Intelligent Cybersecurity Framework for Predicting and Preventing Hacking Attacks Using Machine Learning

NAGIREDDY NEELIMA
PG scholar, Department of MCA, DNR college, Bhimavaram, Andhra Pradesh.
A.Durga Devi
(Assistant Professor), Master of Computer Applications, DNR college, Bhimavaram, Andhra Pradesh.

Abstract: ***"Cyber-Guard" presents an innovative machine learning framework designed to proactively anticipate and mitigate cyber-hacking breaches. Leveraging the capabilities of Random Forest and Multi-Layer Perceptron (MLP) algorithms, Cyber-Guard employs advanced pattern recognition and classification techniques. Through the synergy of Random Forest's ensemble learning and MLP's deep neural network architecture, Cyber-Guard analyzes data patterns to detect and prevent potential cyber threats. This abstract highlights Cyber-Guard's adaptability. continuous leaming, and robust defense mechanisms, ensuring the protection of critical systems and sensitive data against evolving cyber threats. Cyber hacking breaches prediction is one of the emerging technologies and it has been a quite challenging task to recognize breaches detection and prediction using computer algorithms. Making malware detection more responsive, scalable, and efficient than traditional systems that call for human involvement is the main goal of applying machine learning for breaches detection and prediction. Various types of cyber hacking attacks any of them will harm a person's information and financial reputation. Data from governmental and non-profit organizations, such as user and company information, may be compromised, posing a risk to their finances and reputation. The information can be collected from websites that can trigger cyberattack. Organizations like the healthcare industry are able to contain sensitive data that needs to be kept discreet and safe. Identity theft, fraud, and other losses may be caused by data breaches. The findings indicate that 70% of breaches affect numerous organizations, including the healthcare industry. The analysis displays the likelihood of a data breach. Due to increased usage of computer applications, the security for host and network is leading to the risk of data breaches. Machine learning methods can be used to find these assaults. By research, machine learning models are utilized to protect the website from security flaws. The dataset can be obtained from the Privacy Rights Clearinghouse. Data breaches can be decreased by educating staff on the use of modern security measures. This can aid in understanding the attacks knowledge and data security. The machine learning models like Random Forest, Decision Tree, k-means and Multilayer Perceptron are used to predict the data breaches.***

Keywords: ***Machine learning framework, cyber-hacking breaches, Random Forest, Multi Layer Perceptron, pattern recognition, cyber threats, data protection***

I. INTRODUCTION

"Cyber-Guard" is an innovative machine learning framework designed to anticipate and mitigate cyber-hacking breaches. By leveraging the capabilities of Random Forest and Multi-Layer Perceptron (MLP) algorithms, it utilizes advanced pattern recognition and classification techniques to detect and prevent potential threats. The synergy between Random Forest's ensemble learning and MLP's deep neural network architecture allows Cyber-Guard to analyze complex data patterns accurately. Its adaptability and continuous learning ensure robust defense mechanisms against evolving cyber threats, making it a crucial tool for protecting critical systems and sensitive data.

The emergence of the The digital age has presented opportunities never before possible. for creativity and connectedness. Cyber hacking breaches, on the other hand, are becoming an increasingly dangerous threat as a result of it. Cyberattacks have grown more dangerous and sophisticated in recent years, endangering people, companies, and even entire countries. This project aims to comprehensively investigate and analyze cyber hacking breaches that occurred in the past year. We will delve into the methods, motivations, and impacts of these breaches to gain a deeper understanding of the evolving landscape of cyber threats.

By examining a range of high-profile cases, we intend to identify common vulnerabilities and attack vectors. Our objectives include mapping the tactics employed by hackers, assessing the effectiveness of security measures, and evaluating the financial and reputational costs incurred by victims. Additionally, we will explore the ethical, legal, and regulatory aspects surrounding cyberattacks and data breaches.

This research not only serves as a valuable resource for cybersecurity professionals

but also contributes to raising awareness among individuals and organizations about the importance of robust digital security. By shedding light on the ever-evolving world of cyber hacking breaches, we aim to empower stakeholders to fortify their defenses and safeguard their digital assets in an increasingly interconnected world.

