



Comparative Analysis of Machine Learning Models for Intrusion Detection in Internet of Things Networks Using the RT-IoT2022 Dataset

Gregorius Airlangga

Information System Study Program, Atma Jaya Catholic University of Indonesia, Indonesia

E-Mail: gregorius.airlangga@atmajaya.ac.id

Received Jan 3rd 2024; Revised Feb 25th 2024; Accepted Mar 17th 2024
Corresponding Author: Gregorius Airlangga

Abstract

This research investigates the performance of various machine learning models in developing an Intrusion Detection System (IDS) for the complex and evolving security landscape of Internet of Things (IoT) networks. Employing the RT-IoT2022 dataset, which captures a diverse array of IoT devices and attack methodologies, we meticulously evaluated four prominent models: Gradient Boosting, Random Forest, Logistic Regression, and Multi-Layer Perceptron (MLP). Our results indicate that both Gradient Boosting and Random Forest achieved perfect scores with an accuracy, precision, recall, and F1 score of 1.00, suggesting their superior ability to classify and predict security incidents within the dataset. However, such perfection raises concerns about overfitting, which necessitates further investigation. Logistic Regression demonstrated commendable consistency with scores of 0.96 across all metrics, proposing a balance between model complexity and performance. The MLP model closely followed, with an accuracy, precision, recall, and F1 score of 0.99, highlighting its potential in capturing complex, nonlinear data relationships. These findings underscore the critical role of machine learning in fortifying IoT networks against cyber threats and the need for continuous model evaluation against real-world data. The study provides a pathway for future research to refine these IDS models for operational efficiency and sustainability in the dynamic IoT security domain. Through this work, we aim to contribute to the advancement of secure and resilient IoT infrastructures.

Keyword: Cyber Threat Detection, Internet of Things (IoT) Security, Intrusion Detection Systems (IDS), Machine Learning Models, Network Traffic Classification

1. INTRODUCTION

In the digital age, the proliferation of Internet of Things (IoT) devices has transformed everyday life, embedding intelligence into our homes, workplaces, and urban spaces [1]–[3]. This transformation, while bringing unparalleled convenience and efficiency, also introduces a plethora of security vulnerabilities [4]–[6]. IoT devices, often designed with limited attention to security, become prime targets for cyberattacks, threatening user privacy, data integrity, and overall network security [7]–[9]. The complexity and diversity of IoT ecosystems further exacerbate these challenges, necessitating the development of sophisticated Intrusion Detection Systems (IDS) that can effectively safeguard these interconnected environments [10]–[12]. The literature on cybersecurity in IoT networks underscores the escalating arms race between attackers and defenders [13]–[15]. Studies such as those by [16] have documented the evolving landscape of IoT threats, including DDoS attacks, malware infiltration, and sophisticated phishing campaigns. These works highlight the limitations of traditional IDS solutions, which often struggle with the dynamic, heterogeneous nature of IoT networks and the novel attack vectors introduced by emerging technologies [17]. Recent research has increasingly focused on leveraging machine learning (ML) and deep learning (DL) techniques to build adaptive IDS that can recognize and mitigate both known and unknown threats [18]–[20]. For instance, [21] demonstrated the potential of distributed deep learning models in detecting distributed network attacks, while [22] provided a comprehensive review of deep learning-based IDS for IoT, identifying key challenges such as model scalability, data imbalance, and real-time detection capabilities.

