**Predicting Cyber-Attacks and Identifying Perpetrators Using Machine Learning Techniques**

1**Senthil G. A***Associate Professor,  
Dep. of Information Technology,*

*Agni College of Technology*,

Chennai, India.

[senthilga@gmail.com](mailto:senthilga@gmail.com)

2**Geerthik S***Associate Professor,  
Dep. of Information Technology,*

*Agni College of Technology*,

Chennai, India.

[geerthiks@gmail.com](mailto:geerthiks@gmail.com)

3**Tharagairani G***Assistant Professor,  
Dep. of Information Technology,*

*Agni College of Technology*,

Chennai, India.

[tharagairani.it@act.edu.in](mailto:tharagairani.it@act.edu.in)

4**Sachin L***,*

*Student,  
Dep. of Information Technology,*

*Agni College of Technology*,

Chennai, India.

[sachin143258@gmail.com](mailto:sachin143258@gmail.com%0d)

5**Dinesh Kumar S***Student,  
Dep. of Information Technology,*

*Agni College of Technology*,

Chennai, India.

[dineshbtech002@gmail.com](mailto:dineshbtech002@gmail.com)

*Abstract*—The generation of technology has grown significantly in current years, evolve. Instead, as technology advances, the number of crimes also rises. It will result in financial difficulties for the nation and its citizens. Cyberattacks account for the majority of reported attacks. It will target the privacy data of the people and the nation. As a result, it is challenging to determine what kind of attack would be employed, and the length of the inquiry would also rise. Numerous systems have been presented that use different machine-learning techniques to analyze and forecast threats. Consequently, we presented the machine learning and regression approach for identifying the assault. It determines the type of cyberattack that will be launched and how to foresee it. According to our research, the probability of a cyberattack declines as potential victims' income and educational levels rise. We believe that our suggested model will be helpful to law enforcement cybercrime units in identifying and stopping cyberattacks, increasing their efficacy in thwarting these threats. In conclusion, this study emphasizes how important machine learning is to improving the ability to detect cyberattacks and identify their perpetrators. By employing data-driven insights and advanced analytics, companies can enhance their capacity to anticipate, identify, and mitigate cyber threats, thereby fortifying their defences in a constantly changing digital landscape. Going forward, sustaining a proactive posture against cyber adversaries and safeguarding the security of our global community that is interconnected depends critically on continuous research and machine learning developments.

***Keywords-- Artificial intelligence, data analysis, machine learning, cyberattack crimes, security and privacy, and criminal prediction.***

# **Introduction**

In the consistently developing scene of advanced innovation, the danger of digital assaults poses a potential threat, presenting huge difficulties to people, organizations, and states around the world. These assaults, going from information breaks to malware diseases, can bring about serious results, including monetary misfortune, reputational harm, and, surprisingly, public safety dangers. With digital hoodlums turning out to be progressively modern in their strategies, there is a squeezing need for cutting-edge devices and procedures to anticipate and forestall such assaults. AI, a subset of man-made consciousness, has arisen as an amazing asset in the battle against digital wrongdoing. Overwhelmingly of information, AI calculations can recognize examples, oddities, and patterns that might demonstrate potential digital dangers. Researchers and practitioners have been focusing on using machine learning to predict cyberattacks and identify their perpetrators, allowing for proactive risk mitigation and cybersecurity enhancement. It, first and foremost, requires the assortment and preprocessing of different information sources, including network traffic logs, framework occasion logs, and authentic assault information. Highlights designing assumes a critical part in this stage as it describes various parts of digital assaults and their culprits. Predictive models can be created using a variety of machine-learning algorithms once the data has been prepared. These algorithms range from more conventional approaches like logistic regression and decision trees to more sophisticated ones like ensemble methods and neural networks. The particular characteristics of the data and the desired predictive performance determine which algorithm is used. In addition, the predictive models' accuracy, robustness, and generalizability can only be assessed through model evaluation and validation. Digital assaults are generally intriguing contrasted with genuine client conduct, making it hard for AI model to gain from imbalanced datasets. One more basic part of anticipating digital assaults is the attribution of assaults to their culprits. Identifying the individuals, groups, or organizations responsible for particular cyber incidents is a frequently difficult and complex process known as cyber attribution. AI can support this interaction by dissecting different advanced scientific antiques, including IP addresses, malware marks, and standards of conduct, to induce the probable beginning and thought processes of the assailants. Additionally, in order to improve the efficiency of cyber-attack prediction and attribution efforts, cybersecurity professionals, judicial bodies, and privately owned companies must work together and disseminate insights. By pooling assets, ability, and information, partners can work on the exactness and practicality of prescient models, accordingly fortifying aggregate protections against digital dangers. All in all, AI offers promising open doors for anticipating digital assaults and recognizing their culprits. Cybersecurity professionals can better safeguard digital assets and infrastructure by utilizing cutting-edge algorithms and data-driven strategies. However, in order to guarantee a safer and more secure cyberspace for everyone, ongoing research and innovation are required to address the inherent difficulties and complexities of cyber-attack prediction.

