International Journal of Engineering &Technology, 13 (2) (2024) 367-371

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Predicting cyber hacking breaches using machine learning: a proactive approach to cyber security

Ms. Manisha B. Thombare[1]*, Mr. Sunil T. Datir[1]

[1]Assistant Professor, MVPS's KBTCOE, Nashik
**Corresponding author E-mail:thombare.manisha@kbtcoe.org*

Abstract

Cyber security breaches pose significant threats to organizations and individuals, necessitating proactive measures for prediction and prevention. This paper proposes a novel approach for Cyber Hacking Breaches Prediction using Machine Learning (CHBPM), leveraging state of-the-art techniques to anticipate and mitigate potential breaches. Building upon previous research, we integrate machine learning algorithms, specifically Support Vector Machine (SVM) and Random Forest (RF), into a robust framework for real-time monitoring and analysis of website security. The system is trained on historical attack data, adaptively learning from successful and unsuccessful attempts to breach security protocols. Our methodology encompasses ensemble classification methods, including Random Forest, for efficient classification of network traffic data, thereby facilitating the identification of anomalous behaviour indicative of potential cyber threats. Through experimental analysis and evaluation owe demonstrate the efficacy of the proposed CHBPM in predicting cyber hacking breaches with improved accuracy and proactive response capabilities. This research contributes to the advancement of cyber security measures by providing a comprehensive framework for predicting and mitigating cyber threats using machine learning techniques, thereby safeguarding digital assets and enhancing resilience in an increasingly interconnected world. Real-time detection mechanisms are devised, employing the best-performing models to continuously monitor network traffic and issue alerts when suspicious activity or potential breaches are identified. The paper explores ensemble methods, such as Random Forest, to fortify detection accuracy and delves into deep learning techniques, including Neural Networks, to unearth intricate patterns within network data. Interpretability and visualization techniques aid security analysts in comprehending the models' decision-making processes. A feedback loop ensures on-going system refinement, adapting to emerging threats and incorporating real-world feedback. Ethical considerations guide responsible data handling and model deployment, upholding privacy and consent standards. By project completion, "Cyber Hacking Breaches Prediction and Detection using Machine Learning" aims to equip organizations and individuals with an adaptive, real-time defence against cyber-attacks. It aspires to render network breaches an increasingly formidable challenge, reinforcing our collective cyber security posture amidst a dynamic threat landscape.

Keywords: *Cybersecurity; Machine Learning; Cyber Hacking; Breach Prediction; Network Anomalies; Intrusion Detection.*

1. Introduction

In today's digital age, cyber security breaches have become a pervasive threat, endangering both individuals and organizations alike. The unauthorized acquisition or loss of sensitive information due to data breaches can lead to significant financial and reputational damage. These breaches often occur due to weak security measures or vulnerabilities in software systems, highlighting the critical need for effective predictive mechanisms to detect and mitigate cyber hacking breaches. Machine learning has emerged as a promising approach for enhancing cyber security measures, leveraging its ability to analyze large volumes of data and identify patterns indicative of potential breaches. By harnessing the power of machine learning algorithms, such as Support Vector Machines (SVM) and Random Forest, it becomes feasible to develop proactive systems capable of predicting and preemptively responding to cyber hacking breaches.

This paper aims to explore the application of machine learning techniques in Cyber Hacking Breaches Prediction (CHBP), focusing on the detection and identification of patterns associated with cyber-attacks. Unlike traditional approaches, which may rely on manual analysis or rule-based systems, machine learning offers the advantage of adaptability and scalability, allowing for real-time monitoring and analysis of website security. Previous research has demonstrated the effectiveness of machine learning algorithms in detecting unauthorized users through classification techniques such as logistic regression, decision tree learning, and neural networks.

However, the choice of algorithm is crucial, considering factors such as interpretability, efficiency, and handling of outliers. By maintaining extensive logs from websites and employing machine learning algorithms, we aim to develop a robust CHBP model capable of efficiently analyzing and interpreting vast amounts of data. The proposed model will undergo rigorous evaluation using diverse datasets to assess its effectiveness in predicting cyber hacking breaches with high accuracy and facilitating timely response actions. Through this research, we endeavour to contribute to the advancement of cyber security measures by providing a comprehensive framework for Cyber Hacking Breaches Prediction using Machine Learning. Ultimately, the goal is to safeguard digital assets and enhance resilience against cyber threats in today's interconnected world.

