Phishing Site Detection Using ML Algorithms

Aman Anand[1], Aayush Gupta[2], Abinash Dubey[3], K Chakradhar Naidu[4]

[1,2,3,4]Student, CSE (AI&Ml) Department, Dayananda Sagar University

Abstract

This study delves into the growing threat of online phishing frauds by evaluating the efficacy of diverse machine learning (ML) algorithms in pinpointing malicious websites. Giventhe substantial risks phishing attacks pose to user privacy and security, it emerges as a promising solution due to its adaptability and the ability to glean insights from extensive datasets. Past research underscores the potential of Support Vector Machines (SVM) and Random Forests in phishing detection. Nevertheless, challenges persist in optimal algorithm selection and feature prioritization. The proposed system integrates Gradient Boosting and Cat Boost alongside Random Forest, leveraging features from the UC Irvine Machine Learning Repository. The study's relevance lies in its performance analysis, which steers the selection of the most effective algorithm for detecting phishing websites, making it pertinent for automated systems addressing the evolvinglandscape of online threats.

Keywords: HTTPS, SVM, ML model, URL, SSL, Domain Age, Content Analysis.

1. Introduction

In the chaotic world of cyber threats in 2023, we faced a serious uptick – 1.2 billion recordscompromised, ransomware incidents doubling, a 15% surge in phishing attacks, and a staggering 3 billion (about 9 per person in the US) lost in crypto hacks [1]. This surge laidbare the chinks in our cybersecurity armor, wreaking havoc on organizations worldwide.

Amid this digital era, where technology offers immense benefits, the rise in phishing attacks demands a fresh, Innovative solution. This research project steps up to the challenge,suggesting real-time detection using machine learning as a response to the shortcomings of traditional methods and an enhancement of overall cybersecurity.

Our trustworthy machine learning (ML) model diligently looks at various features suchas URL structure, SSL certificates, domain age, user behavior, and content anomalies. Through a careful evaluation of these aspects, our system builds a sturdy defense against phishing attempts. The project zeroes in on a binary classification task – distinguishing between websites that mean harm and those that are harmless. To supercharge real- time detection, we bring in a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM), capitalizing on their knack for efficiently storing andprocessing data.

Beyond the technicalities, the main aim here is to be proactive in spotting and thwartingphishing risks by weaving in advanced ML algorithms. The architecture we have built isnot just about nailing immediate accuracy; it is about staying a step ahead of emerging threats. This marks a shift in the game of phishing detection, contributing to a safer online world. As we move forward, subsequent sections will unwrap the methodologies,our careful selection of ML algorithms, how we gathered data, and the potential impact on cybersecurity measures. This highlights our project's role in crafting a secure digital future.

At its core, this project is fueled by the mission to strengthen cyber defenses by tapping intothe power of machine learning (ML). The big picture? Developing a resilient system that can independently identify and neutralize the risks tied to phishing websites. The urgency is clear when we think about the aftermath of falling prey to phishing – it is not just financial losses; it is about safeguarding personal and confidential data from being compromised Magine we aretraining digital detectives to spot the good guys from the bad in the vast online world. It is a bit like teaching them to predict if a website is trustworthy or not – a detective's job for the digital age. Instead of relying on old-school methods, we are using fancy tools from the machine learning toolbox. It is like giving our detectives high-tech gear to navigate the com- plex online landscape. Think of it as a high-tech map that helps us find hidden dangers.

Cybersecurity Now, let us talk about cybersecurity – the never-ending game of keeping our digital space safe. The traditional defenses, like static rules and signature-based detection, sometimes struggle to keep up with the ever-changing threats. That is where our superhero, machine learning (ML), comes into play. ML wears different hats – from spotting unusual activities and predicting future threats to automating the search for cyber dangers and makingour overall threat intelligence sharper. Machine learning is not just a fancy term; it is like a superhero toolbox for solving real-world problems. It powers web search engines, acts as the brains behind automated phishing detection, and becomes the sharpshooter in identifying phishing websites. It is also the guardian against credit card fraud, the analyst in the stock market, and the wizard behind speech recognition, turning spoken words into text.

