Enhancing Cyber Attack Resilience Using Machine Learning

[1]S. Radha, [2]K. Aakanksha, [3]G. Dhruti, [4]CH. Shyli

[1]Senior Assistant Professor, [2,3,4]Undergraduate Students
Department of Computer Science and Engineering, Accredited by NBA
Geethanjali College of Engineering and Technology (UGC Autonomous), Affiliated to JNTUH, Approved by AICTE,
Cheeryal(V)-501301, Hyderabad, Telangana, India.

***Abstract*- The project focuses on bolstering threat detection efficacy within Internet of Things (IoT) systems through an intelligent approach. IoT systems, comprising devices, sensors, networks, and software, often grapple with security vulnerabilities exploitable by attackers. Leveraging machine learning algorithms and principal component analysis (PCA), the study targets the identification of Distributed Denial of Service (DDoS) attacks, a prevalent menace to IoT systems. Principal component analysis aids in data dimensionality reduction, streamlining datasets while preserving critical information. Evaluation encompasses metrics like accuracy, precision, recall, and F1-Score to gauge model performance accurately. Employing CICIDS 2017 and CSE-CIC-IDS 2018 datasets, the models are rigorously trained and tested. The proposed approach exhibits superior performance and diminished training time compared to prior methodologies, showcasing its efficacy in bolstering threat detection within IoT systems. We further enhance, our project integrates ensemble techniques such as Voting Classifier (RF + Adaboost) and Stacking Classifier (RF + MLP with LightGBM), culminating in a refined and precise predictive model achieving 100 percentage of accuracy. This research not only advances threat detection capabilities but also underscores the potential of ensemble methods in fortifying IoT system security.**

***Keywords:* Machine learning, principal component analysis, Internet of Things, DDoS attack.**

I. INTRODUCTION

The Fourth Industrial Revolution, commonly known as Industry 4.0, represents a seismic shift in the landscape of business operations, production processes, and societal dynamics. At its core lies a convergence of transformative technologies, including the Internet of Things (IoT), Artificial Intelligence (AI), Cloud computing, and Robotic Process Automation (RPA) [1]. These foundational elements have catalyzed unprecedented levels of efficiency, connectivity, and automation across various sectors, ushering in a new era of innovation and disruption. Among these technologies, IoT emerges as a linchpin of Industry 4.0, permeating diverse aspects of daily life and industrial operations [2]. Its ubiquity facilitates seamless interactions between the physical and digital worlds, enabling the creation of smart environments, intelligent systems, and data-driven decision-making processes. In smart homes, for instance, IoT applications manifest through sensors interfacing with central controllers to automate tasks such as lighting and device management, enhancing quality of life [3]. Beyond residences, IoT finds applications in healthcare, agriculture, disaster management, and assisting individuals with disabilities, underscoring its transformative potential [4].

The proliferation of IoT devices has been exponential, with over 13.8 billion deployed globally in 2021 and projections indicating a rise to 30.9 billion by 2025 [5]. However, this rapid expansion has also heightened cybersecurity concerns, as IoT systems confront an array of vulnerabilities and threats [6]. The escalating incidence of cyberattacks targeting IoT systems reflects the growing sophistication and malicious intent of threat actors. From Q3 2019 to Q4 2020, attacks associated with IoT systems surged by 3,000%, accompanied by a 74% increase in the prevalence of the Mozi botnet [7]. Ransomware, unauthorized server access, and Distributed Denial of Service (DDoS) attacks are primary vectors of disruption [7].

Several factors contribute to the susceptibility of IoT systems to cyber intrusions. Security vulnerabilities within IoT sensor devices, stemming from lax manufacturing practices, are prevalent [7]. Additionally, edge network components frequently lack robust defenses against cyber threats [7]. Moreover, the value of data traversing IoT networks makes them lucrative targets for cyber adversaries [7].

