IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cyber Hacking Breaches Detection Using Machine Learning

Prof. Kurhe. P. V1 , Aishwarya Antre2 , Piyush Deore3 , Aniket Thakare4 , Tushar Salunkhe5
Guide, Department of Computer Engineering
Students, Department of Computer Engineering
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 causesignificant 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, butexisting systems struggle with aggressive anecdotes.

II. RELATED WORK

Many researchers have explored different methods to detect and understand cyber hacking breaches, using machine learning and data analysis techniques. Xu et al. studied 12 years of cyber hacking incidents (from 2005 to 2017) and found that the time between hacking attacks and the size of each breach cannot be explained using simple mathematical distributions. Instead, they behave like random processes that change over time. They used special models called stochastic processes to predict when the next attack might happen and how big it could be. Their analysis showed that hacking attacks are becoming more frequent, although the damage per attack is not necessarily increasing. Fernandez Maimo et al. developed a deep learning system to detect anomalies in 5G mobile networks. Since 5G introduces new communication technologies, traditional security systems might not work well anymore. Their system can analyze realtime network traffic and adjust itself automatically depending on how much traffic is coming in. This helps it run efficiently and maintain good detection accuracy, even during high traffic times. Kantarcioglu and Ferrari looked into how big data is changing the world and the new security and privacy problems that come with it. As big data is collected and used across industries like healthcare and government, it is important to protect this information. They explained that the entire big data process from storing to sharing needs strong security and privacy measures to prevent misuse, such as what happened during the Cambridge Analytical scandal.Hammouchi et al. analyzed over 9,000 public data breaches from 2005 onwards, which exposed around 11.5
billion personal records. They focused especially on hacking breaches, identifying which types of organizations are most often targeted and how hackers' focus changes over time. Their findings showed that human error is causing fewer breaches now, thanks to better awareness and training. This research helps organizations understand where to focus their security efforts.

Depuru et al. proposed a machine learning method to detect and classify malware. Since more devices are now connected to the internet, cybercriminals use malware to attack and steal data. Their system uses visual representations of malware and convolutional neural networks (CNNs) to identify and classify different types of malware effectively. They tested various models and found the one with the best accuracy for detecting these threats.

III. 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 wellsuited for this project due to its ability to handle highdimensional 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.