# **RELATED WORK**

Zahid Anwar, et al. (2019) [1] Amid rising cyber data breaches, manual log analysis proves error-prone and slow. An innovative machine learning architecture is put forth that associate assault objectives with opponent tactics, techniques, and procedures (TTPs). It builds a semantic network that connects threat vectors, detection mechanisms, and gateways. Trained on a TTP dataset, it achieves 92% accuracy, swiftly detecting incidents in 0.15 seconds on average. This framework offers automated threat analysis, enhancing cybersecurity against diverse and sophisticated threats.

Amos, Brandon et al. (2013) [2] Smartphone systems are a growing source of worry due to their broad usage and contextually sensitive nature. Using a dataset of hundreds of actual applications, this research evaluates several classifiers already in use. Our STREAM framework, designed to facilitate quick, extensive validation of mobile malware machine learning classifiers, is also shown.

Q. K. A. Mirza, et al. (2018) [3] Enterprises and users heavily rely on antivirus software, yet its methodologies often fall short in detecting malicious activities and consumes excessive host machine resources. This paper proposes a machine learning approach using a feature-rich dataset of harmful and benign files that was extracted using a unique method. utilizing decision trees, support vector machines, and boosting on decision tree, we achieve optimal detection rates. Additionally, a scalable, cloud-based architecture on Amazon Web Services is introduced to support the detection process. Testing across various scenarios yields high-performance results with minimal host machine energy consumption.

Martin Husak, et al. (2018) [4] This survey explores prediction and forecasting methods in cybersecurity across forecasting network security situations, identifying intentions, projecting attacks, and predicting intrusions are the four primary objectives. Both discrete models and continuous models are examined and compared. Additionally, machine learning and data mining approaches are discussed for their potential in the dynamic cybersecurity landscape. Practical usability and evaluation challenges of these methods are also addressed.

# **METHODOLOGY**

## **A. Data Gathering:** In cyber-attack refers to the process of collecting information relevant to launching or defending against cyber-attacks. This includes gathering intelligence on potential targets, vulnerabilities, and attacker tactics. Methods range from passive reconnaissance, such as monitoring publicly available information to active techniques like scanning networks for weaknesses. Data gathering is essential for understanding the threat landscape, identifying potential risks, and developing effective security measures to mitigate cyber threats.

**B. Data Preprocessing:** It involves several steps to prepare raw data for analysis. This includes cleaning data to remove noise and inconsistencies, such as irrelevant attributes for analysis, reducing dimensionality. Additionality, data normalization standardizes variables to a common scale, ensuring fair comparison. Finally, data encoding coverts categorical variable into numerical representations for machine learning algorithms. Overall, preprocessing enhances the quality and effectiveness of cyber-attack analysis.