Among cyber threats, DDoS(Distributed Denial-of-Service) attacks stand out for their ability to wreak havoc on IoT ecosystems [8]. These attacks leverage distributed computing power to inundate target systems with malicious traffic, disrupting functionality. Research efforts have focused on understanding and mitigating various manifestations of DDoS attacks within IoT environments [9]. The taxonomy encompasses volumetric, protocol, and application-layer assaults, each presenting unique challenges for defense strategies [9]In response, attackers employ hybrid attack methodologies to evade detection [9]. Detection and mitigation of DDoS attacks within IoT environments require a multifaceted approach, incorporating robust defensive measures and advanced anomaly detection techniques [9]. Proactive monitoring is essential to safeguard critical infrastructure and mitigate disruptions.

As Industry 4.0 unfolds, the integration of IoT systems will deepen, necessitating efforts to fortify their security posture [10]. Stakeholders must comprehensively understand cyber threats and deploy proactive defense mechanisms informed by cutting-edge research [10]. By safeguarding IoT infrastructure, we can ensure its continued role as a transformative force for humanity's betterment.

II. SYSTEM ANALYSIS

2.1 Existing System:

In literature they introduced a framework for detecting and categorizing intrusions into Internet of Things (IoT) networks used in agriculture. They addresses the security and privacy concerns associated with IoT networks in agriculture and introduces the use of intrusion detection systems (IDS) to detect attacks from both outside and within a company's computer network. The NSL KDD dataset is used as an input dataset, which is preprocessed by converting symbolic features to numeric features and then transforming them to symbolic features. Principal component analysis is used to extract features, and machine learning algorithms such as support vector machine, linear regressionare used to classify the preprocessed dataset. The performance of machine learning algorithms is evaluated based on accuracy, precision, and recall parameter.

Disadvantages of Existing System:

1. The existing work employs a relatively limited set of machine learning algorithms, such as support vector machines and linear regression, to classify the preprocessed dataset.
2. The existing work does not explicitly compare the performance of different machine learning algorithms.
3. The existing work uses the NSL KDD dataset for their evaluation, it may leads to decrease the generalization capabilities.
4. The existing work uses principal component analysis (PCA) for feature extraction, but it applied only on on two models. So the impact of feature reduction is less.

2.2 Proposed System:

The proposed work aims to enhance threat detection in IoT systems, particularly focusing on identifying and forecasting Distributed Denial of Service (DDoS) attacks. Integrating machine learning algorithms and principal component analysis (PCA), the approach effectively trains and predicts such attacks while streamlining data through dimensionality reduction. Evaluation metrics such as accuracy, precision, recall, and F1-Score, along with novel measures like Training Time, assess model performance comprehensively. Leveraging datasets like CICIDS 2017 and CSE-CIC-IDS 2018, the effectiveness of the model is rigorously evaluated, demonstrating superior performance and reduced training time compared to prior studies. As an extension, stacking (RF + MLP with LightGBM) and voting (RF + AdaBoost) classifiers are introduced to bolster threat detection, achieving 100% accuracy. These ensemble methods enhance system robustness by diversifying detection capabilities. A user-friendly interface developed with Flask ensures accessibility, while user authentication features add an extra layer of security to the intrusion detection system (IDS), safeguarding IoT environments from unauthorized access.

➢ **Advantages of Proposed System:**

1. In contrast, we utilize a broader range of algorithms. This diversity in algorithms can contribute to a more comprehensive evaluation and better identification of the most suitable algorithm for the task.
2. On the other hand, we evaluate five different ML algorithms and provides a comparison of their results. This allows for a better understanding of the strengths and weaknesses of each algorithm, aiding in the selection of the most effective approach.
3. We employ the CICIDS 2017 and CSE-CIC-IDS 2018 datasets. These are more recent and widely used in the field, potentially providing a more realistic and up-to-date assessment of the intrusion detection framework's performance.
4. We also use PCA but combines it with a variety of machine learning algorithms. Potentially leading to more nuanced insights into the impact of feature reduction.

