International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Cyber Hacking Breaches Predic Tion and Detection Using Machine Learning

Subalakshmi. M[1], Theyjakshaya. Ds[2], Vishnu Vardhini. Am[3]

[1,2,3]Department of Information Technology, Batchelor of Technology, Sri Shakthi Institute of Engineering andTechnology (Autonomous) Coimbatore-641062

ABSTRACT

Predicting cyber-hacking breaches through ML (random forest classifier) is one of the newest technologies, and utilizing computer algorithms to identify and anticipate breaches have shown to be a difficult challenge. The primary focus of employing machine learning for breach detection and prediction is to make malware detection more rapid, scalable, and efficient than traditional systems that require human input. Websites that have the potential to launch a cyberattack can provide the information. Data breaches might end in identity theft, fraud, and other damages. According to the data, 70% of breaches have an impact on many firms. The analysis shows the probability of a breach in data. Security breaches are becoming more likely as a result of increasing use of computer applications and host and network security.

INTRODUCTION

An increasing number of companies are keen in applying machine learning (ML) for the prediction and detection of cyber hacking breaches due to the digital age, ML represents a part of artificial intelligence that is adept at analysing large datasets to identify abnormalities and patterns that may go unnoticed by human analysts. Incorporating ML algorithms into cybersecurity frameworks can significantly improve an organization's ability of avoiding cyber hacking breaches, that can cause substantial financial losses, reputational damage, and compromise of sensitive information.

OBJECTIVE

This project intends to build a user-friendly and reliable system for anticipating and identifying cyber-attacks. The system will make use of a Random Forest classifier for predicting threats appropriately. It will also feature an HTML and Flask web interface which makes it simple for users to interact with the system. By offering real-time analysis and early warning capabilities, the system seeks to improve organizational cybersecurity through providing proactive measures against prospective threats. In the end, the project will protect digital assets and enhance overall cybersecurity resilience by ensuring data security, scalability, overall acceptance of best practices.

LITERATURE REVIEW

Machine learning is essential for identifying and preventing cyberattacks, as demonstrated by recent developments in the field of cybersecurity. When it comes to handling complicated datasets and achieving classification accuracy, Random Forest an automated learning technique has showed to be especially successful. Random Forest's resilience to overfitting and its suitability for use in cybersecurity scenarios

are highlighted in studies notably Breiman's (2001) seminal studies. Using UNSW-NB15 and other data sets, Moustafa and Slay (2016) proved its superiority over standard methods in intrusion detection and showed how well it could detect deviations in networks.Current work on following through machine learning models in accessible formats has emphasized the benefits of integrating these models into web frameworks such as Flask, resulting in real-time threat detection and user-friendly interfaces. It still remains as a complication with guaranteeing data, though.

METHODOLOGY

Preparing and Gathering Data: Compile a range of information from publicly available datasets, system logs, and network logs. Transform categorical data into numerical format, deal with missing values, and eliminate duplicates to clean up the data. Divide the dataset into training, validation, and test sets after extracting pertinent characteristics.

Model Choosing: Owing to its resilience and capacity to manage high-dimensional data, choose for the Random Forest classifier. Choose the model that works best for cyber threat prediction after evaluating several models using performance indicators.

Training of Models: Use methods such as Grid Search or Random Search to optimize the model's parameters. To avoid overfitting, validate the performance of the model using k-fold cross-validation. Using the pre-processed training data, train the Random Forest model and make adjustments based on validation outcomes.

Evaluation of the model: Metrics including accuracy, precision, recall, and F1 score are used to evaluate the performance of the model. To assess the generality of the model, test it on different test sets and real-world events. Take user comments into account and make necessary model adjustments to increase efficacy.

Ongoing Enhancement: Update the model frequently with fresh information to accommodate changing risks. Keeping an eye on the model's performance and tweak its parameters as necessary. Make sure it can adapt and scale to meet growing data volumes and emerging risks. Give users thorough instructions and training so they can maintain and operate the device efficiently.

EXISTING METHODS

Presently available machine learning approaches for anticipating and identifying cyber hacking breaches include a range of approaches designed to tackle distinct categories of cyber threats. Techniques for detecting anomalies, including autoencoders and isolation forests, can be used to find unknown attacks since they can spot departures from the usual. They are especially helpful in situations when more novel threat patterns may go undetected via conventional methods.

Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) are two examples of signature-based techniques which perform best in contexts with well-defined attack signatures because they classify threats based on known attack patterns. However, because these techniques rely on established signatures, they are not resistant to zero-day flaws.

Deep learning models, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), are particularly effective in identifying sophisticated cyber threats because they have the capacity to identify intricate patterns in data, such as anomalies in network traffic or consecutive log entries.In order to improve detection skills and produce deeper and more accurate threat detection solutions, hybrid approaches that combine various methodologies utilize each technique's advantage and keep up with the

changing cybersecurity threat landscape, these techniques improve cyber breach detection and prediction when taken as a whole.

DISADVANTAGES

Data Dependency: Accurate forecasts depend on high-quality data, and data management and compliance may be hampered by privacy issues.
Resource Intensity: Training models can be expensive due to the large amount of computing power needed as well as ongoing maintenance.
Accuracy Issues: Models may lead to major security issues through producing false positives or negatives and having challenges generalizing to new threats.
Integration Challenges: Compatibility problems can arise during the complicated and expensive process of integrating with current systems.

PROPOSED SYSTEM

By utilizing a strong machine learning framework, the suggested method improves cyber breach detection. To provide thorough threat coverage, it gathers information from a variety of sources, such as network logs and threat intelligence feeds. While a Random Forest classifier successfully detects threats, automated preprocessing enhances the quality of the data.
Continuous model training ensures adaptation to new threats, while real-time data streams enable prompt threat identification and action. The system integrates automated reactions to quickly mitigate threats and offers real-time monitoring and instant alarms.
The system can handle massive volumes of data and smoothly integrates with current cybersecurity technologies because it is built for scalability. Ongoing progress is ensured by thorough examination of metrics such as accuracy and precision in conjunction with user feedback. Effective deployment and administration are supported by documentation and training.

SYSTEM REQUIREMENTS

Hardware Requirements:

- Devices: An modern computer.
- CPU: Intel Core i5 or above.
- OS: Linux, macOS, or Windows.
- RAM: For deployment, 16 GB is advised, however at least 8 GB is needed.
- Storage: 20 GB for datasets/logs, 500 MB for software.
- Network Connection: Reliable network.

Software Requirements:

- Python
- HTML
- CSS

Libraries:

- NumPy
- Pandas
- Scikit-learn

Module Description:

1. **DATA COLLECTION AND PRE-PROCESSING:**

- Gather information from a variety of sources, such as system and network logs and the data is clean, transformed, and ready for analysis.
- Deal with missing values, eliminate duplicates and transform categorical data into numerical representation.

2. **MODEL SELECTION:**

- Select the proper machine learning methods for cyber hacker breach prediction
- Compare and contrast various models according to performance indicators and task appropriateness.

3. **MODEL TRAINING:**

- Using the pre-processed data, train the chosen machine learning model
- Improve model parameters with methods such as hyperparameter tuning

4. **MODEL EVALUATION:**

- Use a variety of measures to evaluate the performance of the trained model.
- To guarantee accuracy and resilience, validate the model using different test sets.

5. **ONGOING ENHANCEMENT:**

- Establish systems for regular model upgrades and modifications.
- Over time, improve the system's efficacy by incorporating fresh data and user feedback.

CONCLUSION

Our organization's protection against cyber hacking breaches has been greatly strengthened by the implementation of this machine learning-based solution, which provides a strong and dependable framework for detection and prediction. Employing HTML, Flask, Python, and a Random Forest Classifier, we have boosted our capacity to protect resources and retain a secure working environment. Going ahead, we'll be concentrating on ongoing growth. We'll strive to be attentive in adjusting to the shifting cybersecurity landscape, improve our models, and broaden our sources of information to capture emerging threats. By keeping our defence ahead of new threats, we will safeguard the integrity and resilience of our organization against intrusions.

REFERENCES

1. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
2. Moustafa, N., & Slay, J. (2016). The evaluation of network anomaly detection systems Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 dataset. *Information Security Journal: A Global Perspective*, 25(1-3), 18-31.
3. Flask. (n.d.). *Flask Documentation*. Retrieved from https://flask.palletsprojects.com/en/2.0.x/
4. W3Schools. (n.d.). *HTML Tutorial*. Retrieved from https://www.w3schools.com/html/
5. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.