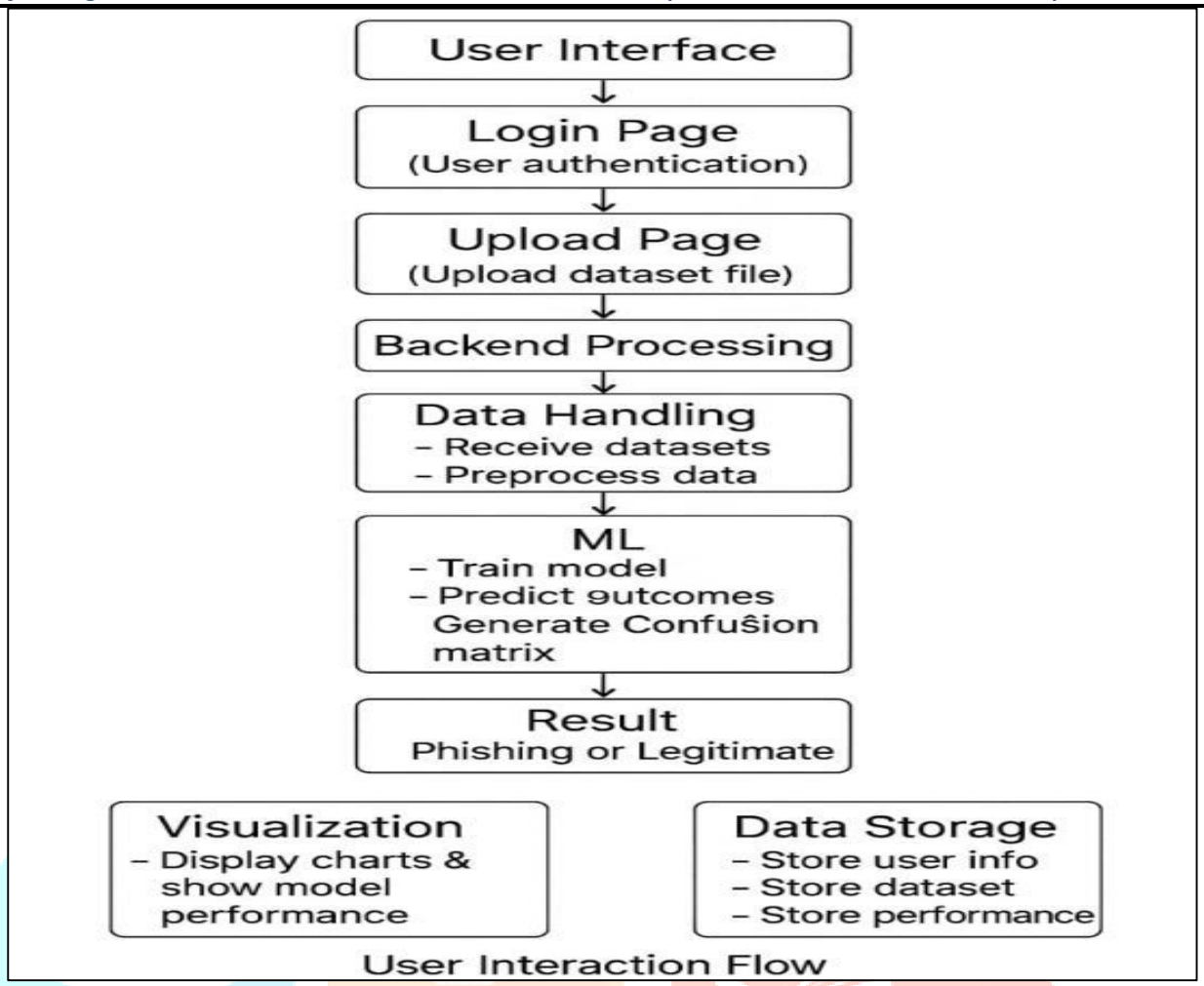


Fig. Proposed system

1.User Interface :- Entry point for user interaction.

2.Login Page :- Authenticates users before access.

3.Upload Page :- Allows users to upload dataset files.

4.Backend Processing :- Manages system operations behind the scenes.

5.Data Handling :- Receives dataset Preprocesses the data for model training.

6.ML(Machine Learning) :- Trains the model. Predicts outcomes based on input data.

7.Performance Metric :- Calculates Precision, Recall, and F1 Score. Generates a Confusion Matrix.

8.Result Phishing or Legitimate

9.Visualization:- Displays charts and graphs. Shows model performance visually.

10.Data Storage :-Stores user data,datasets and performance

11.User Interaction Flow :-Supports prediction and analysis for end users.

IV. IMPLEMENTATION

1.Data Collection:- The process begins by collecting data. In this project, data is taken from Kaggle, specifically dataset containing information about URLs labeled asphishing or legitimate. This is the raw material for training the machine learning mode.

2.Data Preparation:- This step involves cleaning and organizing the data. Unwanted records (e.g., duplicates or
missing values) are removed. The dataset is balanced to ensure equal distribution of phishing and legitimate samples using oversampling/under sampling.Feature selection is done to pick only the most useful data columns. In this case, only two features were used: URL and Status.

3.Dataset Splitting:- The prepared dataset is split into two parts: 80% for training , 20% for testing. This helps in building a reliable model and evaluating its performance separately.

4.Model Selection:- The chosen machine learning algorithm is the Random Forest Classifier. This algorithm builds many decision trees and combines their results to improve accuracy. It is effective and gives good results in classification problems like phishing detection.