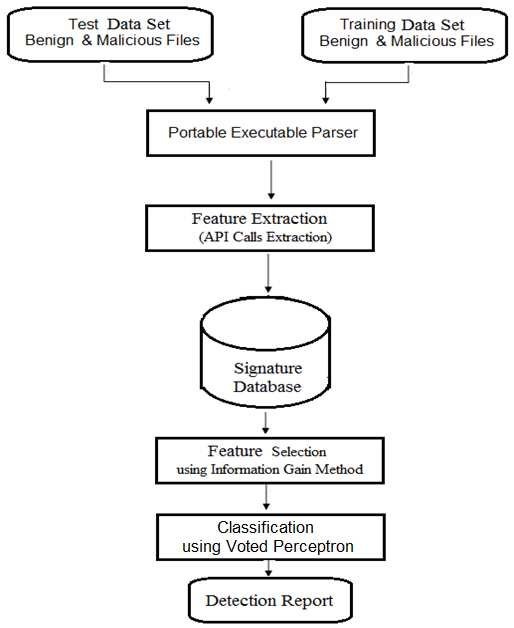
**C. Training and Testing:** Training involves using labeled data to teach machine learning models to recognize patterns indicative of cyber-attacks. Testing evaluates the trained model’s performance on new, unseen data to assess its accuracy in identifying attacks. The training phase aims to optimize the model’s ability to generalize from the training data to detect attacks effectively in real-worlds scenarios. Testing validates the model’s efficacy ensuring it can reliably identify and differentiate between malicious and benign activity without overfitting or underfitting.

**D. Predictive Modelling:** Predictive modelling in cyber-attack content involves using various techniques to anticipate and forecast potential cyber threats. It encompasses analyzing historical attack data, identifying patterns, and developing models to predict future attacks. This process enables organization to proactively strengthen their cybersecurity measures, detect vulnerabilities, and mitigate risks before they materialize. By leveraging predictive modeling, security teams can enhance their ability to respond effectively to evolving cyber threats, thereby safeguarding sensitive data and systems from potential breaches and intrusions.

# **Proposed Work**

**Machine Learning:** Within the field of artificial intelligence, machine learning focuses on developing algorithms that allow computers to learn from data and make judgments or predictions without explicit programming. It entails using a dataset to train a model to identify patterns and relationships so that it may utilize that knowledge to forecast or decide on incoming data in an informed manner. Supervised learning, in which the model learns from labeled data; unsupervised learning, in which the model looks for patterns in unlabeled data; and reinforcement learning, in which the model learns by feedback from mistakes made to iterative trials, are common approaches in machine learning. Numerous industries, including marketing, banking, and healthcare, use machine learning.

# **System Architecture:**



**Fig. 1 System Architecture of cyber-attack prediction**

**Implementation:**

In our project we gather datasets that would go to the training. In training phase, we give set of data that relevant to our data set and that would be do it again and again.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S. No | Crime | Harm | Attack | Attack Method | Perpetrator |
| 0 | Misuse of Banking Cards | Fraud | Hacking of card number | Fake Shopping site | Unknown |
| 1 | Capture Data | Without Experience Use of site | Hacking Social Media account | Social  Engineer | Uknown |
| 2 | Information Theft | Fraud | Hacking Social Media account | Receive Data on social media | known |
| 3 | Information Theft | Moral Harm | Hacking Social Media account | Phishing Attack | Unknown |

**Fig. 2 Data Gathering**

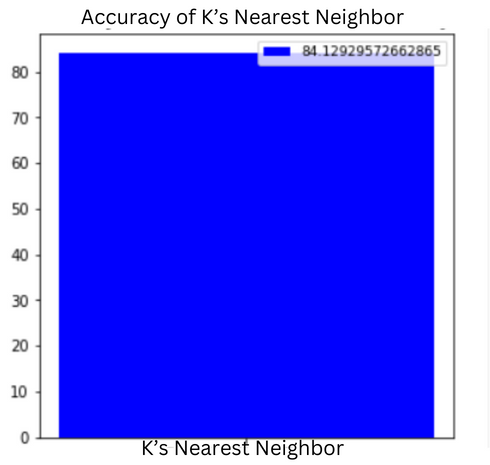
Another set will go the testing phase or data preprocessing, collected data is correct or not, that entire process would do until we get desired result.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Attack Method | Card Copying/ Generating Devices | Creating a Fake Shopping Site | Hacking Tools or Malware | Phishing Attack | Receiving Public Data on Social Media | Social Engineering |
| Copying Bank / ATM Card | 0 | 0 | 0 | 0 | 224 | 0 |
| Hacking Social Media Accounts | 120 | 222 | 0 | 0 | 0 | 608 |
| Obtaining Electronic Bank Accounts | 0 | 0 | 108 | 0 | 0 | 0 |
| Obtaining and Using Data in Digital Environment | 14 | 58 | 0 | 0 | 0 | 356 |

