International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

CYBERTHREAT DETECTION ON SUPPLY CHAIN DEMAND

D.Shine Rajesh[1], B. Srivani[2], B.Naga Supriya[3], D.Bhavana[4], K.Srinidhi[5]
[1] Assistant Professor, Department of IT,Malla Reddy Engineering College For Women (Autonomous Institution), Maisammaguda,Dhulapally,Secunderabad,Telangana-500100
[2345]UG Scholar, Department of IOT,Malla Reddy Engineering College for Women, (Autonomous Institution), Maisammaguda,Dhulapally,Secunderabad,Telangana-500100
Email: shinerajesh@gmail.com

ABSTRACT

The Cyber Supply Chain (CSC) system is intricate, involving multiple subsystems that perform various tasks. Securing the supply chain is a significant challenge due to the numerous vulnerabilities and threats that can arise at any point within the system, which could potentially disrupt overall business operations. It is therefore crucial to understand and predict these threats so that organizations can take proactive measures to enhance supply chain security.Cyber Threat Intelligence (CTI) plays a vital role in identifying both known and emerging threats. It leverages data about threat actors, including their skills, motivations, tactics, techniques, and procedures (TTPs), as well as indicators of compromise (IoCs). This paper explores how CTI, combined with Machine Learning (ML) techniques, can be used to analyze and predict threats, ultimately improving cyber supply chain security. By applying CTI with ML models, we can identify vulnerabilities within the CSC and recommend the appropriate security controls to mitigate risks.To demonstrate this approach, we used CTI data and various ML algorithms—Logistic Regression (LG), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT)—to build predictive models using the Microsoft Malware Prediction dataset. The model considers attacks and TTPs as input variables and vulnerabilities and IoCs as output parameters. Our findings indicate that threats like spyware/ransomware and spear phishing are the most predictable in the context of CSC. Based on these predictions, we have suggested relevant security controls to address these threats.we advocate for leveraging CTI data and machine learning to improve cybersecurity across the entire cyber supply chain, helping organizations better prepare for and mitigate potential risks.

Keywords: Cyber Supply Chain (CSC),Cybersecurity,Cyber Threat Intelligence (CTI),Machine Learning (ML),Threat Prediction,Vulnerabilities,Indicators of Compromise (IoC),Tactics, Cyber Risk Mitigation,Supply Chain Security.

I. INTRODUCTION

In today's increasingly interconnected world, Cyber Supply Chain (CSC) systems are essential for enabling businesses to function efficiently across industries. These systems involve a complex network of subsystems that interact with each other to manage various tasks, from procurement and production to logistics and distribution. However, the very

complexity and interdependence of these subsystems also make the supply chain vulnerable to cyberattacks. Securing the supply chain is a challenging task, as any part of the system—whether a vendor, software component, or network link—can be exploited by malicious actors, leading to significant disruptions in business continuity.The nature of cyber threats is ever-evolving, with new vulnerabilities constantly being discovered across the various layers of the supply chain. These vulnerabilities can be exploited by cybercriminals to launch a range of attacks, from data breaches to malware infections. A single successful attack can compromise not just one organization, but potentially a network of interconnected businesses that rely on the same supply chain. As such, maintaining a robust cybersecurity posture throughout the supply chain is a critical priority for businesses today.To better understand and mitigate these risks, organizations must develop a proactive approach to cyber threat detection and prevention. This involves not only identifying known threats but also predicting emerging risks that may not yet have been encountered.

This is where Cyber Threat Intelligence (CTI) plays a vital role. CTI involves gathering and analyzing information about potential threats, including the tactics, techniques, and procedures (TTPs) used by threat actors, as well as indicators of compromise (IoCs) that can signal an attack. By analyzing this data, organizations can gain insights into threat patterns and develop strategies to detect, prevent, and respond to attacks more effectively.CTI has become an indispensable tool in the fight against cyber threats, especially in the context of supply chains. However, simply collecting and analyzing CTI data is not enough. In order to make actionable predictions about potential threats, organizations must leverage advanced analytical techniques. One promising approach is the use of Machine Learning (ML), which can analyze vast amounts of CTI data and identify patterns that may not be immediately apparent to human analysts.