2. Related work

Title of Paper	Authors, Publication, andYear	Short Description of Paper	Limitations
Cyber ThreatDetection usingMachine Learning Techniques:A PerformanceEvaluation Perspective	KamranShaukat et al.,Published in2021	This paper assesses machine learning techniqueslike deep belief networks,decision trees, and RandomForest Classifier for cyberthreat detection, particularly in spam, intrusion and malware scenarios.	The paper concludesthat machine learningis crucial for tackling cyber threatsbut highlights limitations. There's no one-size-fits-all solution, datasets lack diversity, and customized models for security are needed.
Modeling andPredicting Cyber HackingBreaches	R. Raja Subramanian et al.,Published in2021	This paper presents a machine learning model,mainly utilizing SupportVector Machines, to detect and prevent data breachesin real-time. It focuses onlearning from historicalattack data, continuouswebsite monitoring, andclassifying access patternsfor improved cyber security.	The research shouldbe moved forwardto form an identical structure for allvulnerable situationsthat can come ahead.
Detection andIdentification ofCyber-Attacksin Cyber-Physical Systems Based onMachine Learning Methods	Zohre NasiriZarandi et al.,Published in2020	This research paper focuseson the growing vulnerability of cyber-physicalsystems to intelligent andstealthy cyber attacks andstudies different neural networks for early attack detection.	The study's limitations include a narrow focus on certain attack types and the need for broader applicability testing, especially in larger andmore complex cyberphysical systems.

3. Methodology

3.1. Data collection

The first step involves gathering a comprehensive dataset of historical cyber breaches. This data can be obtained from public repositories or through collaborations with security firms.

The dataset should include information like:

Network traffic data (e.g., volume, source, destination, protocols)

System logs (e.g., login attempts, file access, application activity)

Vulnerability scan results

Known breach indicators (e.g., malware signatures, suspicious IP addresses)

Information about the type and severity of past breaches (if available)

3.2. Data pre-processing

The collected data may need cleaning and pre-processing to ensure its quality and consistency.

This might involve:

Handling missing values

Scaling numerical features

Encoding categorical features

Feature engineering (creating new feature based on existing ones)

Model Selection and Training:

Various machine learning algorithms can be explored for breach prediction. Here are some potential options with mathematical models:

1) Logistic regression:

This is a supervised learning algorithm that models the relationship between independent variables (features) and a binary dependent variable (breach or no breach). Mathematically, it uses a sigmoid function to estimate the probability of a breach:

P(breach) = 1 / (1 + exp(-(β0 + β1x1 + β2x2 + ... +βnxn)))

Where:

- P(breach) is the probability of a breach occurring.
- β0 is the bias term.
- βi (i = 1 to n) are the coefficients for each feature (xi).

2) Random forest:

This is an ensemble learning method that combines multiple decision trees for improved prediction accuracy. Ensemble Learning: Random Forest is a popular ensemble learning method used for both classification and regression tasks in machine learning. Ensemble learning combines multiple individual models to improve predictive performance and robustness.

Decision Trees: At the core of Random Forest are decision trees. Decision trees are hierarchical structures that recursively partition the feature space into regions, making decisions based on the values of input features.

Bootstrapping: Random Forest employs a technique called bootstrapping to create multiple subsets of the original dataset. Each subset, known as a bootstrap sample, is generated by randomly sampling observations with replacement from the original dataset. This introduces diversity among the trees in the forest.

Random Feature Selection: In addition to bootstrapping, Random Forest introduces randomness by selecting a random subset of features at each node of the decision tree. This prevents individual trees from relying too heavily on any single feature, reducing overfitting and improving generalization performance.

Building Trees: With the bootstrap samples and random feature subsets, Random Forest constructs a predetermined number of decision trees. Each tree is grown independently, typically using techniques such as recursive binary splitting to partition the feature space based on feature values that minimize impurity or maximize information gain.