Every superhero has its challenges. ML needs to be quick, cost-effective, and not gettoo obsessed with its training data – a fancy way of saying it should not focus too much on pastexperiences. It is also tackling speech-based phishing, making sure it understands and identifiesthreats in spoken words. Our project is ambitious. We are building a system that goes beyond rigid rules, analyzing everything from URL structures and user behavior to content nuances.Our mission? To cut down on phishing frauds, protecting people from financial losses, data breaches, and a loss of trust in online interactions. At the same time, we are strengthening cybersecurity, making advanced threat detection, reducing cyber risks, and diving into excitingresearch in AI and cybersecurity.

Based Detection Using Gradient and Cat Boost Classifier: This study explores how machine learning plays a vital role in spotting phishing websites, a critical aspect of cybersecurity. Conventional methods often struggle with new attacks, pushing the need for machine learning- driven detection. The emphasis is on using ensemble learning and gradient boostingtechniques, with a keen eye on metrics like accuracy and precision. The goal is to reduce falsenegatives, ensuring a better catch rate for phishing attempts. The study suggests diving deeper into deep learning and transfer learning to enhance detection capabilities. Understanding Detection and Defense Against Phishing: Leveraging Machine Learning and Deep Learning Techniques This research underscores the crucial role of machine learning approaches, such as neural net- works and Support Vector Machines (SVM), in improving phishing detection. Techniques like ensemble learning, especially incorporating XGBoost, along with careful feature selection significantly contribute to overall accuracy. Evaluation metrics like accuracy, precision, recall, and F1-score are crucial for ensuring the reliability and effectiveness of phishing detection models. Further Insights into Phishing Websites Classification using a HybridSVM and KNN Approach: Researchers passionately delve into using machine learning (ML) capabilities for automatic detection of phishing websites, aiming for improved accuracy. Theirapproach involves a multifaceted strategy, including fuzzy logic, the robustness of SVM, and synergistic ensemble methods like XGBoost. The commitment extends to using sophisticated techniques like identity discovery, keyword retrieval, and nuanced content-based feature analysis. The goal is not just superior accuracy in phishing detection but also developing comprehensive strategies to fortify cybersecurity against evolving online threats.

This thorough examination explores various methodologies to enhance phishing detection. Intelligent phishing detection, integrating ML for identifying phishing attacks, and applying supervised learning algorithms like Multi-layer Perceptron (MLP), Decision Tree (DT), and Naïve Bayes (NB) are key aspects. The analysis acknowledges the importance of identity andfeature extraction processes, emphasizing their pivotal role in ensuring effective detection mechanisms. The aim is to scrutinize and optimize these processes, paving the way for more robust and sophisticated systems in the realm of cybersecurity. A Hybrid Two-level Framework for Feature Selection and XGBoost Tuning: This paper introduces a sophisticated hybrid framework integrating the powerful XGBoost model. The framework exhibits outstanding performance through extensive experiments on datasets related to phishing websites. The strategy involves a two-level approach, incorporating advanced feature selection techniques and meticulous tuning of XGBoost parameters. The study highlights the effectiveness of this approach in boosting the accuracy and efficiency of phishing website detection. Together, theseresearch endeavors contribute significantly to advancing strategies for detecting phishing withinthe cyber- security landscape.

We have selected ten unique machine learning classification models to detect those annoying phishing websites. These include Logistic Regression, k–nearest neighbor, SupportVector Machine, Naive Bayes Classifier, Decision Tree, Random Forest, Gradient Boosting Classifier, Cat boost Classifier, XGBOOST Classifier, and Multilayer Perceptron (MLP).

2. Related Work

2.1 URL-Based Detection Using Gradient and Cat Boost Classifier:

The review explores machine learning's role in detecting phishing websites, crucial for cybersecurity. Traditional methods falter against evolving attacks, leading to a shift to ML- driven detection using techniques like ensemble learning and gradient boosting. Key metricssuch as accuracy, precision, recall, and false positives/negatives gauge model performance, with a focus on reducing false negatives to catch more phishing attempts. Ensemble modelsand gradient boosting show promise, while further exploration of deep learning and transferlearning is recommended to enhance detection capabilities.