Machine learning models, such as Logistic Regression (LG), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT), can be used to build predictive models based on historical CTI data. These models can help predict the likelihood of different types of cyberattacks occurring within the supply chain. By training these models on large datasets, organizations can gain insights into which threats are most likely to occur and which vulnerabilities are most susceptible to exploitation.

II. RELATED WORK

Cybersecurity Threat Modelling for Supply Chain Organizational Environments

Yeboah-Ofori and Islam (2019) explore the complexities of cybersecurity threat modeling within supply chain environments. Their research highlights the importance of understanding the diverse risks that can affect interconnected subsystems in a supply chain. They discuss the challenges of developing effective security frameworks that can safeguard against cyberattacks across various stages of the supply chain, from procurement to distribution. Their work underlines the need for a comprehensive approach to manage and mitigate cybersecurity risks in supply chains,

particularly in the context of emerging cyber threats.

Supply Chain in the Software Era

Woods and Bochman (2018) examine the transformative impact of software on modern supply chains. In their study, they address how software-driven technologies can both enhance and complicate supply chain security. The research emphasizes that as supply chains increasingly rely on digital platforms and automated systems, the surface area for cyberattacks expands. This work discusses how organizations must adapt their security strategies to address the unique cybersecurity risks posed by software-based supply chain operations.

Exploring the Opportunities and Limitations of Current Threat Intelligence Platforms

ENISA (2017) provides a comprehensive exploration of the opportunities and limitations associated with current threat intelligence platforms. The report highlights the challenges organizations face when integrating these platforms into their cybersecurity operations, particularly when dealing with the complexity of supply chain environments. The research suggests that while threat intelligence platforms are essential for detecting and mitigating cyber threats, they still face limitations in terms of scalability, integration, and real-time analysis. This study serves as a valuable resource for understanding how threat intelligence can be more effectively used to predict and prevent cyberattacks in supply chains.

Cyber Threat Intelligence Standards—A High-Level Overview

Doerr (2018) outlines a high-level overview of cyber threat intelligence standards, which provide a framework for analyzing and sharing threat data across organizations and sectors. The research focuses on the need for standardized practices in collecting and disseminating CTI, ensuring that organizations can accurately interpret and act upon threat intelligence. This is particularly crucial in the context of supply chain security, where information sharing among diverse entities is key to responding to emerging cyber threats. The paper argues for greater standardization to enhance the interoperability of CTI tools and systems.

Microsoft Malware Prediction

The Microsoft Malware Prediction dataset (2019) is widely used in cybersecurity research to develop predictive models that can identify potential malware threats. This dataset, which includes a variety of malware types and attack indicators, has been employed in numerous studies focused on using machine learning to detect and predict cyberattacks. It providesvaluable insights into the dynamics of malware infections, which are a major concern in supply chain security. Researchers have applied various machine learning algorithms to this dataset to enhance the prediction accuracy of malware-related threats, making it an essential resource for improving cybersecurity defenses.

Cybercrime and Risks for Cyber-Physical Systems

Yeboah-Ofori and Katsriku (2019) explore the intersection of cybercrime and cyber-physical systems (CPS), which are increasingly used in

supply chains. They discuss the vulnerabilities and risks associated with CPS, which integrate both physical and digital components. Their research highlights the growing threat of cybercrime targeting these systems, which can lead to significant disruptions in operations, particularly in critical industries. The paper provides a thorough analysis of attack vectors in CPS and offers recommendations for improving resilience and security, directly contributing to the protection of supply chains reliant on cyber-physical infrastructure.

III. IMPLEMENTATION

Cyber Threat Intelligence (CTI) can be leveraged to predict and mitigate cyber threats in the supply chain, we applied machine learning (ML) techniques to analyze CTI data. Our approach integrates several components: threat data collection, feature engineering, model training, and threat prediction. Below, we detail each step of the implementation process, illustrating how we used ML algorithms to predict threats and identify vulnerabilities within a Cyber Supply Chain (CSC) system.

