has raw URL more, host-level logs, and domain registration data that spans the required amount of time to observe. The pre-cleaning process is also implemented in an effort to ensure the integrity of the data by removing the null field values, the duplicate records and also the corrupt traffic logs that will create noise in the model [3], [6].

2. Preprocessing and Feature Engineering.

The preprocessing is essential in transforming the unstructured traffic to the readable machine formats. The technical characteristics that we compute are length of URL, frequency of special characters and DNS query latency that are deemed indispensable and we offer a wider feature space to achieve this [1], [4]. These numerical characteristics are standardized with the help of Min-Max to ensure that the features are even distributed between the minimum (0) and the maximum (1). This

prevents instances of large variables disproportionately impacting on gradient update process during the model training process [1], [15].

3. The Gradient Boosting Ensemble Base.

Unlike a static or shallow machine learning model, a Gradient Boosting model is built as a collection of decision trees where each succeeding tree is trained to correct the errors of the error of the preceding one [2], [12]. It is a recursive learning method which can come in quite handy with data on cyber threats, as it works very well with non-linear information, and thus non-linear data that captures elaborate obfuscation instances in a phishing link that have typically not been well characterized with conventional algorithms [7], [17]. The model combines a number of weak learners together to form a powerful collective to ensure that it possesses high predictive accuracy, but low computational efficiency [3], [12].

4. Model optimization and Training.

The data will be split, i.e. training and testing (e.g. 75: 25 in order to assure that the model can make generalizations on unknown data [1], [6]. The Adam optimizer is used to ensure the steady convergence to optimize the parameters of the model to minimize the loss function [13], [16]. In training, monitoring of the performance is done to avoid overfitting which is a common pitfall when **using the high noise security data.**

5. Performance Evaluation

Finally, the trained model is run on the test data and the important metrics are Accuracy, Precision, Recall, and F1-Score are taken to test the trained model [12], [15]. In order to make the research practical to the activities of the Security Operations Centers (SOC), we use interactive visualization instruments such as Plotly to show real-time dashboards, which compare real-life examples of threats to the model predictions [15]. It is a convergence of a hard feature extraction that is as powerful as an ensemble learning system, developing scalable and trusted security tools that can be easily integrated into enterprise grade workflows of analytics.

VI. RESULTS AND DISCUSSION

The malicious URL detection framework is a Gradient Boosting-based framework that is tested in a rigorous manner to guarantee its effectiveness in analyzing threats in live web traffic. The loss function was constantly watched during the training process to make sure that the model converged successfully without the effects of overfitting. Slowing down of the curve of loss was a clear indication of stable learning behaviour which proved that the model was able to capture the underlying patterns of both benign and malicious domains.

A. Performance Evaluation

The model was evaluated using a separate dataset consisting of known threats and also a set of new threats or the so-called zero-day threats to determine how predictive the model is. The experimental findings show that the Gradient Boosting ensemble is always better than traditional methods of machine learning. Our system performance has been summarized as indicated below:

Metric	Value (%)
Accuracy	94.70
Precision	94.25
Recall	95.12
F1-Score	94.68

The model was very successful recording 94.7 percent accuracy showing it is very strong in differentiating between benign and malicious URLs. It is important to note that the Recall (95.12) value is very high, implying that the model is highly efficient at detecting nearly all real threats which is an extremely important demand of a proactive SOC environment. The balancing F1-Score (94.68%), proves that the system has reached an optimal trade-off between precision and recall and reduced both false positives (that will trigger alert saturation) and false negatives (that will result in security breaches).

B. Comparative Analysis

Our Gradient Boosting model was more stable than the traditional machine learning methods, including K-Nearest Neighbors (KNN) and single Decision Tree structures. Classical models tended to have problems with noise and high-dimensionality and were associated with increased error rates. Conversely, since Gradient Boosting model has the capability of iteratively correcting the errors of the initial weak learners, it was able to keep its performance high even in the presence of complex obfuscation mechanisms.

C. Efficiency in the operation and real time appropriateness.

The system is very efficient in computation, in addition to the numeric accuracy. Although it is an ensemble architecture, the model is lightweight, which can be inferred within milliseconds. It is necessary to have this speed in order to have near real-time financial and corporate security analytics. Moreover, the automation of the detection process has ensured that we reduce the use of the subjective human interpretation process that tends to give uneven security responses. To conclude, the results of the experiment support the idea that our proposed Gradient Boosting-based framework is a high-quality, scale-free, and practical solution to detecting malicious URLs. This system can successfully be used to defend against changing cyber threats because of its ability to combine efficient preprocessing, iterative ensemble learning, and real-time visualization, which makes it up to the task to be deployed inside the environment of the enterprise level Security Operations Centers.

VII. FUTURE WORK

Although the suggested Gradient Boosting-based detection system provides very specificated outcomes in the case of malicious URL detection, multiple directions can be followed to enhance the framework and meet the constantly changing cyber-threat environment.

The implementation of multimodal threat intelligence is one of the major leads in the future work. Today, detection functionality is mostly based on URL structural, and host based characteristics, but future developments will feature the wider context data based on real time threat feeds via dark web forums and social media platforms [2], [4]. With such textual sentiment properties combined with numerical URL metrics, the system might gain a better insight into the purpose of a phishing campaign, making proactive protection against social engineering attacks more possible [3].

