IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, November 2024

Cyber Hacking Breaches Detection Using Machine Learning

Prof. Kurhe. P. V[1], Aishwarya Antre[2], Piyush Deore[3], Aniket Thakare[4], Tushar Salunkhe[5]
Guide, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2,3,4,5]
SND College of Engineering & Research Center, Yeola, Maharashtra, India

Abstract: *Cyber hacking breach prediction is one of the emerging technologies and detecting and predicting breaches through computer algorithms has become a very challenging task To make malware detection more effective, scalable and more efficient than traditional system calls human involvement Machine learning to be used for breach detection and prediction The main goal is a series of cyber hacking attacks each of which will harm the person's information and financial reputation. Government and non-profit organizations' data, such as user and company information,, can be compromised, posing a risk to their finances and reputation if they collect information from websites and social networks it can trigger a cyber attack. Organizations such as the healthcare sector are capable of holding sensitive information that must be handled discreetly and securely. Data breaches can lead to identity theft, fraud, and other losses. The findings show that 70% of breaches affect a wide range of organisations, including healthcare providers. The investigation indicates a possible data breach. Due to the heavy usage of computer programs and security on the host and network, there is a risk of data breach. Machine learning can be used to detect these attacks. Research uses machine learning models to protect against web security flaws. The data set is available from the Privacy Rights Clearing House. Teaching employees how to use modern security measures can reduce data breaches. This can help to understand attack detection and data security. Machine learning models such as Random Forest, Decision Tree, k-means and Multi- layer Perceptron are used to predict data violations.*

Keywords: Cyber hacking breaches, Machine learning, Algorithms, Prediction

I. INTRODUCTION

Over the years, companies have become a prime target for cyberattacks, particularly ransomware, which cause significant financial and reputational damage. Millions of cyberattacks occur every day, making it increasingly difficult to maintain system security, including protecting corporate and personal data. This research focuses on three main objectives: developing prediction techniques for cybercrime using actual cybercrime data, testing whether the available data can identify cybercriminals, and analyzing the impact of these attacks on organizations.

Cyber attacks often lead to data breaches, especially in industries such as healthcare, where sensitive information is at risk. Data breaches can lead to identity theft, fraud, and lawsuits against organizations. Despite advances in technology and data collection, the risk of breaches remains. The Privacy Rights Clearinghouse (PRC) listed several breaches from 2005 to 2019, affecting a wide range of organizations. Breach detection using traditional methods is complicated by the large amount of information involved. Technical instruction, where there are motorways and traffic, and the discipline of carelessness and negligence and the discipline of lawlessness, where the inner dhokhadhadis are trained. Have The PRC dataset is used to analyze these breaches, revealing that negligent breaches are more often due to human error than malicious ones but the analysis specifically examines hacking-related breaches to understand its impact and improve detection methods. Breach detection involves constant monitoring of networks for access and potential incidents, but existing systems struggle with aggressive anecdotes.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, November 2024

II. RELATED WORK

Modelling and predicting cyber hacking breaches

AUTHORS: M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu

Analyzing cyber incident data sets is an important method for deepening our understanding of the evolution of the threat situation. This is a relatively new research topic, and many studies remain to be done. In this paper, we report a statistical analysis of a breach incident data set corresponding to 12 years (20052017) of cyber hacking activities that include malware attacks. We show that, in contrast to the findings reported in the literature, both hacking breach incident inter-arrival times and breach sizes should be modeled by stochastic processes, rather than by distributions because they exhibit autocorrelations. Then, we propose particular stochastic process models to, respectively, fit the inter-arrival times and the breach sizes. We also show that these models can predict the interarrival times and the breach sizes. In order to get deeper insights into the evolution of hacking breach incidents, we conduct both qualitative and quantitative trend analyses on the data set. We draw a set of cybersecurity insights, including that the threat of cyber hacks is indeed getting worse in terms of their frequency, but not in terms of the magnitude of their damage.

A self-adaptive deep learning-based system for anomaly detection in 5G networks

AUTHORS**:** Fernandez Maimo et al.