2.2 Detection and Defense Against Phishing: Machine Learning and DeepLearning Techniques:

Machine learning (ML) approaches, including the application of neural networks and Support Vector Machines (SVM), play a pivotal role in enhancing the effectiveness of phishing detection. The utilization of ensemble learning techniques, specifically incorporating XGBoost, coupled with meticulous feature selection, contributes significantly to the overall improvement in detection accuracy. In the evaluation process, metrics such as accuracy, precision, recall, and F1-score hold paramount importance, serving as crucial benchmarks for ensuring the reliability and effectiveness of the phishing detection model.

2.3 Phishing Websites Classification using Hybrid SVM and KNN Approach:

Researchers ardently focus on harnessing the capabilities of machine learning (ML) to facilitate the automatic detection of phishing websites, striving for heightened accuracy in the process. Their endeavors encompass a multifaceted approach involving the strategic

application of fuzzy logic, the robustness of Support Vector Machines (SVM), and the synergistic potential of ensemble methods such as XGBoost. This concerted effort towardsprecision is complemented by the incorporation of sophisticated techniques like identity discovery, keyword retrieval, and the nuanced analysis of content-based features. These intricate methodologies not only display a commitment to achieving superior accuracy in phishing detection but also underscore the dedication to developing comprehensive and nuanced strategies for fortifying cybersecurity in the evolving landscape of online threats.

2.4 Efficient prediction of phishing websites using supervised learning Algorithms:

A comprehensive examination delves into diverse methodologies aimed at bolstering the field of phishing detection. These methodologies span intelligent phishing detection, the integration of machine learning (ML) for the identification of phishing attacks, and the application of supervised learning algorithms, including but not limited to Multi-layer Perceptron (MLP), Decision Tree (DT), and Naïve Bayes (NB). Recognizing the significance of identity and feature extraction processes, this in-depth analysis underscores their pivotal role in ensuring the efficacy of detection mechanisms. By scrutinizing and optimizing thesecritical processes, researchers and practitioners can pave the way for more robust and sophisticated systems, contributing to the advancement of effective phishing detection strategies within the realm of cybersecurity.

2.5 Improving Phishing Website Detection Using a Hybrid Two-level Frameworkfor Feature Selection and XGBoost Tuning:

This paper introduces a sophisticated hybrid framework that integrates the powerful XGBoost model. This framework exhibits exceptional performance, as demonstrated throughextensive experiments conducted on datasets related to phishing websites. The approach in-volves a two-level strategy, incorporating advanced feature selection techniques and meticulous tuning of XGBoost parameters, highlighting its effectiveness in bolstering the accuracy and efficiency of phishing website detection.

2.6 Phishing Website Detection Using Machine Learning Techniques:

This paper sheds light on the considerable potential of various methods, including pruned decision trees, Support Vector Machines (SVMs), Naïve Bayes, neural networks, associative classification, and fuzzy inference systems, in elevating the effectiveness of phishing detection.The paper emphasizes the need for future research endeavors to delve deeper into these techniques, suggesting that exploring and refining these approaches could significantly contribute to advancing the field of phishing detection and cybersecurity.

3. METHODOLOGY

Machine Learning Algorithm:

Ten machine learning classification model Logistic Regression, k–nearest neighbor, Support Vector Machine, Naive Bayes Classifier, Decision Tree, Random Forest, Gradient Boosting Classifier, Cat boost

Classifier, XGBOOST Classifier, Multilayer Perceptron (MLP) been selectedto detect phishing websites.

3.1 Logistic Regression:

It is like a statistical superhero predicting the probability of binary outcomes. Logistic Regression is your go-to for classification tasks, applying a logistic function to a mix of inputfeatures.

3.2 k-Nearest Neighbor (k-NN):

Think of it as the friendly neighbor who helps you classify a data point based on the majorityclass of its k closest friends in the feature space.

3.3 Support Vector Machine (SVM):

The superhero of supervised machine learning, SVM finds a hyperplane to separate classes in ahigh-dimensional space, creating maximum space between different class datapoints.