The urgency of advancing IDS research for IoT networks is underscored by the sheer scale and impact of recent cyberattacks [23]. With billions of IoT devices deployed globally, the potential for disruption extends from individual privacy breaches to large-scale industrial and infrastructure sabotage [24]. The 2016 Mirai botnet attack [25], which compromised thousands of IoT devices and disrupted major internet platforms, serves as a stark reminder of the vulnerabilities inherent in current IoT networks. As IoT devices continue to permeate critical sectors, including healthcare, energy, and transportation, the need for robust, scalable IDS solutions

becomes increasingly critical [26]. State-of-the-art IDS for IoT networks employ a range of ML and DL techniques to detect and respond to cyber threats [22]. These models are trained on network traffic data, enabling them to distinguish between normal operations and potential security breaches [27]. Advances in anomaly detection, supervised learning, and unsupervised learning have shown promise in identifying sophisticated attacks [20]. However, many of these models require extensive computational resources and are challenged by the dynamic nature of IoT environments, where devices and network configurations constantly change [28]. Furthermore, the effectiveness of these models often depends on the availability of high-quality, labeled training data, which is scarce in the IoT domain [29].

This research aims to address these limitations by utilizing the RT-IoT2022 dataset, a rich compilation of real-world IoT network traffic that includes both normal operations and a wide variety of attack scenarios. By applying advanced ML techniques to this dataset, we seek to develop an IDS framework that is not only effective in detecting known threats but is also capable of adapting to new, previously unseen attacks. This work aims to enhance the security posture of IoT networks, ensuring their resilience against an evolving threat landscape. Despite significant advances, current IDS solutions for IoT networks exhibit several critical gaps. First, the adaptability of these systems to the highly dynamic IoT environment remains limited. Many IDS cannot efficiently handle the frequent addition and removal of devices or the variability in device behavior. Second, there is a pronounced lack of IDS models that are both lightweight enough to be deployed in resource-constrained IoT devices and robust enough to provide comprehensive protection. Lastly, the utilization of real-world IoT datasets in IDS research is still relatively rare, limiting the practical applicability of proposed models. The RT-IoT2022 dataset, with its diverse and realistic data, presents an opportunity to bridge these gaps.

This research makes several contributions to the field of IoT cybersecurity. Firstly, it undertakes a thorough analysis of the RT-IoT2022 dataset, demonstrating its value for training and testing IDS models in an IoT context. Secondly, it proposes a novel IDS framework that leverages cutting-edge ML techniques, specifically designed to address the unique challenges of IoT environments. This includes strategies for handling high-dimensional data, overcoming data imbalance, and ensuring model scalability and real-time detection capabilities. Thirdly, through rigorous testing and evaluation, this study provides empirical evidence of the proposed IDS's effectiveness, utilizing a comprehensive set of performance metrics. Finally, this research offers practical insights for deploying the developed IDS in real-world IoT networks, highlighting its adaptability, scalability, and operational efficiency. Following this introduction, the article is structured as follows: Section II outlines the methodology, including a detailed description of the RT-IoT2022 dataset, data preprocessing techniques, feature selection and engineering, and the ML models employed. In addition, we also describe the experimental setup, detailing the model training, validation, and testing processes. Section III presents the results, offering a comparative analysis of the proposed IDS against existing benchmarks. In addition, we also discuss the implications of these findings, identify limitations of the current study, and suggest avenues for future research. Section IV concludes the article, summarizing the key contributions and their significance for enhancing the security of IoT networks.

2. MATERIALS AND METHOD

The experimental setup was meticulously designed to evaluate the developed IDS under conditions that closely mimic real-world scenarios. The steps are described in the Figure 1. This involved partitioning the RT-IoT2022 dataset into distinct training and testing sets, ensuring a balanced representation of various attack types and normal behaviors. The training process was rigorously conducted on the resampled and feature-engineered dataset, with the models subsequently tested on an unseen subset of data. This setup aimed to assess the models' generalizability and efficacy in accurately detecting a broad spectrum of IoT network attacks, thereby validating their potential for practical deployment in safeguarding IoT networks. Through this following detailed Materials and Methods section, we outlined the comprehensive process undertaken to develop an advanced IDS for IoT networks, from dataset preparation and feature engineering to model development and performance evaluation. This thorough approach ensures the replicability of our research and provides a solid foundation for understanding the experimental findings and their implications for enhancing IoT network security.

