IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MALICIOUS WEBSITE DETECTION USING MACHINE LEARNING

[1]Subin Sunil, [2]Sreedeep D, [3]Mathews Varkey
[1]Final Year BCA Student, [2]Final Year BCA Student, [3]Final Year BCA Student
[1]PG Department of Computer Applications & AI, [2]PG Department of Computer Applications & AI, [3]PG Department of Computer Applications & AI
[1]Saintgits College of Applied Sciences,Kottayam,India, [2]Saintgits College of Applied Sciences,Kottayam,India, [3]Saintgits College of Applied Sciences,Kottayam,India

***Abstract**:* The role that the World Wide Web plays in enabling illegal actions including spam, fraud, and the distribution of malware is examined in this study three times.The focus is on malicious URL detection using machine learning, with Random Forest and MLP demonstrating high accuracy on a dataset of 2.4M URLs. The study addresses challenges posed by dynamic HTML development and proposes a resilient method for accurate malicious webpage detection in cybersecurity. Our approach, overcoming the limitations of traditional antivirus methods, analyzes webpage characteristics to identify malicious intent. In response to the escalating cyber threat landscape, we present a robust cybersecurity method that classifies websites based on URL features using supervised machine learning. The models, trained on a dataset of malicious and benign URLs, are evaluated with Random Forests, Gradient Boosted Decision Trees, and Deep Neural Networks. The paper also tackles challenges like the availability of training data to attackers and the dynamic nature of malicious websites, introducing a paradigm for detecting and countering induced concept drifts."

***Keywords*:**Malicious URL, detection,machine, learning,random forest,cyber security,detection accuracy,,MLP(multi-layer perceptron),deep neural networks .

I. INTRODUCTION

The paper highlights the surge in web users due to factors like online business migration and improved internet connectivity, leading to increased vulnerability to cyber threats. It proposes a real-time classification system using machine learning to validate URLs, focusing on features [8] such as Lexical, Host-Based, and Content-Based attributes. The study evaluates classifiers like Random Forests and Neural Networks[10], acknowledging challenges in dealing with the dynamic nature of websites and potential adversarial interference.

Additionally, the paper addresses the security concerns associated with malicious DHTML codes in web browsers. Traditional antivirus methods relying on signature-based techniques are deemed ineffective against the dynamic and easily transformable nature of DHTML codes. The paper introduces a machine learning-based method for detecting malicious DHTML codes, emphasizing the importance of feature analysis for classification.

Overall, the paper underscores the need for advanced cybersecurity measures in the face of evolving cyber threats, proposing innovative machine learning solutions to enhance detection accuracy and adaptability.

II. RELATED WORKS

Various research efforts have leveraged machine learning (ML) to combat the proliferation of malicious websites. Ma et al. [1] developed a real-time system that gathers URL features and pairs them with labeled data from a webmail provider to train a classifier with 99% precision in detecting malicious websites. Yadav et al. [2] mined patterns in domain names generated by botnets using statistical methods, achieving satisfactory performance across different scenarios. By employing a taxonomy specifically created for domain generation algorithms (DGAs) to precompute future domain names, Plohmann et al. [3] successfully discovered 5 fraudulent domain names.In 2001, Bergeron et al. presented a static detection method for executable malware. Their approach comprises three main steps: building an intermediate representation, examining programme behaviour focused on security, and confirming vital actions in a stationary way. 2005; Christodorescu et al. provided a malware-detection algorithm that describes the semantics of instructions using a template. Their semantically-aware detection system resists popular obfuscation techniques utilised by adversaries. Blum et al.proposed a confidence-weighted classification system combined with phishing URL detection to dynamically detect present and emergent types of phishing domains, offering improved protection against zero-hour threats. Additionally, Buczak et al. conducted a literature survey on ML and data mining (DM) methods for intrusion detection, addressing the complexity and challenges of using ML/DM for cybersecurity. They provided recommendations on method selection based on data importance and complexity. In order to improve security performance and dependability in Internet of Things applications, researchers have also looked into cutting-edge techniques 1. These include robust certificateless lightweight signature (CLS) methods and stochastic modelling programming for neural networks[12]. Recent studies also focus on feature selection[8] and model optimization. Hall et al. introduced a correlation-based filtering algorithm for feature selection[9], outperforming traditional methods in reducing data dimensionality. Radford et al. demonstrated the efficacy of deep convolutional adversarial pairs[11] in learning hierarchical representations for image tasks. Finally, leveraging ML for web page classification, researchers have systematically explored features of Dynamic HTML (DHTML) code, striking a balance between detection accuracy and resilience to code obfuscation .