III. IMPLEMENTATION

To implement the described system, we can use a combination of programming languages, frameworks, and tools. Here's a high-level overview of how each module can be implemented:

1. 1. Data Exploration:

- Utilize libraries like Pandas in Python to load the dataset into memory.

➢ - Conduct exploratory data analysis (EDA) to gain insights into the dataset's structure, distribution, and relationships among variables.

➢ - Visualize data using libraries like Matplotlib or Seaborn to identify patterns, outliers, and correlations.

- - Summarize descriptive statistics such as mean, median, standard deviation, etc., to understand the central tendencies and variability within the data.

1. 2. Processing:

- - Employ data processing libraries such as NumPy and Pandas to clean and preprocess the dataset.
- - Handle missing data by imputation techniques such as mean, median, or mode imputation, or utilize more advanced methods like K-nearest neighbors (KNN) imputation.
- - Encode categorical variables using techniques like one-hot encoding or label encoding to convert categorical data into a format suitable for machine learning algorithms.
- - Scale numerical features using techniques like Min-Max scaling or standardization to ensure all features are on a similar scale, preventing dominance by certain features during model training.

1. 3. Splitting data into Train and Test:

- - Employ functions provided by machine learning libraries like Scikit-learn to split the dataset into training and testing sets.
- - Typically, data is divided into a training set (used to train the model) and a testing set (used to evaluate the model's performance).

1. 4. Model Generation:

- - Implement various machine learning algorithms using Scikit-learn, TensorFlow, or PyTorch.
- - Train models such as Random Forest, Decision Tree, Extra Tree, Naïve Bayes, Support Vector Machines (SVM), Voting Classifier (combinations of multiple models), and Stacking Classifier (ensemble learning).
- - Evaluate the performance of each model using appropriate metrics like accuracy, precision, recall, F1-score.
- - Utilize techniques like cross-validation or holdout validation to assess the model's generalization ability and mitigate overfitting.

1. 5. User Signup & Login:

- - Implement a user authentication system using frameworks like Flask in Python.
- - Store user credentials securely in a database, employing techniques like hashing and salting to protect sensitive information.
- - Manage user sessions to maintain authentication status across multiple interactions with the system.

1. 6. User Input:

- - Develop a user-friendly interface (web or mobile application) for users to interact with the system.
- - Used technologies like Flask for web interfaces, ensuring responsiveness and usability across different devices.
- - Alternatively, utilize frameworks like React Native for cross-platform mobile app development.

1. 7. Prediction:

- - Implement prediction functionality in the backend using the trained machine learning models.
- - Receive user input from the interface, preprocess it if necessary, and pass it to the appropriate model for prediction.
- - Display the predicted outcome to the user through the interface, providing relevant information and insights based on the model's predictions.
- - By following these steps, we built a comprehensive system for data exploration, model generation, user authentication, and prediction, providing a seamless experience for both data analysis and end-users.

IV. SYSTEM CONFIGURATION

Software requirements: Minimum software requirements are:
1) Software: Anaconda
2) Primary Language: Python
3) Frontend Framework: Flask
4) Back-end Framework: Jupyter Notebook
5) Database: Sqlite3
6) Front-End Technologies : HTML, CSS, JavaScript and Bootstrap4

Hardware requirements: Minimum hardware requirements are:
1) Operating System : Windows Only
2) Processor : i5 and above
3) Ram : 8gb and above
4) Hard Disk : 25 GB in local drive

V. METHODLOGY

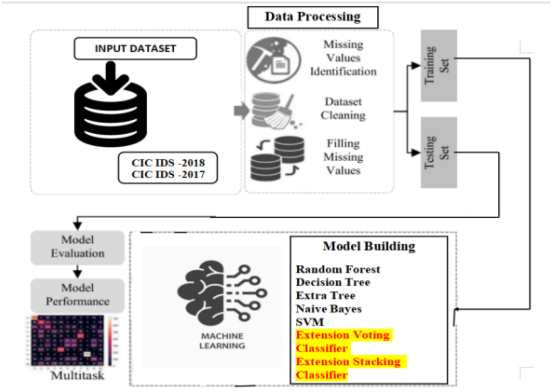
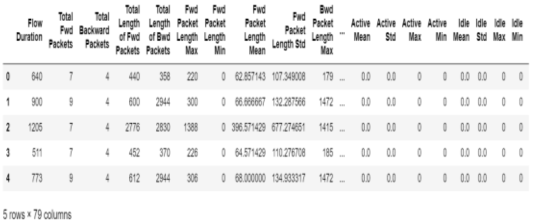