**Fig. 3 Data Preprocessing**

**K’s Nearest Neighbor Algorithm:**

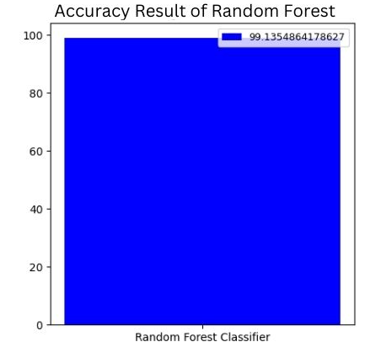
For accuracy finding we used an “K’s Nearest Neighbor” algorithm that would check our data and find the data which is nearest to our problem it will give 84% of accuracy for the nearest value of the given data.



**Fig. 4 K’s Nearest neighbor Algorithm**

**Random Forest Algorithm:**

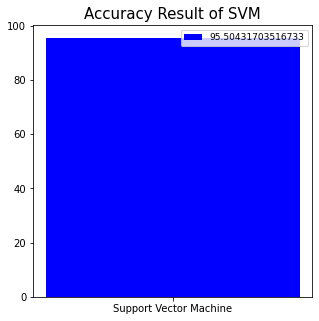
Another algorithm we used is an “Random Forest” algorithm, it is a classification and regression algorithm this method used to classify our program and go the training set our given set data would be get trained and test would be done then it goes to the regression class that our given set of data would be separated and then desired result would be given, but sometimes it can’t do the regression so it gives the accuracy of 99%.



**Fig. 5 Random Forest Algorithm**

**Support Vector Machine Algorithm:**

Third algorithm we used “Support Vector Machine” in this algorithm we use training to the AI that will helps to find the desired data, in this training set data we already trained some that data that will helps the AI we enter new data that is related our previous data it gives the desired result it will give the accuracy of 95% but sometimes it gets confused when new data get not clear or new data that time it gets confused.



**Fig. 6 Support Vector Machine Algorithm**

**Gradient Boost Algorithm:**

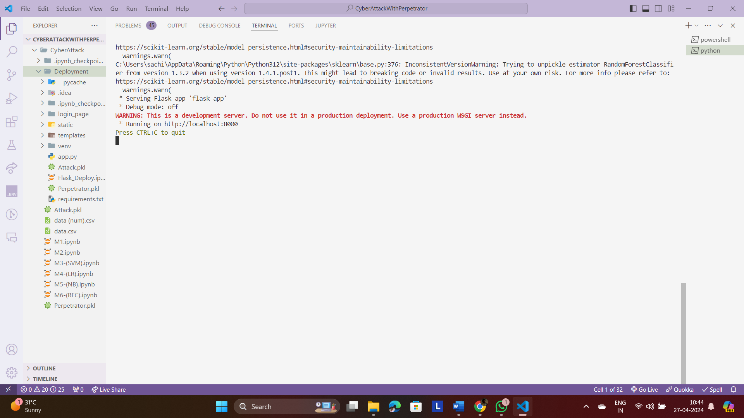
Final algorithm we used “Gradient Boost” algorithm in this algorithm it goes to the trained model and then calculated the errors in trained data and then it predicted the which values would be set for the result and then it repeat the training itself and calculated the errors it helps to find the accurate value and it would be preferable now a days, because it give the best result and it predicted the dataset it would be learned from himself of the give desired set of trained datasets and then it would be corrected.

# **Result And Conclusion**

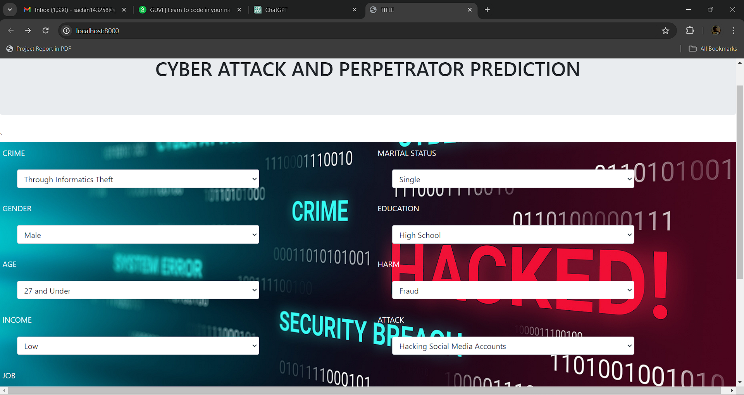
This paper provides numerous surveys on cyber-attack analysis and methodologies for prediction. This problem is set to find the what kind of attack would be used and analyzed. And this method would be compared and analyzed by the prediction methods. We can separate in three findings method.