The upcoming fifth-generation (5G) mobile technology, which includes advanced communication features, is posing new challenges on cybersecurity defense systems. Although innovative approaches have evolved in the last few years, 5G will make existing intrusion detection and defense procedures become obsolete, in case they are not adapted accordingly. In this sense, this paper proposes a novel 5G-oriented cyberdefense architecture to identify cyberthreats in 5G mobile networks efficient and quickly enough. For this, our architecture uses deep learning techniques to analyze network traffic by extracting features from network flows. Moreover, our proposal allows adapting, automatically, the configuration of the cyberdefense architecture in order to manage traffic fluctuation, aiming both to optimize the computing resources needed in each particular moment and to fine tune the behaviour and the performance of analysis and detection processes. Experiments using a well-known botnet data set depict how a neural network model reaches a sufficient classification accuracy in our anomaly detection system. Extended experiments using diverse deep learning solutions analyze and determine their suitability and performance for different network traffic loads. The experimental results show how our architecture can self-adapt the anomaly detection system based on the volume of network flows gathered from 5G subscribers' user equipments in real-time and optimizing the resource consumption.

Research Challenges at the Intersection of Big Data, Security and Privacy.

AUTHORS**:** Kantarcioglu M and Ferrari E

As reports from McKinsey Global Institute (McKinsey et al., 2011) and the World Economic Forum (Schwab, 2016) suggest, capturing, storing and mining “big data” may create significant value in many industries ranging from health care to government services. For example, McKinsey estimates that capturing the value of big data can create $300 billion dollar annual value in the US health care sector and $600 billion dollar annual consumer surplus globally (McKinsey et al., 2011). Still, several important issues need to be addressed to capture the full potential of big data. As shown by the recent Cambridge Analytica scandal (Cadwalladr and Graham-Harrison, 2018) where millions of users profile information were misused, security and privacy issues become a critical concern. As big data becomes the new oil for the digital economy, realizing the benefits that big data can bring requires considering many different security and privacy issues. This in return implies that the entire big data pipeline needs to be revisited with security and privacy in mind. For example, while the big data is stored and recorded, appropriate privacy-aware access control policies need to be enforced so that the big data is only used for legitimate purposes. On the other hand, while linking and sharing data across organizations, privacy/security issues need to be considered. Below, we provide an overview of novel research challenges that are at the intersection of cybersecurity, privacy and big data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, November 2024

Digging deeper into data breaches: An exploratory data analysis of hacking breaches over time

AUTHORS: H. Hammouchi, O. Cherqi, G. Mezzour, M. Ghogho, and M. El Koutbi

Data breaches represent a permanent threat to all types of organizations. Although the types of breaches are different, the impacts are always the same. This paper focuses on analyzing over 9000 data breaches made public since 2005 that led to the loss of 11,5 billion individual records which have a significant financial and technical impact. Also, since the most devastating breaches are hacking breaches, we shed the light on type to unveil the most targeted organizations, and examine how the interest of hackers changes over time. On the other hand, the breaches caused by human factor are decreasing which can be explained by the awareness of employees and the application of security standards. This study would improve the state of knowledge about hacking breaches and help in securing organizations' data by prioritizing the most attacked sectors.

A Machine Learning based Malware Classification Framework

AUTHORS: S. Depuru, P. Hari, P. Suhaas, S. R. Basha, R. Girish and P. K. Raju

Rapid developments in the field of computer technology introduced new factors such as sophistication, effectiveness, and increased connectivity. The evolution in computer technology involuntarily leads to the evolution of malware as more devices are connected, leading to the development of malware that can spread through interconnected devices. The increase in the use of computers in almost every part of our life, people rely more on them by storing all the crucial details and information on the computers. This will put us in a vulnerable state where cybercriminals can attack us and exploit us using our data against us. These attacks can be performed by injecting various kinds of malware into our computers which puts us at risk. This creates the necessity to discover the different malicious attacks and classify them before it takes place to be safe. Representing the malware in a visual representation makes the process of classification much easy and more efficient. This research study proposes a neural network model to evaluate the performance of different combinations of various methods available for representing malware and different models of convolutional neural networks (CNN) and selects the best model that has highest accuracy.