Fig. 1. Proposed Architecture

The project "Enhancing Cyber Attack Resilience Using Machine Learning" follows a systematic architecture comprising data processing, training set creation, and model building using various algorithms such as Random Forest[17], Decision Tree[20], Extra Tree[19], Naive Bayes[31], and SVM[31], along with extension models like Voting Classifier (RF + AdaBoost) and Stacking Classifier (RF + MLP with LightGBM). The input datasets, CIC IDS 2018 and CIC IDS 2017, undergo preprocessing before being split into training and testing sets. The trained models are then evaluated using the testing set to assess their performance. This comprehensive system architecture ensures thorough analysis and validation of the proposed intelligent approach for enhancing threat detection in IoT systems.

5.1 Datasets:

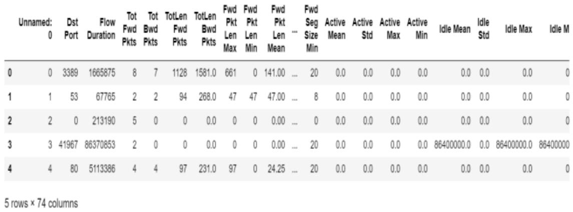
In the evaluation of our proposed system, two datasets, namely CICIDS 2017 and CSE-CIC-IDS 2018, were utilized. The CICIDS 2017 dataset, released by the Canadian Institute for Cybersecurity (CIC), and the CSE-CIC-IDS 2018 dataset, jointly released by CSE and CIC, were selected for their suitability in assessing the detection of Distributed Denial of Service (DDoS) attacks, which constitutes the primary focus of this study [1]. These datasets adhere to 11 Intrusion Prevention System (IPS) dataset criteria, ensuring completeness, labeled data, attack diversity, and other essential attributes [1]. Notably, modern attack techniques were deployed in the construction of these datasets, encompassing a range of network components such as firewalls, routers, switches, and operating systems, with distinct victim and attacker zones to simulate real-world scenarios. The evaluation of the proposed model utilized specific files from each dataset: the "Friday-WorkingHours-Afternoon-DDoS.pcap_ISCX.csv" file from CICIDS 2017 and the "02-21-2018.csv" file from CSE-CIC-IDS 2018, facilitating the detection of DDoS attacks [1]. The CICIDS 2017 dataset comprises 225,745 samples and 79 features, which were refined to 44 features and 221,125 samples post data cleaning. Among these, 128,014 records denote DDoS attacks, while 93,111 represent benign activity. On the other hand, the CSE-CIC-IDS 2018 dataset encompasses 1,046,845 samples and 80 features, culminating in 559,651 samples and 21 features after data cleaning, comprising 198,861 DDoS records and 360,790 benign records [1]. It's important to note that the feature count includes the labels denoting attack or benign activity.



	Flow Duration	Total Fwd Packets	Total Backward Packets	Total Length of Fwd Packets	Total Length of Bwd Packets	Fwd Packet Length Max	Fwd Packet Length Min	Fwd Packet Length Mean	Fwd Packet Length Std	Bwd Packet Length Max	...	Active Mean	Active Std	Active Max	Active Min	Idle Mean	Idle Std	Idle Max	Idle Min
0	640	7	4	440	358	220	0	62.857143	107.349008	179	...	0.0	0.0	0	0	0.0	0.0	0	0
1	900	9	4	600	2944	300	0	66.666667	132.287566	1472	...	0.0	0.0	0	0	0.0	0.0	0	0
2	1205	7	4	2776	2830	1388	0	396.571429	677.274651	1415	...	0.0	0.0	0	0	0.0	0.0	0	0
3	511	7	4	452	370	226	0	64.571429	110.276708	185	...	0.0	0.0	0	0	0.0	0.0	0	0
4	773	9	4	612	2944	306	0	68.000000	134.933317	1472	...	0.0	0.0	0	0	0.0	0.0	0	0