III. CLASSIFICATION TECHNIQUES

Pattern classification is indeed fundamental in real-world scenarios, serving to assign categories to new observations based on their attributes. This process is essential across various fields such as finance, healthcare, and marketing. In pattern classification, a supervised approach is commonly employed, where the model learns from historical data that includes both the attributes of the observations and their corresponding known categories or labels. This supervised learning paradigm enables machines to make informed decisions, automate tasks, and solve complex problems across diverse domains.In a supervised learning approach, historical data containing both input features[9] and corresponding output labels are utilized to train the model. This approach is commonly employed in pattern classification tasks, where the goal is to predict the correct category or label for new observations based on their features. Supervised learning algorithms include various techniques such as decision trees, logistic regression, support vector machines, and neural networks. These algorithms differ in their underlying mathematical principles and complexity but share the common goal of learning from labeled data to make predictions on new, unlabeled data.Once trained, the supervised learning model can be evaluated on a separate validation or test dataset to assess its performance and generalization ability. If the model performs well, it can then be deployed to make predictions on new, real-world data, effectively automating the classification task based on patterns learned from historical data.

Classification Models: Several popular classifiers are discussed, including Decision Trees, k-Nearest Neighbors (kNN), Bayesian Networks, Random Forests, Support Vector Machines (SVM), and Multi-layer Perceptrons (MLP).

Decision Tree

Decision Trees are hierarchical structures used for classification tasks, employing attribute tests at internal nodes to classify instances at leaf nodes. The C4.5 algorithm is a popular choice for constructing decision trees, utilizing the concept of information entropy to select attributes that effectively split the data.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (kNN is an instance-based learning method used for classification tasks. In kNN, classification is based on the majority class among the k-nearest neighbors of a given instance.
Here's how kNN works:
Instance Representation: Each instance in the dataset is represented by a vector of features in a multidimensional feature space.
Distance Calculation: To find the k-nearest neighbors of a given instance, the distance between that instance and all other instances in the dataset is calculated. Common distance metrics include Euclidean distance, Manhattan distance, or cosine similarity.
Neighbor Selection: The k instances with the smallest distances to the given instance are selected as its nearest neighbors.
Majority Voting: The class label of the given instance is determined by a majority vote among the class labels of its k-nearest neighbors. That is, the class that occurs most frequently among the neighbors is assigned as the predicted class for the given instance.
Decision Rule: In case of ties (i.e., multiple classes with the same number of neighbors), a decision rule, such as selecting the class with the smallest distance-weighted vote, can be applied.
kNN is a simple yet powerful classification algorithm that does not require training on the entire dataset. However, its performance can be sensitive to the choice of the distance metric and the value of k. Additionally, kNN tends to be computationally expensive for large datasets, as it requires calculating distances to all instances in the dataset for each prediction.

Bayesian Networks

Bayesian Networks are graphical models representing conditional dependencies between variables using a directed acyclic graph (DAG). Inference involves computing posterior probabilities given evidence, while learning involves structure and parameter estimation from data.Bayesian Networks are widely used for modeling complex systems with uncertain relationships between variables, including in fields such as medicine, finance, and artificial intelligence. They provide a principled framework for representing and reasoning about uncertainty and probabilistic dependencies in a structured and interpretable manner.

Random Forests

Random Forests are an ensemble learning method comprising multiple decision trees. It employs bootstrap sampling and random feature selection to construct diverse trees. Each tree in the forest is trained on a random subset of the data and a random subset of features. The final prediction is typically the mode (for classification) or mean (for regression) of the predictions of individual trees. Random Forests are known for their robustness and high performance across various datasets.

Support Vector Machines (SVM)

Support Vector Machines (SVM) are binary linear classifiers that aim to find a hyperplane that best separates classes in a high-dimensional feature space. The optimal hyperplane maximizes the margin, which is the distance between the hyperplane and the nearest data points of each class, known as support vectors. SVMs are effective in dealing with linearly separable data.Moreover, SVMs can be extended to handle non-linear boundaries by using kernel functions. These kernel functions implicitly map the input data into higher-dimensional spaces, where the data may become linearly separable. Common kernel functions include polynomial kernels, Gaussian radial basis function (RBF) kernels, and sigmoid kernels. SVMs are widely used in various applications such as text classification, image recognition, and bioinformatics, due to their ability to handle both linear and non-linear classification tasks.