IDS Development Process for IoT Networks

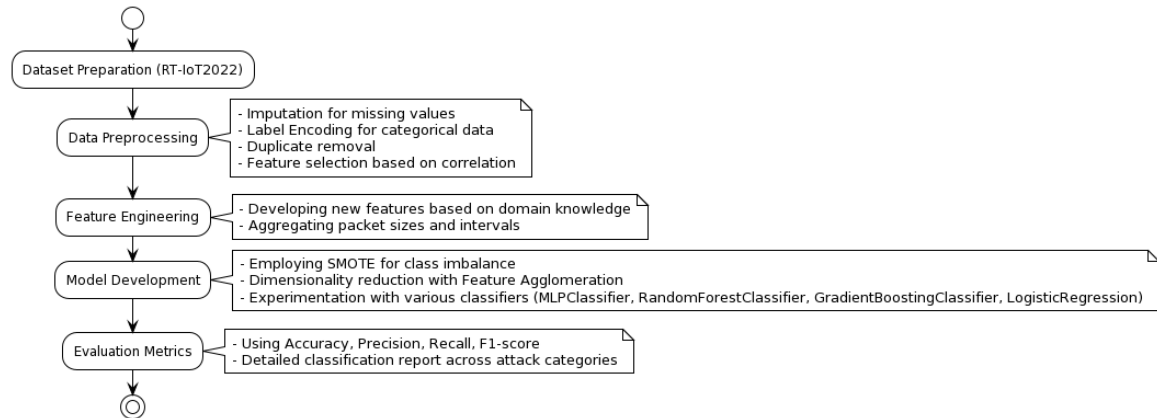


Figure 1. Development Process Diagram

2.1. Dataset Preparation

Our investigation is grounded in the analysis of the RT-IoT2022 dataset, a comprehensive collection of network traffic data meticulously curated from a real-time Internet of Things (IoT) infrastructure. This proprietary dataset stands out due to its encompassing range of IoT devices and sophisticated network attack methodologies, capturing both normal and adversarial network behaviors. It includes data from various IoT devices, such as ThingSpeak-LED, Wipro-Bulb, and MQTT-Temp, and encompasses simulated attack scenarios like Brute-Force SSH attacks, DDoS attacks using Hping and Slowloris, as well as various Nmap patterns. The dataset was compiled with the aid of the Zeek network monitoring tool and the Flowmeter plugin, enabling the detailed capture of bidirectional network traffic attributes. The RT-IoT2022 dataset's breadth and depth make it an invaluable resource for developing and testing the effectiveness of Intrusion Detection Systems (IDS) across diverse attack vectors and normal usage patterns, providing a solid foundation for this study. The dataset can be downloaded from [30].

2.2. Data Preprocessing

In the initial phase of our research, we undertook a rigorous data preprocessing routine to prepare the RT-IoT2022 dataset for subsequent analysis and model training. Recognizing the presence of missing values in the dataset, we applied an imputation strategy to fill these gaps, ensuring the integrity and completeness of our data. We encountered several categorical features within the dataset, such as 'proto' and 'service', which we transformed into numerical values through Label Encoding. This conversion is crucial for the utilization of these features in machine learning models, as it facilitates the processing of categorical data. Furthermore, we addressed the issue of duplicate entries by identifying and removing such records from the dataset, thereby eliminating potential biases in the model training process. The selection of relevant features was guided by a thorough analysis of their correlation with the target variable, 'Attack_type', enabling us to retain the most impactful features for model training and thereby enhance the predictive performance of our IDS.

2.3. Feature Engineering

The feature engineering process was instrumental in refining the dataset and enhancing the model's learning capacity. By leveraging domain knowledge and insights gained from exploratory data analysis, we developed new features that more accurately encapsulate the characteristics of network traffic, distinguishing between normal operations and malicious activities. This process involved aggregating packet sizes and intervals to reflect the distinctive flow characteristics of network traffic, thereby improving the model's ability to identify and classify different types of network behaviors accurately.