Voting or Averaging: For classification tasks, each decision tree in the Random Forest independently predicts the class label of a new instance. The final prediction is determined through a majority voting mechanism, where the class with the most votes across all trees is selected as the final prediction. For regression tasks, the final prediction is often the average of the predictions made by individual trees.

Hyperparameter Tuning: Random Forest includes several hyperparameters that influence its performance and behavior, such as the number of trees in the forest, the maximum depth of each tree, and the number of features considered at each split. Hyperparameter tuning techniques, such as grid search or random search, can be employed to find the optimal combination of hyperparameters for a given dataset.

Feature Importance: Random Forest provides a measure of feature importance, indicating the contribution of each feature to the predictive performance of the model. Feature importance is calculated based on metrics such as the mean decrease in impurity or the mean decrease in accuracy when a particular feature is randomly permuted.

Parallelization: Random Forest is inherently parallelizable, as each decision tree in the forest can be grown independently of the others. This makes Random Forest suitable for parallel and distributed computing environments, enabling efficient training on large-scale datasets.

Robustness: One of the key strengths of Random Forest is its robustness to noisy data and outliers. By aggregating predictions from multiple trees, Random Forest reduces the risk of overfitting and improves resilience to variance in the data.

3) Support vector machine (SVM)

This algorithm creates a hyper plane in high dimensional space that separates breach and non-breach data points.

In training phase following two functions are used in SVM according to the data.

- When the data is sequential then linear kernel function is used.
- When data is not sequential manner then Radial Basis Function (RBF) is used.

3.3. Training data

The labeled dataset is used for training the model, in which the hyperplane which separate the breaches form the linear data with largest margins.

For non linear data SVM uses regularization parameter(C) to balance the classification between maximum and minimum margins.

4. Long short-term memory (LSTM) networks

This type of recurrent neural network can be used for sequential data like network traffic logs, capturing temporal dependencies in the data.

The chosen algorithms will be trained on the pre-processed dataset. This involves splitting the data into training and testing sets. The training set is used to fit the model parameters, while the testing set is used to evaluate the model's performance on unseen data.

5. Mathematical model

Consider set D as a dataset of data points of the previous cyber breaches activities.

Let D represent the dataset consisting of N observations (data points) related to past cyber activity:

$$D = \{(Xi, yi)\}i \overset{N}{=} 1$$

Where:

$X_i \in R^d$_is a vector of features for the i^{th} observation, representing characteristics such as network traffic, user behavior, and system vulnerabilities.

$y_i \in \{0,1\}$ is the label indicating whether a breach occurred (1) or not (0) for the i^{th}observation.

Thus, the dataset D is used for training the machine learning model to predict y, the breach event.

6. Proposed system implementation details

The proposed system outlined in the text utilizes machine learning for website breach prediction. It has some promising aspects. Leveraging machine learning to analyse historical data from various e-commerce platforms offers a broad perspective on vulnerabilities and allows for continuous monitoring. However, there are limitations. Focusing solely on e-commerce data might miss other attack vectors, relying on a single, unspecified model may not capture all security complexities, and storing scraped data from other websites raises ethical and legal concerns.

Here is a new proposal that addresses these short comings:

First, we'd gather data from diverse sources beyond e-commerce platforms. This includes website logs, network traffic details, vulnerability scan results, public cyber breach datasets, and potentially third-party threat intelligence (with proper permissions).

Importantly, we wouldn't scrape data from other websites without explicit consent.

Next, the data would be cleaned and prepped for analysis. Then, we'd develop two machine learning models: a Random Forest for its ability to handle complex relationships and a Logistic Regression model to estimate breach probability. Both models would be trained and evaluated on separate datasets to determine the one with the best overall performance.

Finally, the chosen model would be deployed in a production environment for real-time monitoring. Integration with security infrastructure would enable automated responses to potential breaches, and continuous monitoring with periodic retraining would ensure the model stays effective against evolving cyber threats.