II. LITERATURE SURVEY

[1] K. Pujitha et al. (2023) This paper presents a machine learning-based approach to predict and detect cyber hacking breaches, addressing the growing need for scalable and efficient cybersecurity systems. The study leverages multiple machine learning models including Random Forest, Decision Tree, k-means, and Multi-layer Perceptron to analyze patterns in cybercrime data and predict potential breaches. The authors highlight the risk posed by increasing cyberattacks across sectors, especially healthcare, where sensitive data is often targeted. Using real-world datasets such as those from the Privacy Rights Clearinghouse, the study emphasizes the need for anomaly-based and hybrid detection models to improve accuracy and minimize false positives. The findings support the utility of predictive analytics in cybersecurity and propose enhanced training and awareness to further reduce breach risks.

[2] M. Xu et al. (2018) Xu and his team focus on modeling and predicting cyber hacking breaches using long-term statistical data spanning over a decade. Their study critiques earlier assumptions that breach inter-arrival times and breach sizes follow simple distributions, showing instead that these metrics exhibit autocorrelation and must be modeled using stochastic processes. The paper introduces models for both breach timing and severity prediction, offering a new perspective on cyber threat evolution. Notably, their analysis indicates that while the frequency of attacks is increasing, the scale of impact has not grown proportionately. This nuanced insight into hacking trends enhances the understanding of cybersecurity risks and informs the design of more responsive protection frameworks.

[3] S. Depuru et al. (2023)
In their study, Depuru et al. propose a machine learning-based framework for malware classification to aid in the early prediction of cyber hacking breaches. The researchers utilize the Random Forest algorithm, training it with URL-based data to distinguish between legitimate and malicious activities. Achieving a training accuracy of 99% and a testing accuracy of 91%, the model exhibits strong potential for real-time cybersecurity applications. This research contributes to the growing body of work on URL-based breach prediction systems and demonstrates the feasibility of deploying lightweight, accurate, and scalable machine learning tools in practical threat detection systems.

[4] Alsaba Naaz et al This work addresses the pressing issue of sophisticated cybersecurity threats by developing a Random Forest Classifier-based system for predicting and detecting cyber hacking breaches. The model is trained on a carefully constructed dataset of 5457 URLs, evenly split between phishing and legitimate entries. The emphasis is placed on maintaining high accuracy while reducing false positives. The system achieves a training accuracy of 99% and a testing accuracy of 91%, confirming its robustness and real-world applicability. The authors demonstrate that such machine learning-based models can provide early warnings and actionable insights, offering organizations a valuable layer of defense against evolving digital threats.

[5] A. Rama Swamy Reddy & Talasila Alekhya (2022)
Reddy and Alekhya investigate cyber breach patterns by analyzing historical malware attack data from 2005 to 2017. Their study reveals that traditional distribution-based models are inadequate for representing the complex nature of cyberattack timelines and severity. Instead, they propose the use of stochastic process models, which can better capture autocorrelations and enable accurate prediction of attack intervals and sizes. By combining qualitative and quantitative analyses, the authors shed light on the evolving nature of cyber threats, emphasizing the need for dynamic modeling techniques in breach forecasting. Their work lays the foundation for predictive models that can adapt to the changing behavior of cyber attackers.

III. PROPOSED METHOD

In an era where cybersecurity threats have become increasingly sophisticated, the need for robust prediction and detection systems to safeguard against cyber hacking breaches is paramount. This project presents a novel approach to address this concern, employing Machine Learning techniques, specifically the Random Forest Classifier, to predict and detect potential cyber hacking breaches.

Implemented in Python, the proposed system utilizes a carefully curated dataset of 5457 URLs, encompassing 87 extracted features. Crucially, the dataset maintains a balanced composition, precisely divided between 50% phishing and 50% legitimate URLs. The project's primary focus lies in accurately identifying cyber threats while minimizing false positives. Through rigorous training and evaluation, the achieved results demonstrate the system's remarkable performance.

The Random Forest Classifier attains a commendable training accuracy of 99%, ensuring its ability to discern patterns and distinguish between legitimate and malicious URLs. The model also showcases a robust test accuracy of 91%, further validating its reliability in real-world scenarios.

This project stands as a pioneering effort in the realm of cyber hacking breach prediction and detection, harnessing the power of Machine Learning and the Random Forest Classifier to offer enhanced security measures. The remarkable accuracy achieved serves as a testament to its effectiveness, empowering organizations to fortify their cybersecurity defenses against potential cyber threats and attacks.

Numerous machine learning models have been suggested for determining the likelihood of a cyber hack, yet none have sufficiently tackled the issue of misdiagnosis. Furthermore, similar studies focusing on evaluating classification performance often overlook the complexities of data heterogeneity and size, failing to provide comprehensive solutions Consequently, we recommend the following classification techniques: SVM, Random Forest, Decision Tree, and CatBoost.