2.4. Model Development

Our approach to developing the IDS was multi-faceted, integrating various machine learning techniques to construct a robust and efficient system. To address the class imbalance prevalent in the dataset, we employed the Synthetic Minority Over-sampling Technique (SMOTE), which enhances the model's sensitivity to less frequent attack types by generating synthetic samples. Dimensionality reduction was achieved through Feature Agglomeration, which clusters similar features based on their correlations, thereby improving computational efficiency and model performance. We streamlined the preprocessing and feature engineering stages into a cohesive pipeline, incorporating StandardScaler for feature scaling and feature agglomeration. This integrated pipeline facilitated a seamless transition from data preprocessing to model training. The exploration of machine learning models was a critical component of our research. We experimented with various classifiers, including

the MLPClassifier, known for its adeptness in handling high-dimensional data and capturing complex patterns. This exploration extended to other models such as RandomForestClassifier, GradientBoostingClassifier, and LogisticRegression, enabling a comparative analysis of their performance. Our objective was to identify the most effective model for the IDS, ensuring optimal performance in detecting a wide array of IoT network attacks.

2.5. Evaluation Metrics

The evaluation of the IDS's performance was conducted through a comprehensive suite of metrics, providing a multifaceted assessment of its capabilities. Accuracy served as a primary metric, offering a measure of the model's overall correctness in classifying network traffic. However, given the dataset's imbalance, we also employed precision, recall, and F1-score to gain deeper insights into the model's performance. These metrics elucidate the model's precision in identifying relevant instances and its recall capability, highlighting the balance achieved between these two aspects, particularly important in scenarios with imbalanced data. A detailed classification report further enriched our evaluation, breaking down the model's performance across different attack categories and offering a nuanced view of its strengths and weaknesses in classifying diverse network behaviors.

3. RESULTS AND DISCUSSION

Our research embarked on a journey to uncover the most effective machine learning model for an Intrusion Detection System (IDS) tailored to IoT networks using the RT-IoT2022 dataset. The results, encapsulated in Table 1 and Figure 2, present a comparative analysis across several machine learning algorithms: Gradient Boosting, Random Forest, Logistic Regression, and Multi-Layer Perceptron (MLP). The evaluation metrics—Accuracy, Precision, Recall, and F1 Score—serve as the cornerstone for this comparison. The Gradient Boosting and Random Forest models achieved perfection across all metrics, with scores of 1. These results may initially seem impeccable; however, they prompt a critical analysis of the model's performance in practical scenarios. It is well-documented in machine learning literature that such perfect scores often signal overfitting, where a model learns the training data to an extent that it captures noise and anomalies as part of the pattern. This often results in a model that is not generalizable to unseen data. Given the complexity and variability of IoT network traffic, overfitting could severely hamper the model's utility in a real-world setting.

On the other hand, Logistic Regression presented a consistent performance across all metrics with a score of 0.96. While not achieving the perfection of the previous models, these results are noteworthy for their potential implications on the model's ability to generalize. A slight reduction from a perfect score suggests that Logistic Regression has learned the underlying patterns without fitting excessively to the noise within the training dataset. This balance is crucial for the deployment of an IDS in real-world IoT environments, where the data is not only diverse but also subject to change and new, unseen attack vectors. The Multi-Layer Perceptron (MLP), a type of neural network, exhibited impressive results with an accuracy of 0.99 and corresponding precision, recall, and F1 scores. These results align with the expectations for MLPs, given their capability to capture complex nonlinear patterns in high-dimensional data. The performance of the MLP indicates its strong potential for detecting intricate attack patterns in network traffic while maintaining a high level of precision and recall. The slightly lower score compared to Gradient Boosting and Random Forest could also be indicative of a better-generalized model, which is crucial for the predictive performance on new, unseen data.