III. SYSTEM ARCHITECTURE AND PROPOSED SYSTEM

Proposed System

The proposed system, "Cyber Hacking Breaches Prediction and Detection Using Machine Learning," introduces a cutting-edge approach to address the challenges of cyber threat detection and prediction. Leveraging the power of Python programming language and the Random Forest Classifier algorithm, this system aims to provide robust, accurate, and adaptive cybersecurity measures.

The core of the proposed system is the Random Forest Classifier, a powerful ensemble learning algorithm widely used for classification tasks. It combines multiple decision trees, each trained on a different subset of the dataset, to improve accuracy and reduce overfitting. The Random Forest model is well-suited for this project due to its ability to handle high-dimensional datasets with numerous features, like the 87 extracted features from the 5457 URLs in the dataset.

The system is developed using Python, a popular and versatile programming language. Python's extensive libraries and frameworks, such as scikit-learn and pandas, make it an ideal choice for implementing machine learning algorithms, data manipulation, and feature engineering.

The dataset used in the proposed system contains 5457 URLs, perfectly balanced between legitimate and phishing URLs, with each category comprising 50% of the data. This balanced representation ensures that the model learns equally from both classes, reducing the risk of bias and improving the system's ability to generalize effectively.

The Random Forest model is trained on the balanced training dataset. Each decision tree in the ensemble learns from a different subset of the data, promoting diversity and reducing the risk of overfitting. The model uses the 87 extracted features to learn patterns associated with both legitimate and phishing URLs. After training, the system evaluates the Random Forest Classifier using the test dataset. The accuracy achieved during evaluation provides valuable insights into the model's performance and its ability to predict cyber hacking breaches accurately

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, November 2024

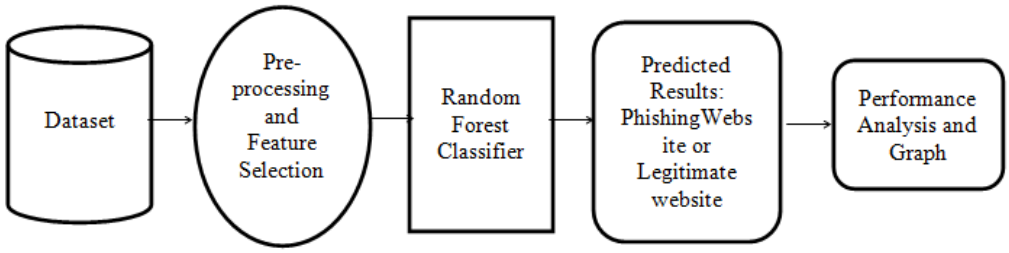
Hardware Requirements

- System :Pentium i3 Processor.
- Hard Disk : 500 GB.
- Monitor: 15'' LED
- Input Devices: Keyboard, Mouse
- Ram: 6 GB