5.Training & Testing:- The training data is used to teach the model to recognize patterns in URLs.The testing data is used to evaluate how well the model learned and how accurately it can classify new, unseen URLs.

6.Evaluation:- The model is evaluated using different metrics: Accuracy: How often the model is right. Precision & Recall: Measures how well the model detects phishing without too many false alarms. F1-Score: A
balance between precision and recall. The model achieved 91.6% accuracy on the test set, which is a strong result

7.Prediction:- Once trained, the model can now predict whether a new URL is phishing or legitimate. This prediction can help in detecting and preventing cyber attacks.

8.Model Saving:- The trained model is saved using Python's pickle module. This saved model can later be loaded and used in real-time applications without retraining

V .SYSTEM REQUIREMENTS

1. Hardware Requirements:

• System :Pentium i3 Processor.

• Hard Disk : 500 GB.

• Monitor: 15'' LED

• Input Devices: Keyboard, Mouse

• Ram: 6 GB

2. Software Requirements:

• Operating system : Windows 10 Pro.

• Coding Language :Python 3.10.9.

• Web Framework :Flask

VI.ADVANTAGES & DISADVANTAGES

1. Real-time Detection: ML can quickly identify threats as they occur, reducing response time .

2.Pattern Recognition: ML excels at detecting anomalies and unknown threats by learning normal behavior.

3. Scalability: It can handle large volumes of data more efficiently than manual systems.

4. Automation: Reduces the need for constant human monitoring and intervention.

VII.APPLICATIONS

1. Banking and Finance (for fraud detection): Uses advanced threat detection systems to monitor transactions and detect unusual patterns that may indicate fraudulent activity.

2.Corporate Email Security: Protects organizations from phishing, malware, and spam by scanning and filtering incoming and outgoing emails.

3.Government Cybersecurity Agencies: Employs cybersecurity applications to protect national infrastructure, sensitive data, and to monitor threats from cyber-attacks.

4. Web Hosting Providers: Utilizes security tools to safeguard hosted websites and servers against DDoS attacks, malware, and unauthorized access.