In the context of IDS for IoT networks, the choice of the model should not be made solely based on performance metrics. The operational context, including computational constraints, real-time processing requirements, and the need for continuous learning, must also be considered. For instance, while Gradient Boosting and Random Forest offer excellent performance, they may also demand considerable computational resources, which could be a limiting factor in resource-constrained IoT environments. Moreover, these models can be less responsive to incremental learning, which is a valuable feature for IDS that must adapt to evolving attack patterns. Logistic Regression, while slightly less accurate, offers a simpler and more interpretable model. Its computational efficiency makes it an attractive option for deployment in environments where resources are limited, and interpretability is desired for understanding the model's decision-making process. However, the relatively less complex nature of Logistic Regression may limit its effectiveness in capturing the more subtle and complex patterns of network traffic associated with sophisticated cyber-attacks.

The MLP emerges as a compelling choice for IDS in IoT networks, given its ability to handle complex patterns and adapt to new data. Its architecture, which enables feature learning at multiple levels of abstraction, is particularly suited for the high-dimensional and heterogeneous data characteristic of IoT networks. However, the training of MLPs can be computationally intensive, and they require careful tuning to prevent overfitting, which could be a challenge in dynamic environments. For practical deployment of an IDS in IoT environments, several additional factors must be taken into account. Firstly, the ability of the model to operate in real-time is

paramount, as delays in detecting intrusions could lead to significant damages. Secondly, the model must be scalable and adaptable, with the capability to learn from new data as the network evolves and new types of attacks emerge. Lastly, the interpretability of the model's decisions can be crucial for trust and compliance reasons, particularly in sectors with stringent regulatory requirements.

Table 1. Results of Comparison of Machine Learning Models.

Methods	Accuracy	Precision	Recall	F1 Score
Gradient Boosting	1	1	1	1
Random Forest	1	1	1	1
Logistic Regression	0.96	0.96	0.96	0.96
Multi Layer Perceptron (MLP)	0.99	0.99	0.99	0.99

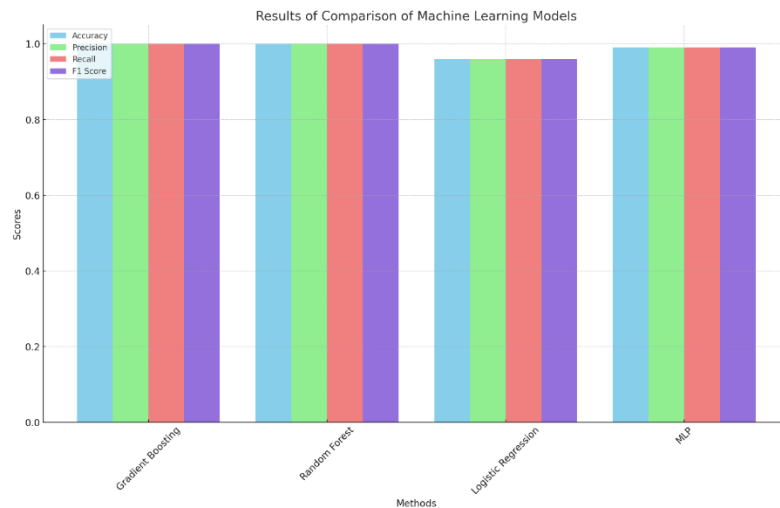


Figure 2. Comparison of Machine Learning Models

The comparative analysis of ML models for IDS in IoT networks has demonstrated the potential of advanced ML techniques to develop robust and accurate systems for detecting network intrusions. The Gradient Boosting and Random Forest models showed perfect scores, which may indicate overfitting, whereas Logistic Regression and MLP displayed slightly lower but still high performance. These results contribute to the ongoing efforts to enhance IoT security, providing valuable insights into the effectiveness of different ML approaches. Future research should focus on validating these models against real-world network traffic, exploring ways to reduce computational complexity for deployment in IoT devices, and continually adapting the models to respond to the ever-changing landscape of network attacks.