3.4 Naive Bayes Classifier:

This one is like a clever detective, using Bayes' theorem and assuming independence betweenfeatures. It is often your go-to for text classification and spam filtering.

3.5 Decision Tree:

Picture a tree that loves making decisions! The Decision Tree model recursively splits databased on features, with each leaf representing the predicted class or value.

3.6 Random Forest:

It is a team effort! Random Forest combines multiple decision trees to improve accuracyand reduce overfitting. Teamwork makes the dream work!

3.7 Gradient Boosting Classifier:

Here comes the ensemble hero! Gradient Boosting builds a strong predictive model by addingweak learners one after the other and adjusting their weights based on errors.

3.8 XGBOOST Classifier:

This one is like the cool kid in town, optimized and scalable. XGBOOST uses fancy regularization techniques to prevent overfitting, making it popular for structured/tabular data.

3.9 Multilayer Perceptron (MLP):

Meet the artistic one in the group! A Multilayer Perceptron (MLP) is like an artistic neural net- work with multiple layers of nodes (neurons), often used for handling complex non-linear tasks.

3.10 Cat Boost Classifier:

The cat lover in the bunch! Cat Boost is a gradient boosting algorithm that deals withcategorical features like apro, delivering high performance with default settings.

This diverse lineup of machine learning superheroes aims to demystify their roles in theexciting world of phishing detection, working towards as after cyber space for all.

4. RESULT AND ANALYSIS

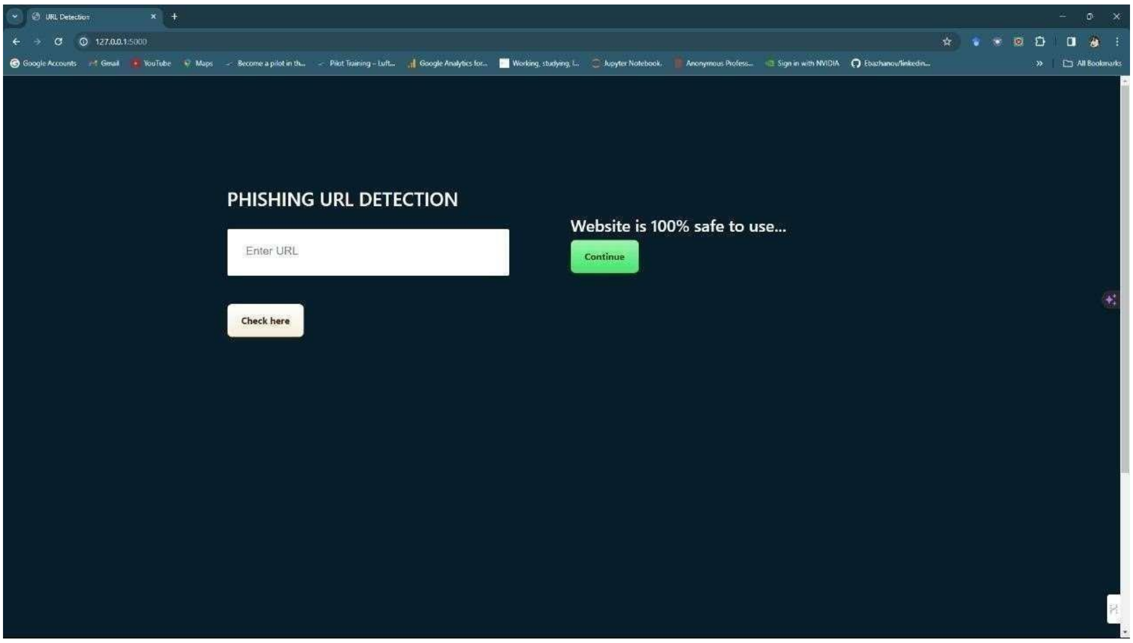
In our journey of implementing machine learning wonders, we have enlisted the help of Scikit- learn to seamlessly integrate these algorithms into action. Our dataset, a treasure trove of insights, undergoes a division into 80% training and 20% testing sets. The training sets serve as the training grounds for our classifiers, while the testing sets play a crucial role in evaluating theirprocess. Our assessment involves a meticulous calculation of accuracy scores, false negative rates, and false positive rates to understand the superhero-like performance of our classifiers.