This new proposal offers several advantages. The ensemble approach with Random Forest and Logistic Regression provides a more robust prediction strategy. The diverse data sources lead to a more comprehensive analysis, and the ethical data collection avoids legal and ethical issues. By implementing this improved system, you can achieve a more accurate and reliable cyber hacking breach prediction system, ultimately enhancing your website's security posture.

Block Diagram:

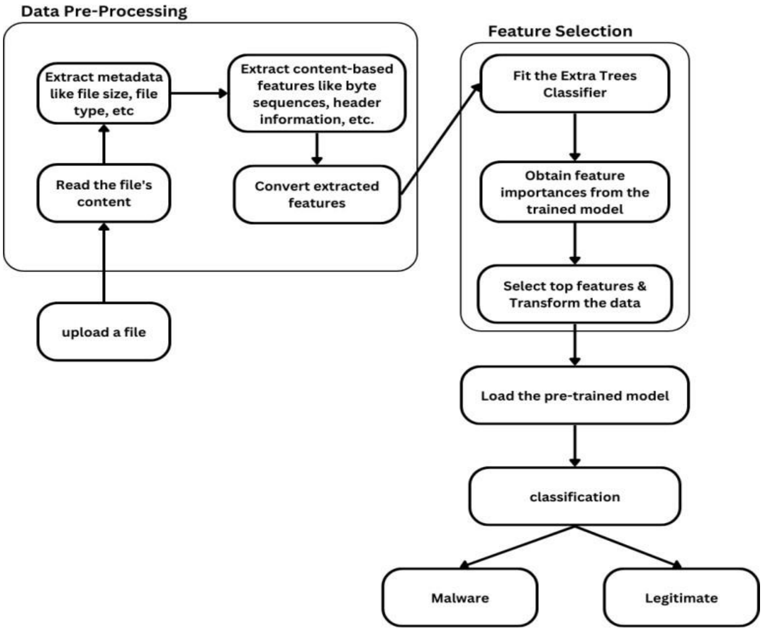

Data Pre-Processing
Extract metadata like file size, file type, etc
Extract content-based features like byte sequences, header information, etc.
Read the file's content
Convert extracted features
upload a file
Feature Selection
Fit the Extra Trees Classifier
Obtain feature importances from the trained model
Select top features & Transform the data
Load the pre-trained model
classification
Malware
Legitimate

Fig.1: Block Diagram Architecture.

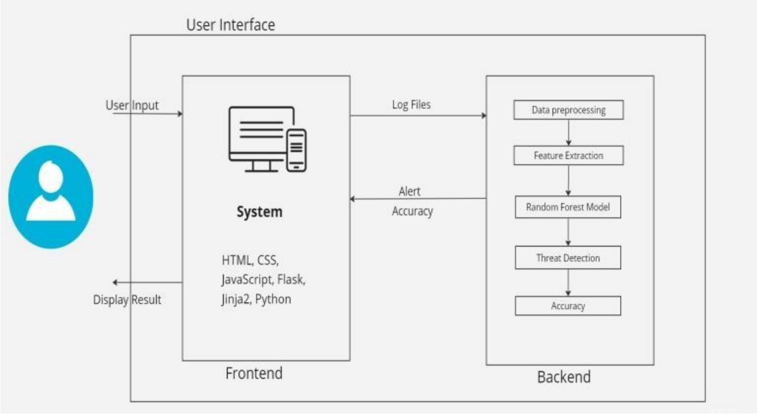
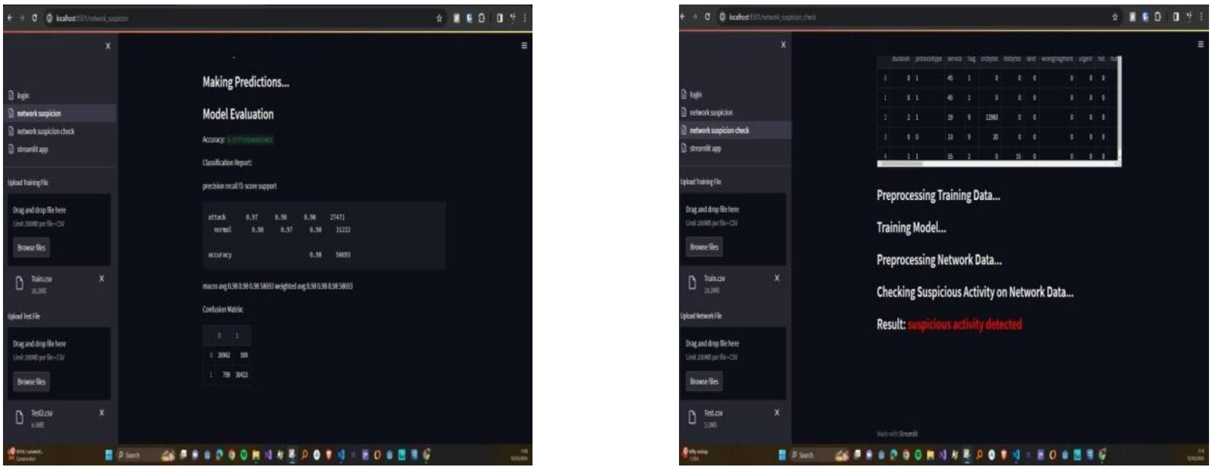