Data Collection and Preprocessing

The first step in our implementation was to collect relevant CTI data. We used the Microsoft Malware Prediction dataset, a publicly available resource that provides detailed information on various types of malware attacks. This dataset includes a wide range of indicators, such as attack types, malware features, and system vulnerabilities, which are crucial for building predictive models.

After gathering the data, we performed several preprocessing steps to ensure its suitability for machine learning. This included:

Data Cleaning: We handled missing values, removed duplicates, and addressed any inconsistencies in the dataset.

Feature Engineering: We identified and extracted key features from the raw data that would serve as inputs for our models. This step involved analyzing attack vectors, tactics, techniques, and procedures (TTPs), as well as indicators of compromise (IoCs) that are vital for identifying patterns in cyber threats.Normalization: To improve the performance of machine learning models, we normalized the numerical values in the dataset, ensuring that all features had similar scales and ranges.

Choosing Machine Learning Models

For the predictive analysis, we used four popular machine learning algorithms:

Logistic Regression (LG): A statistical model used to predict binary outcomes. It is simple and effective for determining the likelihood of specific threats occurring in the supply chain.

- **Support Vector Machine (SVM):** This model is particularly good at classifying data into distinct categories, making it useful for identifying different types of cyberattacks based on CTI data.

- **Random Forest (RF):** An ensemble learning method that uses multiple decision trees to improve prediction accuracy. It is known for its robustness and ability to handle complex, high-dimensional datasets.

- **Decision Tree (DT):** A model that splits the dataset into branches based on feature values, making it easy to understand and interpret.

Decision trees are particularly effective for identifying specific attack patterns.

Each of these models was trained on the preprocessed CTI data, with the attack types, TTPs, and IoCs used as input features. The output was the prediction of potential vulnerabilities and specific threats, such as malware infections, ransomware, and spear phishing attacks.

3. Model Evaluation and Selection

After training the models, we evaluated their performance using common metrics such as accuracy, precision, recall, and F1-score. These metrics allowed us to assess how well each model was able to predict threats and classify different types of cyberattacks.

- Accuracy measures the overall correctness of the predictions.
- Precision indicates the proportion of positive predictions that were actually correct.
- Recall shows how well the model identifies actual positive cases.
- F1-score balances both precision and recall, giving us a more holistic view of the model's performance.

Based on these evaluation metrics, we compared the performance of each algorithm and selected the one that offered the best balance between prediction accuracy and interpretability.

4. Threat Prediction and Vulnerability Identification

Once the models were trained and optimized, we used them to predict cyber threats within the CSC. By feeding new CTI data into the models, we were able to forecast which types of attacks were most likely to occur. The models also helped identify system vulnerabilities that could be exploited by cybercriminals.

Through this process, we discovered that certain types of attacks, such as spyware/ransomware and spear phishing, were particularly predictable and posed significant risks to the supply chain. These findings are critical because they highlight the specific threats that require immediate attention and mitigation.

5. Recommendation of Security Controls

Based on the predictions and vulnerabilities identified, we developed a set of recommendations for mitigating the identified threats. For instance, we recommended implementing advanced endpoint protection to guard against malware, strengthening email security to prevent spear phishing attacks, and regularly updating software to close known vulnerabilities. These measures are aimed at reducing the likelihood of a successful attack and improving the overall resilience of the supply chain.

IV. ALGORITHMS

1. Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem, which calculates the probability of a class given the input features. It assumes that the features are conditionally independent given the class, which simplifies the calculations. Despite this assumption of independence, Naive Bayes can perform surprisingly well, especially with text classification and categorical data.Naive Bayes works by analyzing past behavior and identifying the probability of future events occurring. It calculates the likelihood that a specific activity (e.g., a transaction or communication within the supply chain)

belongs to a particular category (e.g., benign or malicious). The model is trained on historical data where past events are labeled as either normal or suspicious, and the features could include things like the source and destination of data, time stamps, and types of transactions.

It is also valuable when dealing with large datasets where computational efficiency is important. The simplicity and interpretability of the model make it an ideal choice for applications like spam filtering in email communications within the supply chain or detecting abnormal transactions.