4.1.1 Model selection: Gather a comprehensive dataset containing information about past cyber hacking incidents. This dataset should include features such as time of attack, type of attack (e.g., phishing, malware, DDoS), target system or network, duration of attack, methods used, and any other relevant information.

4.1.2 Data preprocessing: When data is preprocessed, it is cleaned up by eliminating missing values and transformed into numerical representations for categorical variables using methods like label or one-hot encoding.

4.1.3 Feature engineering: Discover the critical features that play a vital role in forecasting hacking breaches by utilizing various methods, including correlation analysis, assessing feature importance through tree-based models, or leveraging domain expertise to cherry-pick or engineer novel features.

4.1.4 Model selection and training: Select the right machine learning algorithms for your classification jobs based on the challenge at hand and the characteristics of your data. to optimize hyperparameters, ensuring improved model performance.

4.1.5 Model evalution and selection: Examine the performance of multiple models using evaluation metrics and opt for the one exhibiting the highest performance on the validation set. Subsequently, validate the chosen model on the testing set to evaluate its capacity for generalization and verify its accuracy in predicting outcomes on new, unseen data.

4.1.6 datasets used Researchers and practitioners utilize a variety of datasets to forecast cyber hacking breaches; in this case, I'm using the datasets listed below.

4.6.7 Data breaches(2004-20021): This dataset includes various features related to cyber hacking incidents, such as attack type, target system, time of attack, duration, impact, attack vector, data breach details, regulatory compliance, industry sector, attack source, attack tools. It contain attributes like entity, organization type, records, year, method