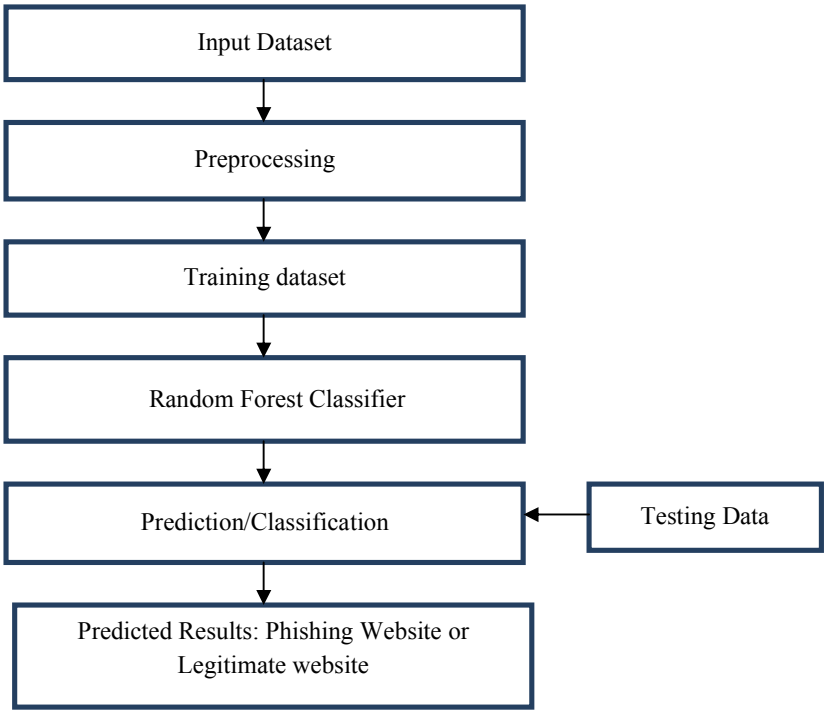
Software Requirements

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

System Architecture:



Data Flow Diagram



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, November 2024

IV. ADVANTAGES AND DISADVANTAGES

Advantages

- Improved Detection Accuracy
- Real-Time Threat Detection
- Automated Processes
- Adaptability to New Threats
- Enhanced Threat Prediction
- Reduction in False Positives

Disadvantages

- Data Requirements
- Complexity and Expertise
- High Computational Cost
- Vulnerability to Adversarial Attacks

V. CONCLUSION

In conclusion, the project "Cyber Hacking Breaches Prediction and Detection Using Machine Learning" provides a pioneering solution for combating cyber threats. By utilizing Python and the Random Forest Classifier, the system achieves high accuracy, adaptability, and significantly reduced false positives, enhancing trust in its predictions. With a balanced dataset of 5457 URLs and feature-rich extraction, the model achieved 99% training accuracy and 91% test accuracy, effectively distinguishing between legitimate and malicious URLs. This adaptable system addresses the limitations of rule-based methods, making it future-proof and capable of detecting emerging threats. Overall, the project represents a sophisticated, scalable advancement in cybersecurity defenses.

VI. FUTURE WORK

While the proposed system has shown strong performance, several enhancements could further improve its capabilities. Exploring advanced machine learning models, like Gradient Boosting, Neural Networks, or Deep Learning, could boost accuracy and generalization. Implementing online learning would allow continuous adaptation to real-time threats, and adding anomaly detection could identify previously unknown breaches. Ensemble methods combining multiple models could enhance robustness, while expanding feature engineering could capture new threat characteristics. Integrating real-time automation for response actions would mitigate attack impacts quickly. Large-scale deployment in real-world environments, continuous dataset updates, domain-specific knowledge from cybersecurity experts, and collaboration with industry partners could validate and strengthen the system's practical effectiveness in diverse scenarios.

REFERENCES

[1] M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu, “Modelling and predicting cyber hacking breaches,” IEEE Trans. Inf. Forensics Security, vol. 13, no. 11, pp. 2856–2871, 2018.
[2] IBM. (2019). Cost of a data breach report. IBM Security, 76. [Online].
Available https://www.ibm.com/downloads/cas/ZBZLY7KL
[3] Fernandez Maimo et al., “A self-adaptive deep learning-based system for anomaly detection in 5G networks,” IEEE Access, vol. 6, pp. 7700–7712, 2018.
[4] Kantarcioglu M and Ferrari E (2019) Research Challenges at the Intersection of Big Data, Security and Privacy.
[5] Verizon, “Data breach investigations report,” 2019. [Online].
Available:https://enterprise.verizon.com/resources/reports/dbir/
[6] H. Hammouchi, O. Cherqi, G. Mezzour, M. Ghogho, and M. El Koutbi, “Digging deeper into data breaches: An exploratory data analysis of hacking breaches over time,” Procedia Computer Science, vol. 151, pp. 1004–1009, 2019.
[7] rack T. Majority of malware analysts aware of data breaches not disclosed by their employers. http://www.threattracksecurity.com/press-release/majority-of-malware-analysts-aware-of-databreaches-not-dis closed-by-their-employers.aspx