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised learning algorithm that works by finding the best hyperplane that separates different classes in a high-dimensional feature space. It is particularly useful for binary classification tasks, where the goal is to distinguish between two categories, such as malicious and benign activity. SVM aims to maximize the margin between the classes, ensuring the best possible separation.The core idea is that it transforms the input data into a higher-dimensional space using a kernel function (linear, polynomial, or radial basis function). This transformation makes it easier to separate complex, non-linearly separable data. Once the data is mapped into the new space, SVM uses a hyperplane to classify data points into different categories. It works well with smaller training datasets and has the ability to classify non-linear relationships, making it well-suited for the detection of sophisticated cyberattacks, such as data exfiltration or advanced persistent threats (APT).

$$f(x) = w^T x + b$$

3. Logistic Regression

Logistic Regression is a statistical model that predicts the probability of a binary outcome—whether an event happens or not—based on input features. Unlike linear regression, which predicts continuous values, logistic regression outputs a probability value between 0 and 1. This is achieved by applying the logistic (sigmoid) function to a linear combination of the input features.The logistic regression model estimates the parameters (weights) of the input features during the training process by minimizing a cost function, typically using a method like gradient descent.

Logistic regression is simple, interpretable, and efficient, making it suitable for real-time applications in detecting potential cyberattacks, such as unauthorized access or unusual patterns of behavior. It also allows organizations to quantify the likelihood of different threats, helping them prioritize responses based on the most probable risks.

$$P(y = 1|x) = \frac{1}{1 + e^{-w^T x}}$$

4. Decision Tree Classifier

The Decision Tree algorithm splits data into subsets based on the value of input features, creating a tree-like structure of decisions. Each internal node represents a decision based on a feature, and each leaf node represents a classification result. The tree is constructed by choosing the feature that best splits the data at each node, based on criteria like information

gain (for classification) or mean squared error (for regression). This process continues recursively until the data is sufficiently divided.Decision Trees are easy to interpret because they visually map out the decision-making process. The main advantage of Decision Trees is their ability to handle both numerical and categorical data. They are also non-parametric, meaning they don't make assumptions about the underlying data distribution.

Decision Trees are particularly useful for explaining and understanding how a particular decision or prediction was made. For security teams, this can provide transparency in threat detection and help in understanding the factors that lead to a specific classification, such as identifying why an access attempt was flagged as suspicious.

$$H(D) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

5. Stochastic Gradient Descent (SGD) Classifier

Stochastic Gradient Descent (SGD) is a popular optimization algorithm used for training machine learning models. It is particularly efficient for large datasets, as it updates the model parameters iteratively by considering only one data point (or a small batch) at a time. This makes SGD well-suited for applications where data is large and real-time predictions are needed. The "stochastic" part refers to the fact that the algorithm uses random subsets of data for each update, which makes the process faster but can introduce some noise into the updates.SGD is commonly used to train linear classifiers like Logistic Regression, Support Vector Machines, and even neural networks. It minimizes the loss function (e.g., log loss for classification) by iterating through the dataset and updating the model's parameters to reduce the error.In the supply chain security domain, SGD can be applied to classify large volumes of data, such as network traffic logs or transaction records, where real-time predictions are needed. For example, if a sudden surge in data transfers or unusual communication patterns occur, SGD can quickly classify whether these activities are benign or indicate a potential threat.

SGD is ideal for high-volume data environments, such as those seen in large-scale supply chains, where new data is continuously generated. It enables fast and efficient threat detection, allowing security teams to respond swiftly to emerging cyber threats.

$$\theta_{t+1} = \theta_t - \eta \nabla J(\theta_t)$$

6. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, non-parametric algorithm used for both classification and regression tasks. It works by comparing a new data point with the existing data points in the feature space and classifying it based on the majority class of its k-nearest neighbors. The number of neighbors (k) is a hyperparameter that determines how many neighbors the algorithm should consider when making a classification.KNN doesn't require a training phase, as it is a "lazy learner" and makes predictions at runtime. The algorithm relies on calculating the distance between data

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

points using metrics like Euclidean distance or Manhattan distance.