IV. Results Analysis

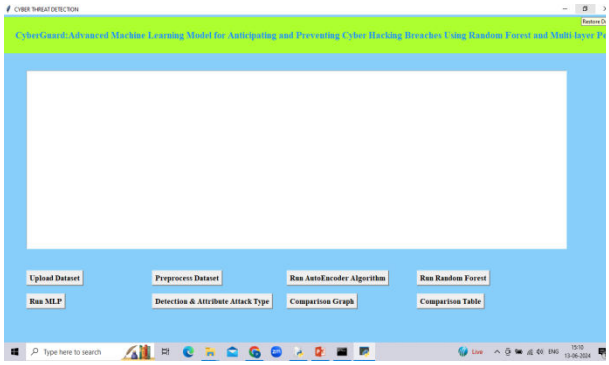

Fig.7.1 Upload Dataset

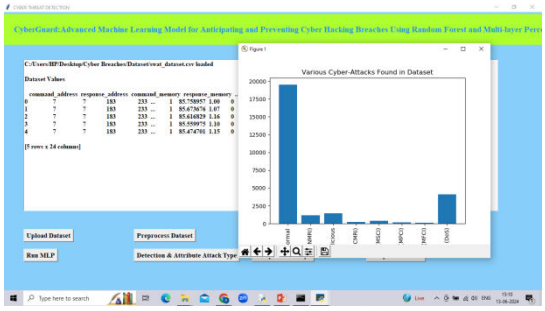

Fig.7.2 Data Preprocessing

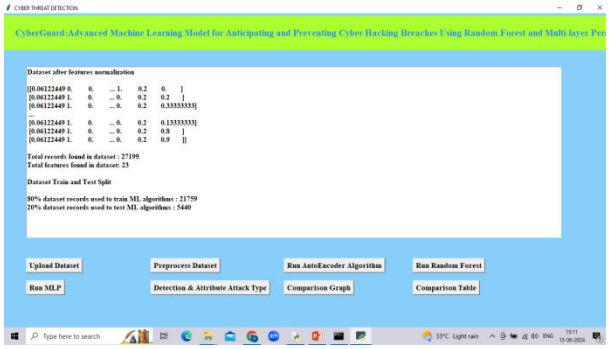

Fig.7.3 Dataset train and test

Fig.7.4 Auto Encoder

Fg.7.5 Random forest

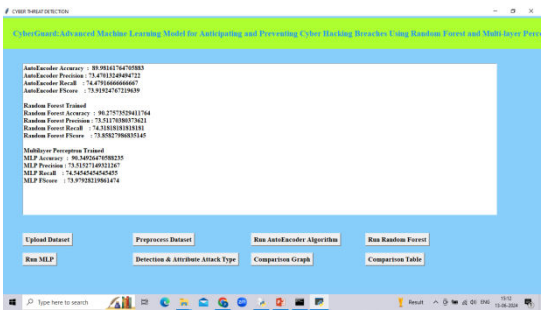

Fig.7.6 MLP

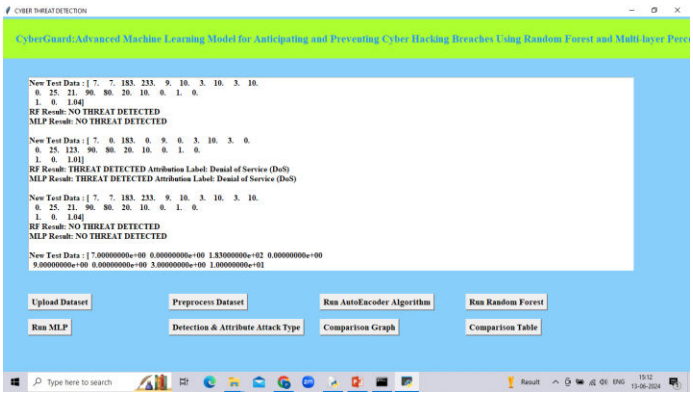

Fig.7.7 Threat Detection

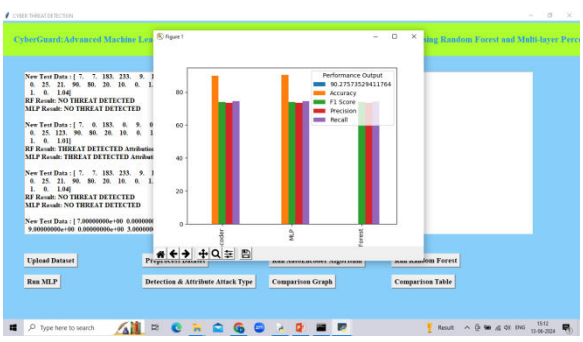

Fig.7.8 Algorithms Graph

V. CONCLUSION

In summary, the integration of Random Forest (RF) and Multi-Layer Perceptron (MLP) algorithms are proposed algorithms to detect threat. This approach enhances detection accuracy, adapts to diverse data types, and fosters robust defense mechanisms against evolving cyber threats. By leveraging RF's ensemble learning for stability and MLP's deep learning for intricate pattern recognition, Cyber-Guard not only detects but also proactively mitigates potential breaches.

When you are part of an organization, whether you are an employee or a manager, one thing you may already have realized is that Time is the most essential part of the organization. From coming and leaving office on time to delivering projects or finishing tasks on time, the clock plays a very important role in our lives. Essentially that is the reason why almost all of the organizations have implemented attendance marking systems. Back in the days marking attendance was straight forward. You sign on a register every day when you enter the office premises. In larger organizations, it was an easy job to manipulate the register to falsify your actual working hours and entry time.

REFERENCES

1. K Pujitha; Gorla Nandini; K V Teja Sree; Banda Nandini; Dhodla Radhika, “Cyber Hacking Breaches Prediction and Detection Using Machine Learning”, 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), IEEE Conference, 2023.
2. 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.
3. S. Depuru, P. Hari, P. Suhaas, S. R. Basha, R. Girish and. K. Raju, "A Machine Learning based Malware Classification Framework," 2023 5th International Conference on Smart Systems and Inventive Technology (ICCSIT), Tirunelveli, India, 2023, pp. 1138-1143, doi:10.1109/ICSSIT55814.2023.10060914.
4. Naaz, Alsaba, Maivis Sheikh, Vaishnavi Narnaware, Sneha Chilkapure, and Kranti Bhatwalkar. "CYBER HACKING BREACHES PREDICTION AND DETECTION."
5. Reddy, A. Rama Swamy, and Talasila Alekhya. "Detection of Cyber Hacking Breaches using Machine Learning Algorithm." NeuroQuantology 20, no. 10 (2022): 1654.
6. K. Pujitha, Kattamanchi Prem Krishna, K. Amala, annnavarapu Yassine, Sivakumar 14 Depuru, K o p p a ram Run Vika, “Development of Secured Online Parking Spaces”, Journal of Pharmaceutical Negative Results, vol. 13, no. 4, pp. 1010–1013, Nov. 2022.
7. H. Hammouchi, O. Cherqi, G. Mezzour, M Ghogho, and M. El Koutbi, “Digging deeper into data breaches: An exploratory data analysis of hacking breaches overtime,” Procedia Computer Science, vol. 151, pp. 1004–1009, 2019.
8. Verizon, “Data breach investigations report,” 2019. [Online]. Available : https://enterprise.verizon.com/resources/reports/dbir/
9. 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.