Model Name	Accuracy	F1-Score	Recall	Precision
Logistic Regression	0.934	0.941	0.943	0.927
K–Nearest Neighbor	0.956	0.961	0.991	0.989
Support Vector Machine	0.964	0.968	0.98	0.965
Naive Bayes Classifier	0.605	0.454	0.292	0.997
Decision Tree	0.961	0.965	0.991	0.993
Random Forest	0.967	0.97	0.992	0.991
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
Cat boost Classifier	0.972	0.975	0.994	0.989
XGBOOST Classifier	0.969	0.973	0.993	0.984
Multi layer Perceptron	0.971	0.974	0.992	0.985

The stars of our show, the Gradient Boosting Classifier, and the Cat boost Classifier, consistently steal the spotlight with their stellar performance, highlighting accuracy ranging from 97.4% to an impressive 99.4%. As you embark on your decision-making journey, factors like model interpretability, computational efficiency, and your application's unique requirements might guide you to choose between these two exceptional models. Happy modeling!

5. APPLICATION

The outcomes obtained through the experimental assessment are displayed on this display. These results are generated through the application of the ML Algorithms utilized in the systems to attain optimal precision.

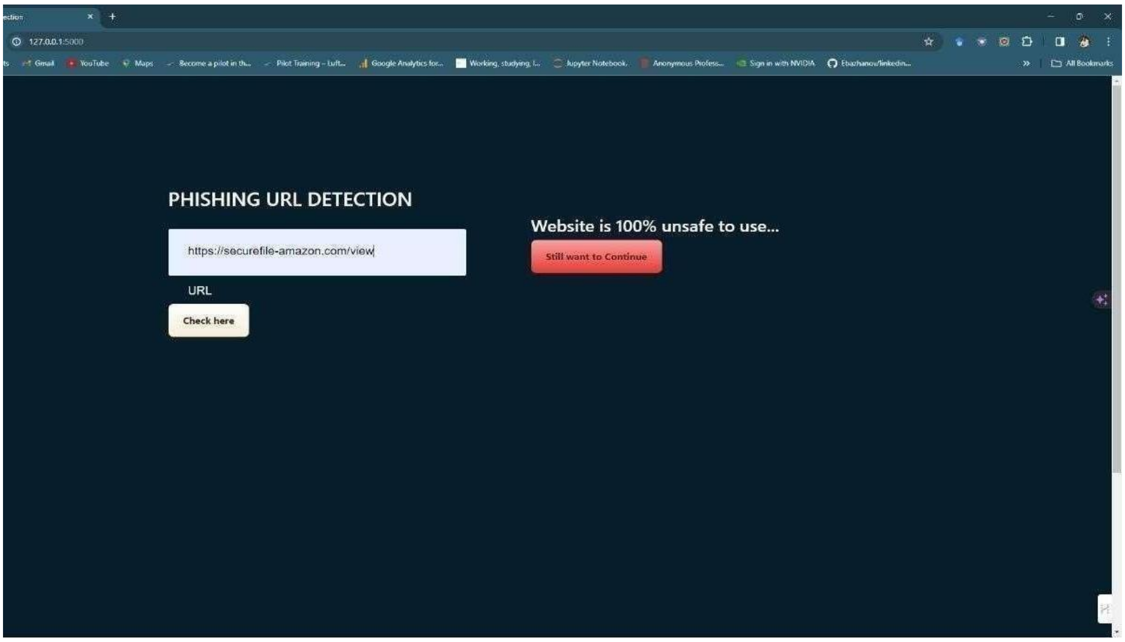


PHISHING URL DETECTION

Enter URL

Website is 100% safe to use...

Detection of Legitimate website



PHISHING URL DETECTION

https://securefile-amazon.com/view

URL

Website is 100% unsafe to use...