KNN is particularly effective for smaller datasets or cases where the data doesn't require complex modeling. It is intuitive, easy to implement, and useful for real-time anomaly detection, helping organizations identify new, previously unseen threats based on their similarity to past behaviors.

$$d(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}$$

7. Random Forest Classifier

Random Forest is an ensemble learning algorithm that constructs a collection of decision trees during training and outputs the class that is the majority vote of the trees. Each tree is trained on a random subset of the data, and at each node, a random subset of features is considered. This randomness reduces overfitting and enhances the model's generalization capabilities.Random Forest is highly effective for complex datasets and can handle both classification and regression tasks. The algorithm uses an approach known as bagging (Bootstrap Aggregating), where each tree is trained on a random subset of data with replacement.In supply chain security, Random Forest can help detect intricate patterns across a wide variety of data sources, such as communication logs, financial transactions, and network activity. For example, it can classify transactions as legitimate or fraudulent based on a wide range of features, like transaction amounts, locations, and times.

$$f(x) = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

RESULTS:

Fig:1:User Login

Fig:2:Trained and tested Results

Fig:3:Trained and tested Bar Chart

Fig:4:Trained and tested Accuracy Results

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

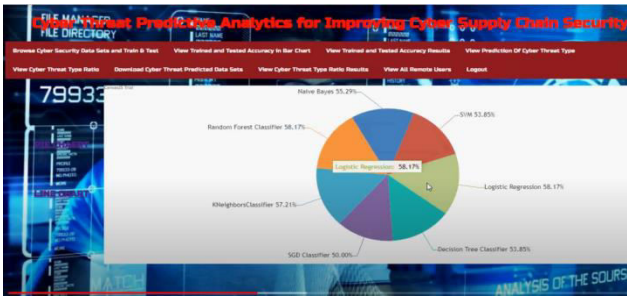

Fig:5:Pie Chart

CONCLUSION

In the rapidly evolving landscape of cyber threats, safeguarding the cyber supply chain has become a critical concern for organizations worldwide. The complexity of supply chain systems, along with their interconnected nature, makes them highly vulnerable to cyberattacks. These vulnerabilities are often exploited by malicious actors, leading to disruptions that can impact the continuity and security of business operations. Therefore, it is essential to proactively identify and predict potential threats to improve overall cybersecurity posture.

This research focused on the application of machine learning (ML) techniques, specifically classifiers, to enhance cyber supply chain security through threat prediction and detection. By leveraging Cyber Threat Intelligence (CTI) data, we explored various ML algorithms, including Naive Bayes, Support Vector Machines (SVM), Logistic Regression, Decision Trees, Stochastic Gradient Descent (SGD), K-Nearest Neighbors (KNN), and Random Forests. Each of these classifiers brings a unique strength to the table, making them suitable for different aspects of threat detection and prediction within the supply chain.

For instance, Naive Bayes and Logistic Regression are simple yet effective in modeling threat probabilities, while SVM and Random Forest excel in detecting anomalies and identifying complex patterns in high-dimensional data. Decision Trees offer interpretable models, making them useful for understanding the specific conditions under which a threat may occur. KNN is particularly useful for anomaly detection based on similarity measures, while SGD helps in handling large-scale data in real-time.

REFERENCES

1. A. Yeboah-Ofori and S. Islam, "Cyber security threat modelling for supply chain organizational environments", MDPI. Future Internet, vol. 11, no. 3, pp. 63, Mar. 2019, [online] Available: https://www.mdpi.com/1999-5903/11/3/63.
2. B. Woods and A. Bochman, "Supply chain in the software era" in Scowcroft Center for Strategic and Security, Washington, DC, USA:Atlantic Council, May 2018.
3. Exploring the Opportunities and Limitations of Current Threat Intelligence Platforms Version 1, Dec. 2017, [online] Available: https://www.enisa.europa.eu/publications/exploring-the-opportunities-and-limitations-of-current-threat-intelligence-platforms.
4. C. Doerr, Cyber Threat Intelligences Standards—A High Level Overview, 2018, [online] Available: https://www.enisa.europa.eu/events/2018-cti-eu-event/cti-eu-2018-