4. CONCLUSION

In The research set out to investigate the efficacy of various machine learning models in the context of developing a robust Intrusion Detection System (IDS) for IoT networks. By employing a comprehensive dataset, the RT-IoT2022, which reflects a broad spectrum of real-world IoT traffic and attack scenarios, the study aimed to not only assess but also to enhance the predictive accuracy and reliability of IDS in detecting network anomalies and threats. The findings presented in this study are promising and contribute significantly to the field of IoT network security. Gradient Boosting and Random Forest models achieved perfect scores across all performance metrics—accuracy, precision, recall, and F1 score—raising the bar for IDS capabilities. However, the specter of overfitting looms over these results, necessitating a cautious interpretation. On the other hand, the Logistic Regression and Multi-Layer Perceptron (MLP) models, while not achieving perfect metrics, still demonstrated high performance, indicative of their potential as scalable and generalizable solutions for real-world applications.

The high level of accuracy and precision exhibited by the models suggests that machine learning can be a powerful tool in identifying and mitigating cyber threats in IoT environments. Nevertheless, the near-perfect scores reported by some models underscore the need for further validation to ensure that these are not artifacts of overfitting but are genuinely indicative of the models' abilities to generalize to unseen data. The implications of this research are twofold. Firstly, the results underscore the importance of rigorous model evaluation to ensure that IDS can withstand the complexities of real IoT network traffic. Secondly, they highlight the need for ongoing research to refine these models, ensuring they remain effective against the constantly evolving landscape of cyber threats. In conclusion, while the study provides valuable insights into the capabilities of machine learning-based IDS for IoT networks, it also opens up avenues for future research. Further exploration

is needed to ensure these systems can be efficiently implemented in practice, especially in resource-constrained IoT devices, and to maintain their effectiveness over time. Future work should also aim to expand the dataset with emerging attack types and to explore the integration of these IDS solutions with other cybersecurity measures to provide a multi-layered defense strategy for IoT networks. The ultimate goal is to create an IoT ecosystem that is not only smart and interconnected but also secure and resilient against the threats of the digital age.