Firstly, we most check the cybersecurity prediction methods employ a framework to analyze and anticipate potential attacks or security incidents in the future.

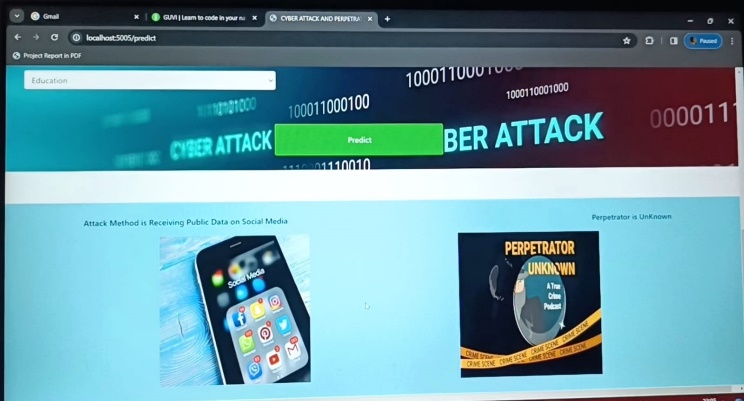
Secondly, we have introduced the new ways for the core basic of machine learning and data mining that would have the likelihood to find which kind of attack would be used and that would help to study about the crimes happen on over the network. It helps to taking the quick response for the attack on the network and how to take action for that. In conclusion, predicting cyber-attacks remains a captivating subject of research that has garnered significant attention from numerous scholars on multiple occasions.



**Fig. 7 URL Execution**



**Fig. 8 Cyber Attack Input Page**



**Fig. 9 Cyber Attack Output Page**

# **future work**

For future endeavors’, the prediction of wrongdoing, criminal behavior, victim profiling, and cyber-attacks can be facilitated through the application of deep learning algorithms, with the outcomes subject to comparison. Collaborations with authorized units possessing crime databases could further enrich the analysis by incorporating cybercrime data from diverse regions. This comparative analysis across different areas can provide valuable insights.