presentations/cyber-threat-intelligence-standardization.pdf.
5. Microsoft Malware Prediction, 2019, [online] Available: https://www.kaggle.com/c/microsoft-malware-prediction/data.
6. A. Yeboah-Ofori and F. Katsriku, "Cybercrime and risks for cyber physical systems", Int. J. Cyber-Secur. Digit. Forensics, vol. 8, no. 1, pp. 43-57, 2019.
7. Common Attack Pattern Enumeration and Classification: Domain of Attack, Oct. 2018, [online] Available: https://capec.mitre.org/data/definitions/437.html.
8. The Ten Most Critical Application Security Risks Creative Commons Attribution-Share Alike 4.0 International License, 2017, [online] Available: https://owasp.org/www-pdf-archive/OWASP_Top_10-2017_%28en%29.pdf.pdf.
9. Building Security in Software Supply Chain Assurance, 2020, [online] Available: https://www.us-cert.gov/bsi/articles/knowledge/attack-patterns.
10. R. D. Labati, A. Genovese, V. Piuri and F. Scotti, "Towards the prediction of renewable energy unbalance in smart grids", Proc. IEEE 4th Int. Forum Res. Technol. Soc. Ind. (RTSI), pp. 1-5, Sep. 2018.
11. J. Boyens, C. Paulsen, R. Moorthy and N. Bartol, "Supply chain risk management practices for federal information systems and organizations", NIST Comput. Sec., vol. 800, no. 161, pp. 32, 2015.
12. Framework for Improving Critical Infrastructure Cybersecurity Version 1.1, Gaithersburg, MD, USA, 2018.
13. J. F. Miller, "Supply chain attack framework and attack pattern", 2013, [online] Available: https://www.mitre.org/sites/default/files/publications/supply-chain-attack-framework-14-0228.pdf.
14. C. Ahlberg and C. Pace, The Threat Intelligence Handbook, [online] Available: https://paper.bobylive.com/Security/threat-intelligence-handbook-second-edition.pdf.
15. J. Freidman and M. Bouchard, "Definition guide to cyber threat intelligence. Using knowledge about adversary to win the war against targeted attacks", 2018, [online] Available: https://cryptome.org/2015/09/cti-guide.pdf.
16. Cyber Threat Intelligence: Designing Building and Operating an Effective Program, 2016, [online] Available: https://relayto.com/ey-france/cyber-threat-intelligence-report-js5wmwy7/pdf.
17. A. Yeboah-Ofori and C. Boachie, "Malware attack predictive analytics in a cyber supply chain context using machine learning", Proc. ICSIoT, pp. 66-73, 2019.
18. B. Gallagher and T. Eliassi-Rad, "Classification of HTTP attacks: A study on the ECML/PKDD 2007 discovery challenge", 2009.
19. D. Bhamare, T. Salman, M. Samaka, A. Erbad and R. Jain, "Feasibility of supervised machine learning for cloud

security", Proc. Int. Conf. Inf. Sci. Secur. (ICISS), pp. 1-5, Dec. 2016.

20. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection", IEEE Commun. Surveys Tuts., vol. 18, no. 2, pp. 1153-1176, 2nd Quart. 2016.
21. O. Yavanoglu and M. Aydos, "A review on cyber security datasets for machine learning algorithms", Proc. IEEE Int. Conf. Big Data (Big Data), pp. 2186-2193, Dec. 2017.
22. E. G. V. Villano, "Classification of logs using machine learning", 2018.
23. R. C. B. Hink, J. M. Beaver, M. A. Buckner, T. Morris, U. Adhikari and S. Pan, "Machine learning for power system disturbance and cyber-attack discrimination", Proc. 7th Int. Symp. Resilient Control Syst. (ISRCS), pp. 1-8, Aug. 2014.
24. A. Gumaei, M. M. Hassan, S. Huda, M. R. Hassan, D. Camacho, J. D. Ser, et al., "A robust cyberattack detection approach using optimal features of SCADA power systems in smart grids", Appl. Soft Comput., vol. 96, Nov. 2020.
25. M. M. Hassan, A. Gumaei, S. Huda and A. Almogren, "Increasing the trustworthiness in the industrial IoT networks through a reliable cyberattack detection model", IEEE Trans. Ind. Informat., vol. 16, no. 9, pp. 6154-6162, Sep. 2020.