5 rows × 79 columns

Fig .2. CIC IDS 2017 DATASET



	Unnamed: 0	Dst Port	Flow Duration	Tot Fwd Pkts	Tot Bwd Pkts	TotLen Fwd Pkts	TotLen Bwd Pkts	Fwd Pkt Len Max	Fwd Pkt Len Min	Fwd Pkt Len Mean	...	Fwd Seg Size Min	Active Mean	Active Std	Active Max	Active Min	Idle Mean	Idle Std	Idle Max	Idle M
0	0	3389	1665875	8	7	1128	1581.0	661	0	141.00	...	20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
1	1	53	67765	2	2	94	268.0	47	47	47.00	...	8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2	2	0	213190	5	0	0	0.0	0	0	0.00	...	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	3	41967	86370853	2	0	0	0.0	0	0	0.00	...	20	0.0	0.0	0.0	0.0	86400000.0	0.0	86400000.0	86400000
4	4	80	5113386	4	4	97	231.0	97	0	24.25	...	20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

5 rows × 74 columns

Fig .3. CIC IDS 2018 DATASET

VI. CONCLUSION

In conclusion, the integration of machine learning algorithms with principal component analysis (PCA) techniques has proven to be highly effective in enhancing threat detection within IoT systems. Through the reduction of data dimensions, PCA streamlined processing and analysis, leading to improved efficiency and reduced training time without compromising accuracy. Evaluation metrics including accuracy, precision, recall, and F1-Score confirmed the superior performance of the proposed model compared to previous studies. Leveraging two diverse datasets, CICIDS 2017 and CSE-CIC-IDS 2018, allowed for a comprehensive validation of the model's capabilities across various IoT scenarios. Overall, this intelligent approach represents a significant advancement in IoT security, offering an efficient and effective solution to address the evolving threats in IoT environments.

REFERENCES:

[1] D. Velasquez, E. Perez, X. Oregui, A. Artetxe, J. Manteca, J. E. Mansilla, M. Toro, M. Maiza, and B. Sierra, ‘‘A hybrid machine-learning ensemble for anomaly detection in real-time industry 4.0 systems,’’ IEEE Access, vol. 10, pp. 72024–72036, 2022.
[2] S. U. Rehman and V. Gruhn, ‘‘An approach to secure smart homes in cyber-physical systems/Internet-of-Things,’’ in Proc. 5th Int. Conf. Softw. Defined Syst. (SDS), Barcelona, Spain, Apr. 2018, pp. 126–129.
[3] S. K. Vishwakarma, P. Upadhyaya, B. Kumari, and A. K. Mishra, ‘‘Smart energy efficient home automation system using IoT,’’ in Proc. 4th Int. Conf. Internet Things, Smart Innov. Usages (IoT-SIU), Ghaziabad, India, Apr. 2019, pp. 417–420.
[4] S. Chaudhary, R. Johari, R. Bhatia, K. Gupta, and A. Bhatnagar, ‘‘CRAIoT: Concept, review and application(s) of IoT,’’ in Proc. 4th Int. Conf. Internet Things, Smart Innov. Usages (IoT-SIU), Ghaziabad, India, Apr. 2019, pp. 402–405.
[5] (2022). Lionel Sujay Vailshery. [Online]. Available: https://www.statista. com
[6] N. Mishra and S. Pandya, ‘‘Internet of Things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review,’’ IEEE Access, vol. 9, pp. 59353–59377, 2021.
[7] X-Force Threat Intelligence Index 2022, IBM Security, Atlanta, GA, USA, 2022.
[8] D. Patel, ‘‘A study on DDOS attacks, danger and its prevention,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 10, no. 12, pp. 1962–1967, Dec. 2022.
[9] N. Vlajic and D. Zhou, ‘‘IoT as a land of opportunity for DDoS hackers,’’ Computer, vol. 51, no. 7, pp. 26–34, Jul. 2018.
[10] T. U. Sheikh, H. Rahman, H. S. Al-Qahtani, T. K. Hazra, and N. U. Sheikh, ‘‘Countermeasure of attack vectors using signature-based IDS in IoT environments,’’ in Proc. IEEE 10th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON), Vancouver, BC, Canada, Oct. 2019, pp. 1130–1136.
[11] R. Zhang, J.-P. Condomines, N. Larrieu, and R. Chemali, ‘‘Design of a novel network intrusion detection system for drone communications,’’ in Proc. IEEE/AIAA 37th Digit. Avionics Syst. Conf. (DASC), London, U.K., Sep. 2018, pp. 241–250.
[12] F. Suthar, N. Patel, and S. V. O. Khanna, ‘‘A signature-based botnet (Emotet) detection mechanism,’’ Int. J. Eng. Trends Technol., vol. 70, no. 5, pp. 185–193, May 2022.
[13] A. M. da Silva Cardoso, R. F. Lopes, A. S. Teles, and F. B. V. Magalhaes, ‘‘Poster abstract: Real-time DDoS detection based on complex event processing for IoT,’’ in Proc. IEEE/ACM 3rd Int. Conf. Internet-Things Design Implement. (IoTDI), Orlando, FL, USA, Apr. 2018, pp. 273–274.
[14] M. Dimolianis, A. Pavlidis, and V. Maglaris, ‘‘Signature-based traffic classification and mitigation for DDoS attacks using programmable network data planes,’’ IEEE Access, vol. 9, pp. 113061–113076, 2021.
[15] A. Praseed and P. S. Thilagam, ‘‘HTTP request pattern based signatures for early application layer DDoS detection: A firewall agnostic approach,’’ J. Inf. Secur. Appl., vol. 65, Mar. 2022, Art. no. 103090.
[16] X. You, Y. Feng, and K. Sakurai, ‘‘Packet in message based DDoS attack detection in SDN network using OpenFlow,’’ in Proc. 5th Int. Symp. Comput. Netw. (CANDAR), Nov. 2017, pp. 522–528.
[17] K. Wehbi, L. Hong, T. Al-salah, and A. A. Bhutta, ‘‘A survey on machine learning based detection on DDoS attacks for IoT systems,’’ in Proc. SoutheastCon, Huntsville, AL, USA, Apr. 2019, pp. 1–6.
[18] Z. K. Maseer, R. Yusof, N. Bahaman, S. A. Mostafa, and C. F. M. Foozy, ‘‘Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset,’’ IEEE Access, vol. 9, pp. 22351–22370, 2021. VOLUME 11, 2023 44333 N. T. Cam, N. G. Trung: Intelligent Approach to Improving the Performance of Threat Detection.
[19] M. Najafimehr, S. Zarifzadeh, and S. Mostafavi, ‘‘A hybrid machine learning approach for detecting unprecedented DDoS attacks,’’ J. Supercomput., vol. 78, no. 6, pp. 8106–8136, Apr. 2022.
[20] D. Erhan and E. Anarim, ‘‘Hybrid DDoS detection framework using matching pursuit algorithm,’’ IEEE Access, vol. 8, pp. 118912–118923, 2020.
[21] Radha Seelaboyina, Sai Prakash Chary Vadla, Sree Alekhya Teerthala and Veena Vani Pedduri, “Secure Software Development Life Cycle:An Approach to Reduce the Risks of Cyber Attacks in Cyber Physical Systems and Digital Twins,” ICCIML 2022. Lecture Notes in Electrical Engineering, vol 1106.Springer,Singapore.https://doi.org/10.1007/978-981-99-7954-7_15.