Cyber Protector: Advanced Machine Learning Algorithm for Phishing Website Monitoring

Ms. S. Indhumathi [1], S.Santhosh Adaikalaraj[2], K. Reshma[3], V. Abinash[4]
[1] *Ms. S. Indhumathi, Department of Software System, Sri Krishna Arts and Science College*
[2,3,4] *PG student of Computer Science, Sri Krishna Arts and Science College*

Abstract- Phishing attacks are one of the most pervasive cyber threats in today's digital era, targeting individuals and organizations by mimicking legitimate entities to steal sensitive data. As phishing tactics grow more sophisticated, traditional detection methods struggle to keep pace. Cyber Protector addresses this challenge by employing machine learning techniques, specifically Gradient Boosting and XGBoost algorithms, to monitor and detect phishing websites effectively. This solution offers superior detection accuracy, reduced false positives, and adaptability, ensuring a secure browsing experience. By integrating real-time monitoring and continuous learning, Cyber Protector presents a scalable approach to combating evolving phishing threats.

I. INTRODUCTION

Phishing attacks have become a major cybersecurity concern, with attackers leveraging increasingly sophisticated tactics to deceive users. Traditional rule-based detection systems rely on static databases or signatures, rendering them ineffective against novel threats. This paper introduces Cyber Protector, a cutting-edge web application that leverages machine learning, specifically XGBoost, to provide an intelligent and dynamic solution for detecting phishing websites. Cyber Protector achieves high accuracy through comprehensive feature engineering and advanced classification techniques. The system supports real-time detection and integrates continuous learning to adapt to new patterns of phishing attacks, thus ensuring an enhanced level of cybersecurity for end-users.

II. DESCRIPTION

Cyber Protector applies a robust machine learning framework to detect phishing websites. The core of the system is powered by XGBoost, an efficient gradient boosting algorithm known for its scalability and high performance in classification tasks. The system analyse a diverse set of features, such as URL patterns, domain characteristics, content-based attributes, and network metadata, to classify websites as either phishing or legitimate. Additionally, the web application offers a user-friendly interface that allows users to input URLs for real-time analysis. The application also maintains a repository of detected phishing websites to enhance its training dataset, facilitating continuous improvement.

III. DATASET COLLECTION

The dataset used for training and testing consists of URLs from reliable sources such as Phish Tank, Open Phish, and Alexa's top websites. Preprocessing steps include cleaning and normalizing the dataset, removing duplicates, and extracting relevant features. The dataset contains a balanced mix of phishing and legitimate websites, ensuring that the model performs well across diverse scenarios. Advanced feature engineering techniques are employed to extract key indicators, such as URL length, number of subdomains, and HTTPS usage. These features are critical for training the model to identify patterns indicative of phishing attempts.

IV. EXISTING SYSTEM

Existing systems for phishing detection primarily rely on blacklist-based approaches or simple heuristic rules. These systems are limited in scope and fail to adapt to new and evolving phishing techniques. Furthermore, they often suffer from high false-positive rates and require frequent manual updates to remain effective. Unlike these static solutions, Cyber Protector leverages machine learning to dynamically learn from new data, offering a robust and adaptive alternative for phishing detection.

V. PROPOSED SYSTEM

Cyber Protector employs XGBoost to classify websites based on extracted features. XGBoost's ability to handle large datasets and its resistance to overfitting make it an ideal choice for this task. The system is designed with the following components:

- Feature Extraction: Extracts and analyzes attributes such as URL structure, domain age, DNS records, and HTML content.
- Model Training: Uses labeled datasets to train the XGBoost model for classification. Hyperparameter tuning ensures optimal model performance.
- Real-Time Analysis: Accepts user-input URLs and classifies them on the fly.
- Continuous Learning: Periodically retrains the model with new data, improving detection capabilities over time.

The system also incorporates a browser extension for seamless integration, enabling users to receive alerts while browsing.

VI. LITERATURE REVIEW

Studies on phishing detection highlight the limitations of traditional systems and the potential of machine learning models. Research demonstrates that ensemble models like XGBoost provide superior performance in classification tasks due to their ability to handle complex patterns in large datasets. For instance, high-accuracy phishing detection systems often utilize features such as lexical analysis of URLs, behavioral patterns of phishing sites, and DNS anomalies. This paper builds on these findings by implementing an XGBoost-based system tailored for real-time detection of phishing websites.

VII. DETECTION METHODOLOGY

The detection process begins with feature extraction, where the system analyzes characteristics such as URL length, presence of suspicious keywords, and domain registration details. The extracted features are then passed to the XGBoost model for classification. The model uses gradient boosting to optimize decision trees, enabling it to handle imbalanced datasets effectively. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess the model's performance. Cyber Protector's architecture ensures low latency and high throughput, making it suitable for real-time applications.

VIII. WORKFLOW

The system operates in a structured workflow:

1. Collect and preprocess dataset, including feature engineering and normalization.
2. Train the XGBoost model with labeled phishing and legitimate websites.
3. Deploy the model to a Flask-based backend for handling URL inputs.
4. Develop a browser extension and a web interface for user interaction.
5. Monitor model performance and periodically update it with new data.

XI. RESULTS

Cyber Protector was tested on a benchmark dataset, achieving an accuracy of over 95%. The system demonstrated a significant reduction in false positives compared to traditional methods. Real-time analysis was completed within milliseconds, showcasing the system's efficiency. Key insights from testing include the importance of HTTPS-related features and the effectiveness of XGBoost in handling imbalanced data. The browser extension successfully integrated detection capabilities into the user's browsing experience, further enhancing usability.

X. CONCLUSION

This paper presents Cyber Protector, a machine learning-based web application for detecting phishing websites. By leveraging XGBoost, the system achieves high detection accuracy and adaptability. Its real-time analysis, user-friendly interface, and continuous learning capabilities make it a comprehensive solution for combating phishing attacks. Future enhancements include expanding feature sets, integrating threat intelligence feeds, and implementing real-time collaboration with cybersecurity platforms.

REFERENCE

[1] Fenton, N., & Pfleeger, S. L., *Software Metrics: A Rigorous and Practical Approach*.

[2] Pressman, R. S., *Software Engineering: A Practitioner's Approach*.

[3] Chidamber, S. R., & Kemerer, C. F., *A Metrics Suite for Object-Oriented Design*.