Detection Of Phishing Website

6. CONCLUSION AND FUTURE WORK

In closing, evaluating these machine learning classifiers reveals salient performance differences pertinent to application-specific strengths and tradeoffs. The contemporary climate has seen an explosion of deceitful websites, critically threatening individuals and institutions while inflicting widespread harm. This malicious epidemic has become an insidious plague permeating daily life and the integrity of global networks. Using clever impersonation, unscrupulous attackers connive to steal sensitive personal information. Thedeceptive resemblance of these fraudulent sites to trusted entities leaves unsuspecting users vulnerable when inadvertently providing data, believing they interact with legitimate financial institutions. This pressing threat underscores the vital need for advanced phishingwebsite detection. Our project pursues a model leveraging specialized feature extraction techniques focused on URLs, validating extracted characteristics against known datasets to reinforce robustness. We harness algorithms like regression-neighbors, vectors, probabilistic classifiers, decision trees, gradient boosting, and random forests to optimize efficacy. Testing proves commendable performance meeting anticipated aims. This paper focuses on our detection methodology using innovative machine learning. Noteworthy successes include 97.4% accuracy with gradient improvement, and 97.2% with feline boosting, maintaining exceptionally low false positives. As the saying goes, classifiers thrive on plentiful training examples. Future work involves implementing hybrid technologies.

7. REFERENCES

1. "Intelligent Ensemble Learning Approach for Phishing Website Detection Basedon Weighted Soft Voting" by Altyeb Taha November 2021.
2. Mohammad, R.M., Thabtah, F. & McCluskey, L. "Predicting phishing websites based on self-structuring neural network". Neural Compute & Applica 25, 443–458 (2014).
3. Malicious URL Detection using Machine Learning: A Survey Doyen Sahoo,Chenghao Liu, Steven C.H. Hoi [Submitted on 25Jan 2017 (v1), last revised 21 Aug 2019 (this version, v3)].
4. A. Maini, N. Kakwani, R. B, S. M K and B. R, "Improving the Performance of Se- mantic-Based Phishing Detection System Through Ensemble Learning Method," 2021 IEEE Mysuru (Mysore) Sub Section International Conference (Mysuru Con), 2021, pp. 463-469.
5. Cat Boost: gradient boosting with categorical features support Anna Veronika Dorogush, Vasily Ershov, Andrey Gulin [v1] Wed, 24 Oct 2018.
6. Bentéjac, C.Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54, 1937– 1967 (2021).
7. Singh and Meenu, "Phishing Website Detection Based on Machine Learning: A Survey," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 398-404.
8. Christou, O., Pitropakis, N., Papadopoulos, P., McKeown, S., & Buchanan, W. J. (2020). Phishing URL detection through top-level domain analysis: A descriptive approach. arXiv preprint arXiv:2005.06599.
9. Safi, A., Singh, S. (2023). A Systematic Literature Review on Phishing Website Detection Techniques
10. Y. Wei and Y. Sekiya, "Sufficiency of Ensemble Machine Learning Methods for Phishing Websites Detection," in IEEE Access, vol. 10, pp. 124103- 124113,2022, doi: 10.1109/ACCESS.2022.3224781.
11. N. Abdelhamid, F. Thabtah and H. Abdel-jaber, "Phishing detection: A recent intelligent machine learning comparison based on models' content and features,"2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing,China, 2017, pp. 72-77, doi: 10.1109/ISI.2017.8004877
12. N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma and H. Fujita, "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions," in IEEE Access, vol. 10, pp. 36429-36463, 2022, doi: 10.1109/AC- CESS.2022.3151903.
13. Bhavsar, V., Kadlak, A., & Sharma, S. (2020). Study on Phishing Attacks.International Journal of Computer Applications,182(33)
14. Wood, T., Basto-Fernandes, V., BoitenE., & Yevseyeva, I. (2022). Systematic Literature Review: Anti-Phishing Defenses and Their Application to Before-the- click Phishing Email Detection. arXiv

preprint arXiv:2204.13054

15. Sánchez-Paniagua, M., Fidalgo, E., Alegre, E., & Alaiz-Rodríguez, R. (2022). Phishing websites detection using a novel multipurpose dataset and web technologies features. Expert Systems with Applications, 207, 118010.