Fig.2: Architecture Diagram

7. Results

Following are some screenshots of the GUI's of the implemented system.

1) Super Admin: Super admin is one of the primary stakeholders of the system. Super Admin will be able to add users, and also be able to view profiles. Able to manipulate profiles of users have access to add, update and delete profiles
2) User Login: User is the secondary stakeholder of the system. User is able to login into the system by using his credentials. User will be able to check malwares and log files securities. In login module users have different use case to like generate reports check efficiency of different algorithms etc.



Making Predictions...

Model Evaluation

Accuracy: [illegible]

Classification Report:

precision recall f1-score support

```
attack       0.97      0.98      0.98     27471
  normal       0.98      0.97      0.98     31222

accuracy                         0.98     58693
```

macro avg 0.98 0.98 0.98 58693 weighted avg 0.98 0.98 0.98 58693

Confusion Matrix:

	0	1
0	26962	509
1	799	30423

	duration	protocoltype	service	flag	srcbytes	dstbytes	land	wrongfragment	urgent	hot
0	0	1	45	1	0	0	0	0	0	0
1	0	1	45	1	0	0	0	0	0	0
2	2	1	19	9	12983	0	0	0	0	0
3	0	0	13	9	20	0	0	0	0	0
4	1	1	55	2	0	15	0	0	0	0

Preprocessing Training Data...

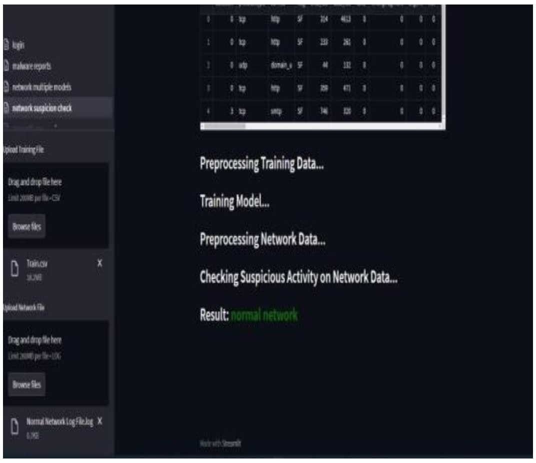
Training Model...

Preprocessing Network Data...

Checking Suspicious Activity on Network Data...

Result: suspicious activity detected

Fig. 3: Model Evaluation.