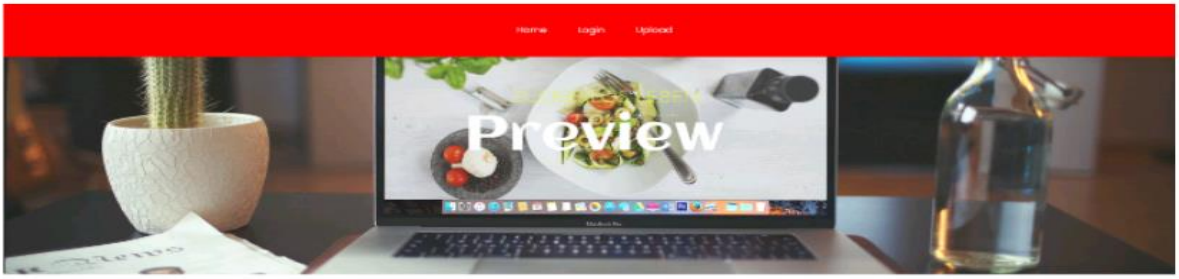
VIII .RESULT

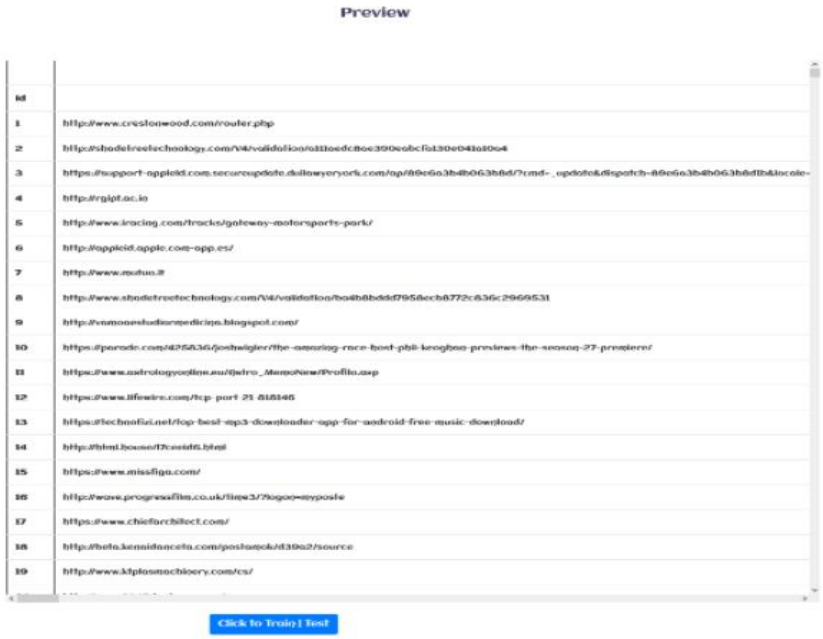
1. Home Page

2. Login Page

3. Upload CSV File

4. **Click for Train &Test**





5. **Show Prediction**

IX.CONCLUSION

The project titled "Cyber Hacking Breaches Prediction and Detection Using Machine Learning" marks a major advancement in cybersecurity Utilizing Python and the Random Forest Classifier, the system offers high accuracy and adaptability. It addresses the drawbacks of traditional rule-based and signature-based methods. These older methods struggled with evolving threats and produced many false positives. Our system uses advanced ML techniques to overcome those challenges. A balanced dataset of 5457 URLs (50% phishing, 50% legitimate) was used. This helped minimize bias and improve prediction accuracy. We extracted 87 relevant features for rich and meaningful analysis.The Random Forest Classifier was selected for its robustness and flexibility. It handles high-dimensional data and avoids overfitting effectively.The system achieved 99% training accuracy and 91% testing accuracy. This strong performance ensures better detection with fewer false positives. It boosts user trust by minimizing disruption to legitimate URLs. The system can adapt to new and evolving threats.Its ability to generalize enables detection of novel attack vectors. This makes it effective even against zero-day attacks. The success highlights the role of ML in modern cybersecurity. Python and Random Forest provide a scalable, intelligent solution. This system is both sophisticated and future-proof. It contributes significantly to cyber threat prediction and prevention. Organizations can better protect sensitive data and assets. The project is a strong step forward in cybersecurity research. It empowers defense strategies in the ever-changing digital landscape.

X.REFERENCES

[1] M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu, “Modeling 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-ofdatabreaches-not-disclosed-by-theiremployers.aspx

[8] K. Pujitha , Kattamanchi Prem Krishna , K. Amala , Annavarapu Yasaswini , Sivakumar Depuru , Kopparam Runvika, “Development of Secured Online Parking Spaces”, Journal of Pharmaceutical Negative Results, vol. 13, no. 4, pp. 1010–1013, Nov. 2022.

[9] Sivakumar Depuru , Anjana Nandam , P.A. Ramesh , M. Saktivel , K. Amala , Sivanantham. (2022). Human Emotion Recognition System Using Deep LearningTechnique. Journal of Pharmaceutical Negative Results, 13(4), 1031– 1035.https://doi.org/10.47750/pnr.2022.13.04.141 (Original work published November 4, 2022)[10] S. Depuru, P. Hari, P. Suhaas, S. R. Basha, R. Girish and P. K. Raju, "A Machine Learning based Malware Classification Framework," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023,pp.1138-1143 doi:10.1109/ICSSIT55814.2023.10060914

[11] S. Depuru, K. Vaishnavi, B. Manogna, K. J. Sri, A. Preethi and C. Priyanka, "Hybrid CNNLBP using Facial Emotion Recognition based on Deep Learning Approach," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 972-980, doi: 10.1109/ICAIS56108.2023.10073918.

[12] Ayyagari, R. (2012). An exploratory analysis of data breaches from 2005-2011: Trends and insights. Journal of
Information Privacy and Security