WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Cyber Breach Prediction System: A Machine Learning Approach

Mohammed Hussain Abid Khan [1] , Md Vazahat Ali [2], Mohammed Mawiz Ahmed [3]

Department of IT, Nawab Shah Alam Khan College of Engineering and Technology , Hyderabad , India
Email: mdhussainabidkhan018@gmail.com

ABSTRACT :

As cybersecurity threats become increasingly complex, there is a critical need for effective prediction systems to prevent cyber hacking breaches. This study introduces an innovative approach using Machine Learning, specifically the Random Forest Classifier, to predict potential cyber breaches based on URLs. Through extensive training and evaluation, the system demonstrated outstanding performance, achieving a training accuracy of 99% and a test accuracy of 91%. These results highlight the system's capability to identify patterns and differentiate between legitimate and malicious instances, offering a reliable tool for cybersecurity.

Keywords: Cyber Breach, Random Forest Classifier, Machine Learning, URL's.

1. INTRODUCTION :

The "Cyber Breach Prediction System: A Machine Learning Approach" project utilizes Machine Learning to bolster cybersecurity defenses. It addresses the shortcomings of traditional methods by employing a Random Forest Classifier on a curated dataset. This system enhances accuracy in detecting cyber threats and minimizes false positives, utilizing Python libraries such as scikit-learn, numpy, and pandas.

1.1 Common Machine Learning Algorithms for Cyber Breach Prediction:

Random Forest Classifier:

The Random Forest algorithm is a robust tree-based learning method in Machine Learning. It constructs multiple decision trees during training, each built from a random subset of the dataset and features. This randomness introduces variability among the trees, reducing overfitting and enhancing overall predictive performance. For predictions, the algorithm aggregates the results from all trees, providing stable and precise outcomes. Random forests are widely used for both classification and regression tasks, known for handling complex data, reducing overfitting, and delivering reliable forecasts in diverse scenarios.

2 Literature Survey

1. Research Challenges at the Intersection of Big Data, Security, and Privacy*(IGI-GLOBAL, 2019)
Authors: Kantarcioglu M and Ferrari E
Big data can generate significant value across industries, but security and privacy issues must be addressed. Appropriate privacy-aware access control policies are crucial, especially during data storage and sharing.

2. Digging Deeper into Data Breaches: An Exploratory Data Analysis of Hacking Breaches Over Time (UGC, 2019)
Authors: H. Hammouchi, O. Cherqi, G. Mezzour, M. Ghogho, and M. El Koutbi
This study analyzes over 9000 data breaches since 2005, highlighting the persistent threat of hacking breaches. It identifies the most targeted organizations and examines changes in hacker interests over time.

3. A Machine Learning-Based Malware Classification Framework (IEEE, 2023)
Authors: S. Depuru, P. Hari, P. Suhaas, S. R. Basha, R. Girish, and P. K. Raju
This research proposes a neural network model for malware classification, evaluating various representation methods and CNN models to achieve high accuracy.

3. Requirements

3.1 SYSTEM REQUIREMENTS:

3.1.1 HARDWARE REQUIREMENTS:

- System : Pentium i3 Processor.
- Hard Disk : 500 GB.
- Monitor : 15'' LED
- Input Devices : Keyboard, Mouse
- Ram : 4 GB

3.1.2 SOFTWARE REQUIREMENTS:

- Operating system : Windows 10.
- Coding Language : Python 3.10.9.
- Web Framework : Flask.

4. System Analysis and Design

4.1 Modules

i. User Interface Module

This module handles user interactions, including uploading an Excel sheet with URLs and displaying results like accuracy, recall, and precision.

ii. URL Processing Module

Responsible for processing URLs from the Excel sheet, this module includes tasks such as URL normalization, validation, and formatting.

iii. Machine Learning Module

This module covers training and testing the machine learning model, using features from legitimate and phishing websites to classify new URLs.

iv. Prediction Module

This module focuses on the model's prediction and detection capabilities.

v. Evaluation Module

Evaluates the model's performance using metrics such as accuracy, recall, and precision, providing feedback on its effectiveness.

4.2 Architecture

The system's architecture is designed to tackle cyber hacking breaches using advanced Machine Learning techniques, particularly the Random Forest Classifier. Implemented in Python, the system comprises various components that work together to ensure effective threat identification while minimizing false positives. A carefully curated dataset of URLs, balanced between phishing and legitimate URLs, is crucial for comprehensive threat coverage. Rigorous training and evaluation processes result in exemplary performance metrics, with the Random Forest Classifier achieving a remarkable training accuracy of 99%.

5. Conclusion

The "Cyber Breach Prediction System: A Machine Learning Approach" project marks a significant advancement in cybersecurity. By leveraging Python and the Random Forest Classifier algorithm, the system demonstrates exceptional accuracy and effectiveness in predicting and detecting cyber hacking breaches. It addresses the limitations of previous rule-based and signature-based methods by integrating advanced machine learning techniques, ensuring adaptability to evolving cyber threats, and reducing false positives.

REFERENCES :

1. L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557-570, 2002.
2. A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in Proceedings of the IEEE Symposium on Security and Privacy, 2008, pp. 111-125.
3. Y. Yang and D. M. Blei, "A supervised topic model for linked document collections," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 527-536.
4. D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature, vol. 393, no. 6684, pp. 440-442, 1998.
5. T. Y. Chen, F. -C. Kuo, R. G. Merkel, and T. Tse, "Adaptive random testing: The ART of test case diversity," Journal of Systems and Software, vol. 83, no. 1, pp. 60-66, 2010.
6. C. Dwork, "Differential privacy," in Proceedings of the International Colloquium on Automata, Languages, and Programming, 2006, pp. 1-12.
7. J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.