0	0	tcp	http	SF	214	4613	0	0	0	0
1	0	tcp	http	SF	233	261	0	0	0	0
2	0	udp	domain_u	SF	44	132	0	0	0	0
3	0	tcp	http	SF	299	471	0	0	0	0
4	3	tcp	smtp	SF	746	320	0	0	0	0

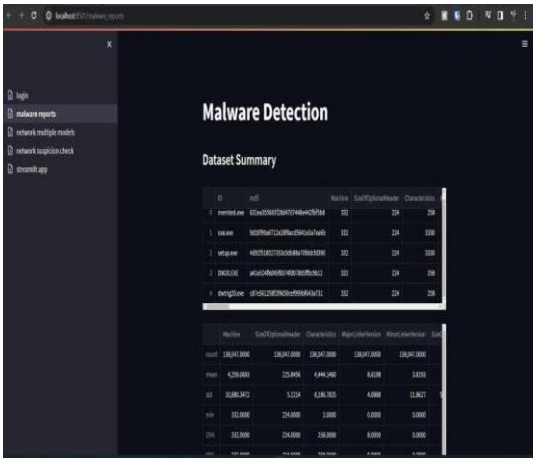
Preprocessing Training Data...

Training Model...

Preprocessing Network Data...

Checking Suspicious Activity on Network Data...

Result: normal network



Malware Detection

Dataset Summary

	ID	md5	Machine	SizeOfOptionalHeader	Characteristics
0	memtest.exe	631ea355665f28d4707448e442fbf5b8	332	224	258
1	ose.exe	9d10f99a6712e28f8acd5641e3a7ea6b	332	224	3330
2	setup.exe	4d92f518527353c0db88a70fddcfd390	332	224	3330
3	DW20.EXE	a41e524f8d45f0074fd07805ff0c9b12	332	224	258
4	dwtrig20.exe	c87e561258f2f8650cef999bf643a731	332	224	258

	Machine	SizeOfOptionalHeader	Characteristics	MajorLinkerVersion	MinorLinkerVersion
count	138,047.0000	138,047.0000	138,047.0000	138,047.0000	138,047.0000
mean	4,259.0693	225.8456	4,444.1460	8.6198	3.8193
std	10,880.3472	5.1214	8,186.7825	4.0888	11.8627
min	332.0000	224.0000	2.0000	0.0000	0.0000
25%	332.0000	224.0000	258.0000	8.0000	0.0000

Fig. 4: Predicted Result as Normal and Suspicious.

8. Evaluation metrics

Given that predicting cyber breaches involves working with imbalanced datasets, appropriate evaluation metrics must be chosen. Let:
TP (True Positives): Correct breach predictions
TN (True Negatives): Correct non-breach predictions
FP (False Positives): Incorrect breach predictions
FN (False Negatives): Missed breach predictions
Precision:
Precision = TP/TP+FP
Recall: Recall=TP/TP+FN
F1-score (harmonic mean of precision and recall):
F1 = 2 ×Precision*Recall/Precision+Recall

9. Conclusion

Machine learning algorithms have great potential of predicting and preventing the cyber flaws by considering the various parameters of the network logs. The different machine leaning algorithms can be implemented and compared with various evaluation metrics to check performance of the system to predict and detect the cyber breaches. By using this system the risk of cyber attacks can be reduced.

References

[1] Liangqiu Meng “College Student Management System Design Using Computer AidedSystem”,2015https://doi.org/10.1109/ICITBS.2015.59.

[2] Kamran Shaukat et al., “Cyber Threat Detection Using Machine Learning Techniques: A Performance Evaluation Perspective”, IEEE Publication, Published in 2021.https://doi.org/10.1109/ICCWS48432.2020.9292388.

[3] Fahima Hossain, Marzana Akter and Mohammed Nasir Uddin, “Cyber Attack Detection Model (CADM) Based on Machine Learning Approach”, IEEE Publication, Published in 2021.https://doi.org/10.1109/ICREST51555.2021.9331094.

[4] R.Raja Subramanian, et al., “Modeling and Predicting Cyber Hacking Breaches”, IEEE Publication, Published in 2021.https://doi.org/10.1109/ICICCS51141.2021.9432175.

[5] Zohre Nasiri Zarandi, et al..,“Detection and Identification of Cyber-Attacks in Cyber- Physical Systems Based on Machine Learning Methods”, IEEE Publication, Published in 2020.https://doi.org/10.1109/IKT51791.2020.9345627.

[6] Prachiti Parkar, Ansh Bilimoria, “A Survey on Cyber Security IDS using ML Methods”, IEEE Publication, Published in 2021.https://doi.org/10.1109/ICICCS51141.2021.9432210.