# **References**

1. Zahid Anwar a, b, Asad Waqar Malik a, Sharifullah Khan a, umara Noor a c, Shahzad Saleem a, “A Machine Learning Framework for Investigating Data Breaches Based on Semantic Analysis of Adversary’s Attack Patterns in Threat Intelligence Repositories”,2019 .
2. Brandon Amos, Hamilton Turner, Jules White Dept. of Electrical and Computer Engineering, Virginia Tech Blacksburg, Virginia, USA, "Applying machine learning classi? ers to dynamic Android malware detection at scale",2013.
3. Q. K. A. Mirza, I. Awan, M. Younas, Cloudintell: An intelligent malware detection system, Future Generation Computer Systems 86 (2018) 1042–1053.
4. Martin Hus´ak, Jana Kom´arkov´a, Elias Bou-Harb, and Pavel?Celeda, “Survey of Attack Projection, Prediction, and Forecasting in Cyber Security”,2018.
5. UmaraNoora,c, ZahidAnwara,b, TehminaAmjadc,Kim-KwangRaymond Chood, "A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise",2019
6. S. More, M. Matthews, A. Joshi, T. Finin, A knowledge-based approach to intrusion detection modelling, in: IEEE Symposium on Security and Privacy Workshops, San Francisco, CA, USA, IEEE, 2012, pp. 75–81.
7. V. Mulwad, W. Li, A. Joshi, T. Finin, K. Viswanathan, Extracting information about security vulnerabilities from web text, in: Proceedings of the 2011 IEEE ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology Workshops, WI-IAT 2011, Lyon, France, IEEE, 2011, pp. 257–260.
8. C. Sabottke, O. Suciu, T. Dumitras, Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits, in: 24th USENIX Security Symposium, USENIX Security 15, Washington, D.C., USA, USENIX, 2015, pp. 1041–1056.
9. I. Gha? r, M. Hammoudeh, V. Prenosil, L. Han, R. Hegarty, K. Rabie, F. J. Aparicio-Navarro, Detection of advanced persistent threat using machine-learning correlation analysis, Future Generation Computer Systems 89 (2018) 349–359.
10. R. A. Ahmadian and A. R. Ebrahimi, “A survey of IT early warning systems: architectures, challenges, and solutions,” Security and Communication Networks, vol. 9, no. 17, pp. 4751–4776, 2016.
11. H. Debar and A. Wespi, "Aggregation and correlation of intrusion detection alerts," in International Workshop on Recent Advances in Intrusion Detection. Springer, 2001, pp. 85–103.
12. S. Shin, S. Lee, H. Kim, and S. Kim, “Advanced probabilistic approach for network intrusion forecasting and detection,” Expert Systems with Applications, vol. 40, no. 1, pp. 315 – 322, 2013.
13. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin,S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, ``Tensor\_ow: A system for large-scale machine learning,'' in Proc. 12th USENIX Symp. Operating Syst. Design Implement., 2016, pp. 265\_283.
14. D. Baylor, ``TFX: A tensor\_ow-based production-scale machine learning platform,'' inProc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2017, pp. 1387\_1395.
15. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi, ``Learning activation functions to improve deep neural networks,'' 2014, arXiv:1412.6830. [Online]. Available: <https://arxiv.org/abs/1412.6830>.
16. Prabha, R., Senthil, G.A., Razmah, M., Akshaya, S.R., Sivashree, J., Cyrilla Swathi, J. (2023). A Comparative Study of SVM, CNN, and DCNN Algorithms for Emotion Recognition and Detection. In: Jacob, I.J., Kolandapalayam Shanmugam, S., Izonin, I. (eds) Data Intelligence and Cognitive Informatics. Algorithms for Intelligent Systems. Springer, Singapore. <https://doi.org/10.1007/978-981-19-6004-8_64>.
17. M. Razmah, S. G. A, R. Prabha, D. B, S. S and A. Naveen, "LSTM Method for Human Activity Recognition of Video Using PSO Algorithm," 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2022, pp. 1-6, doi: 10.1109/ICPECTS56089.2022.10046783.
18. R.M. Asha, P. Pondeepak, R. Prabha, G.A. Senthil, A. Padma Bharrathi, A novel approach effect of ocean acidification on oysters, Materials Today: Proceedings,2023,ISSN22147853,https://doi.org/10.1016/j.matpr.2023.01.194.
19. G. A. Senthil, P. Suganthi, R. Prabha, M. Madhumathi, S. Prabhu and S. Sridevi, "An Enhanced Smart Intelligent Detecting and Alerting System for Industrial Gas Leakage using IoT in Sensor Network," *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2023, pp. 397-401, doi: 10.1109/ICSSIT55814.2023.10060907.
20. G. T. Selvi, R. Prabha, G. A. Senthil, S. R. Krishnan V, S. A and J. R, "An Improved Internet of Things Based Accident Detection System Using Sensor Networks," *2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, Chennai, India, 2022, pp. 1-5, doi: 10.1109/ICPECTS56089.2022.10046887.
21. B. Zhao, D. Wang, G. Shi, D. Liu, and Y. Li, “Decentralized control for large-scale nonlinear systems with unknown mismatched interconnections via policy iteration,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 48, no. 10, pp. 1725–1735, Oct. 2018.
22. V. Narayanan, A. Sahoo, S. Jagannathan, and K. George, “Approximate optimal distributed control of nonlinear interconnected systems using event-triggered nonzero-sum games,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 5, pp. 1512–1522, May 2019.
23. X. Yang and H. He, “Adaptive critic learning and experience replay for decentralized event-triggered control of nonlinear interconnected systems,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 50, no.11, pp. 4043–4055, Nov. 2020.
24. C. Liu, H. Zhang, G. Xiao, and S. Sun, “Integral reinforcement learnin based decentralized optimal tracking control of unknown nonlinear large-scale interconnected systems with constrained-input,” Neurocomputing, vol. 323, pp. 1–11, Jan. 2019.