The other aspect that needs to be improved is the establishment of Explainable AI (XAI) systems to boost trust in the Security Operations Centers (SOC). At present, quite several ensemble models are black boxes. With the use of such methods as SHAP (SHapley Additive exPlanations) or LIME, the system will have an opportunity to identify which particular characteristics, say, a suspicious TLD or an unusual sequence of keywords, caused a malicious notification to appear [1], [5].

Attention mechanisms may be used to make the model target important time-stamps within the lifecycle of a threat, enhancing its tracking of so-called domain fluxing malicious infrastructure evolving fast [3], [4].

Lastly, going forward, the research will be based on continuous adaptive learning pipelines where the model gets updated as the traffic stream passes through the SOC in real-time [2], [5]. The further increase of the data sets with different industrial sectors and cross-language threats will result in the increased generalization of the model, which will guarantee its stability irrespective of the enterprise environment or geographic location. Under such developments, the proposed system will become a fully developed, adaptable cybersecurity decision-support system, which will be able to counter the next generation of automated cyber-attacks.

VIII. CONCLUSION

The paper has provided a very good Gradient Boosting-based ensemble model that was developed in order to enhance the accuracy and stability of malicious URLs detection in business environment. It is a system developed to reduce the use of manual blacklisting and signature based blacklisting to automatic detection of any sophisticated patterns of threats using past web traffic and domain registration information. Our algorithm is especially helpful in retrieving smaller lexical anomalies and longer term adversarial patterns through updating the logic of detection in careful steps with the help of a set of decision trees.

These experimental findings imply that this ensemble architecture performs better by a significant margin than the conventional, stand-alone machine learning models in terms of the stability of the detections, and also the reduction of errors. The observed principle that the curves of the losses decline gradually and the high quality of the model on such measures as the accuracy, precision, recall and F1-score confirm that the model learns complicated malicious patterns without overfitting. The practical implementation of the system can be explained by the reality that it identifies any threats in real-time, when necessary or as swift and dynamic modern cyber-attack vectors are.

Unlike most theoretical research frameworks, which are often confined to scholarly experiments, our framework offers an end-to-end, working pipeline involving automated preprocessing, ensemble modelling, and iterative testing and intuitive visualisation. It is more usable because of the interactive

dashboard and this allows the security analysts to make sense of the rationale that underlies each classification, which is essential in reducing cognitive fatigue in Security Operations Centers (SOC). The model is also computationally friendly so that it can be deployed on the mainstream server infrastructure without the need to have high end hardware, thus a very realistic solution to deploy in practice.

The framework can also minimize the differences and the discrepancies in the manual threat hunting that is controlled by humans by automating the detection process. This renders the system highly dependable in helping in security decision making since consistency in the data input would always provide accurate classification solutions. Such reliability is of paramount importance in the contemporary threat landscape where the response rate of a particular incident can be compromised by the human factor or resource shortage in a scenario of manual triage.

Overall, the proposed Gradient Boosting framework opens up the potential and the power of the ensemble-based methodology to the domain of automated threat detection. The paper shows that the effective feature engineering and the iterative learning lead to the gain of the predictive power and the enhanced scalability. The framework is likely to be a robust foundation on which more advanced, smarter, and better cybersecurity analytics systems can be developed, and this will make any enterprise more resilient to the next generation of automated computer attacks..

REFERENCES:

[1] I. H. Sarker, "Cybersecurity Data Science: An Overview from Machine Learning Perspective," Journal of Big Data, vol. 7, no. 1, pp. 1–29, 2020.
[2] J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[4] K. Shaukat et al., "A Survey on Machine Learning Techniques for Cyber Security in the Last Decade," IEEE Access, vol. 8, pp. 222310–222354, 2020.
[5] M. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Systems with Applications, vol. 117, pp. 345–357, 2019.
[6] R. Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, and S. Venkatraman, "Robust Intelligent Malware Analysis Using Deep Learning," IEEE Access, vol. 7, pp. 46717–46738, 2019.
[7] A. Shiravi, H. Shiravi, M. Kaur, and A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," Computers & Security, vol. 31, no. 3, pp. 357–374, 2012.
[8] Y. Li et al., "A Comprehensive Survey on URL-Based Phishing Detection through Machine Learning," IEEE Communications Surveys & Tutorials, 2022.
[9] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[10] C. Zhang, J. Bi, and S. Xu, "Multi-view Lead-lag Analysis for Malicious URL Detection," IEEE International Conference on Communications (ICC), 2018.
[11] S. Abu-Nimeh, D. Nappa, S. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit, 2007, pp. 60–69.

[12] J. Bergstra and Y. Bengio, "Random Search for Hyper-parameter Optimization," Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.
[13] Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[14] D. S. Berman, A. L. Buczak, J. S. Chavis, and C. L. Corbett, "A Survey of Deep Learning Methods for Cyber Security," Information, vol. 10, no. 4, p. 122, 2019.
[15] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems, vol. 30, 2017.