Multi-layer Perceptrons (MLP)

Multi-layer Perceptrons (MLP) are feed-forward artificial neural networks with multiple layers, typically including input, hidden, and output layers. Each layer consists of interconnected neurons (also called nodes or units), and connections between neurons have associated weights that are learned during training.MLPs are characterized by their ability to learn complex patterns and relationships in data through non-linear transformations. The output of each neuron is calculated by applying an activation function to a weighted sum of the inputs from the previous layer. Common activation functions include sigmoid, tanh, and ReLU (Rectified Linear Unit).The training of MLPs is typically accomplished using the backpropagation algorithm, which involves iteratively adjusting the weights of connections between neurons to minimize the error between the predicted outputs and the true outputs. This is done by propagating the error backwards through the network and updating the weights using gradient descent optimization techniques.MLPs are widely used in various machine learning tasks, including classification, regression, and pattern recognition, due to their flexibility and ability to approximate complex functions. However, they may require careful tuning of hyperparameters and can be prone to overfitting on small datasets.Overall, the document provides a comprehensive overview of different machine learning approaches for pattern classification tasks, along with their advantages .

IV. CONCLUSION

The paper proposes a method for detecting malicious web pages using machine learning. It analyzes the characteristics of malicious web pages and identifies relevant features for machine learning, ensuring resilience against DHTML code obfuscationIn this paper, we propose a robust approach for detecting malicious websites and URLs, achieving a high accuracy of 96.4% on the test dataset. Our algorithm also addresses attempts by attackers to circumvent detection methods, successfully detecting concept drifts. Statistical tests confirm the significance of our method's performance, with Random Forest ranking highest in prediction accuracy, followed by MLP. Random Forest demonstrates balanced prediction results and maximizes malicious URL detection. Feature selection based on features with the highest absolute Pearson coefficient outperforms other sets significantly, with no significant difference observed between highly correlated binary features and only real-value features

V. REFERENCE

1. J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Learning to detect malicious urls,” ACM Transactionson Intelligent Systems and Technology (TIST), vol. 2,no. 3, pp. 1–30, 2011.
2. S. Yadav, A. K. K. Reddy, A. Reddy, and S. Ranjan,“Detecting algorithmically generated malicious domain names,” in Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. ACM, 2010, pp.48–61.
3. D. Plohmann, K. Yakdan, M. Klatt, J. Bader, and E. Gerhards-Padilla, “A comprehensive measurement study of domain generating malware,” in 25th USENIX Security Symposium (USENIX Security 16), 2016, pp. 263–278.
4. A. Blum, B. Wardman, T. Solorio, and G. Warner,“Lexical feature based phishing url detection using online learning,” in Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security. ACM, 2010, pp.54–60.
5. Christodorescu, M., Jha, S., Seshia, S. A., Song, D., & Bryant R. E. (2005). Semantics-aware malware detection. In Proceedings of the IEEE symposium on security and privacy (pp. 32–46). Oakland.
6. A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2015.
7. A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generativeadversarial networks,” arXiv preprint arXiv:1511.06434, 2015
8. NV Balaji,2021, Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach, IOP conference series: materials science and engineering
9. Ambily Merlin Kuruvilla ,2019, Predicting Diabetes Mellitus Using Feature Selection And Classification Techniques In Machine Learning Algorithms, Karpagam JCS
10. Ambily Merlin Kuruvilla,2021, Improved Artificial Neural Network Through Metaheuristic Methods And Rough Set Theory For Modern Medical Diagnosis, Indian Journal of Computer Science and Engineering.
11. PC Sherimon,2023, Customized Hybrid Deep Learning Model for Road Accident Detection Based on CCTV Images, 2023 IEEE International Performance, Computing, and Communications Conference (IPCCC).

12. Manjima Sree, 2023, Estimation of Learners' Levels of Adaptability in Online Education Using Imbalanced Dataset, 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE) DOI: 10.1109/RASSE60029.
13. Bergeron, J., Debbabi, M., Desharnais, J., Erhioui, M. M., Lavoie, Y., & Tawbi N. (2001). Static detection of malicious code in executable programs. In Proceedings of the symposium on requirements engineering for information security, India