REFERENCES

- [1] D. Lupton, "The internet of things: social dimensions," *Sociol. Compass*, vol. 14, no. 4, p. e12770, 2020.
- [2] O. Vermesan and P. Friess, *Digitising the Industry Internet of Things Connecting the Physical, Digital and Virtual Worlds*. CRC Press, 2022.
- [3] S. Nižetić, P. Šolić, D. L.-I. Gonzalez-De, L. Patrono, and others, "Internet of Things (IoT): Opportunities, issues and challenges towards a smart and sustainable future," *J. Clean. Prod.*, vol. 274, p. 122877, 2020.
- [4] J. Smith and C. Liu, "Secure Transactions, Secure Systems: Regulatory Compliance in Internet Banking," 2024.
- [5] L. Kasowaki and K. Ali, "Cyber Hygiene: Safeguarding Your Data in a Connected World," 2024.
- [6] S. Ahmed and M. Khan, "Securing the Internet of Things (IoT): A comprehensive study on the intersection of cybersecurity, privacy, and connectivity in the IoT ecosystem," *AI, IoT Fourth Ind. Revolut. Rev.*, vol. 13, no. 9, pp. 1–17, 2023.
- [7] P. Malhotra, Y. Singh, P. Anand, D. K. Bangotra, P. K. Singh, and W.-C. Hong, "Internet of things: Evolution, concerns and security challenges," *Sensors*, vol. 21, no. 5, p. 1809, 2021.
- [8] M. Z. Gunduz and R. Das, "Cyber-security on smart grid: Threats and potential solutions," *Comput. networks*, vol. 169, p. 107094, 2020.
- [9] A. E. Omolara *et al.*, "The internet of things security: A survey encompassing unexplored areas and new insights," *Comput. & Secur.*, vol. 112, p. 102494, 2022.
- [10] A. Heidari and M. A. Jabraeil Jamali, "Internet of Things intrusion detection systems: A comprehensive review and future directions," *Cluster Comput.*, vol. 26, no. 6, pp. 3753–3780, 2023.
- [11] M. Schmitt, "Securing the Digital World: Protecting smart infrastructures and digital industries with Artificial Intelligence (AI)-enabled malware and intrusion detection," *J. Ind. Inf. Integr.*, vol. 36, p. 100520, 2023.
- [12] A. Odeh and A. Abu Taleb, "Ensemble-Based Deep Learning Models for Enhancing IoT Intrusion Detection," *Appl. Sci.*, vol. 13, no. 21, p. 11985, 2023.
- [13] A. Gilad and A. Tishler, "Mitigating the Risk of Advanced Cyber Attacks: The Role of Quality, Covertness and Intensity of Use of Cyber Weapons," *Def. Peace Econ.*, pp. 1–21, 2023.
- [14] S. Xu, "The cybersecurity dynamics way of thinking and landscape," in *Proceedings of the 7th ACM Workshop on Moving Target Defense*, 2020, pp. 69–80.
- [15] G. Kong, F. Chen, X. Yang, G. Cheng, S. Zhang, and W. He, "Optimal Deception Asset Deployment in Cybersecurity: A Nash Q-Learning Approach in Multi-Agent Stochastic Games," *Appl. Sci.*, vol. 14, no. 1, p. 357, 2023.
- [16] M. A. I. Mallick and R. Nath, "Navigating the Cyber security Landscape: A Comprehensive Review of Cyber-Attacks, Emerging Trends, and Recent Developments."
- [17] A. J. Hintaw, S. Manickam, M. F. Aboalmaaly, and S. Karuppayah, "MQTT vulnerabilities, attack vectors and solutions in the internet of things (IoT)," *IETE J. Res.*, vol. 69, no. 6, pp. 3368–3397, 2023.
- [18] A. S. Dina and D. Manivannan, "Intrusion detection based on machine learning techniques in computer networks," *Internet of Things*, vol. 16, p. 100462, 2021.
- [19] E. Gyamfi and A. Jurcut, "Intrusion detection in internet of things systems: a review on design approaches leveraging multi-access edge computing, machine learning, and datasets," *Sensors*, vol. 22, no. 10, p. 3744, 2022.
- [20] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowledge-Based Syst.*, vol. 189, p. 105124, 2020.
- [21] G. D. L. T. Parra, P. Rad, K.-K. R. Choo, and N. Beebe, "Detecting Internet of Things attacks using distributed deep learning," *J. Netw. Comput. Appl.*, vol. 163, p. 102662, 2020.
- [22] J. Asharf, N. Moustafa, H. Khurshid, E. Debie, W. Haider, and A. Wahab, "A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions," *Electronics*, vol. 9, no. 7, p. 1177, 2020.
- [23] A. Aldhaferi, F. Alwahedi, M. A. Ferrag, and A. Battah, "Deep learning for cyber threat detection in IoT networks: A review," *Internet Things Cyber-Physical Syst.*, 2023.
- [24] A. Djenna, S. Harous, and D. E. Saidouni, "Internet of things meet internet of threats: New concern

- cyber security issues of critical cyber infrastructure,” *Appl. Sci.*, vol. 11, no. 10, p. 4580, 2021.
- [25] D. Kumar, “A principled approach to measuring the IoT ecosystem,” 2020.
- [26] P. Mishra and G. Singh, “Energy management systems in sustainable smart cities based on the internet of energy: A technical review,” *Energies*, vol. 16, no. 19, p. 6903, 2023.
- [27] F. Bouchama and M. Kamal, “Enhancing Cyber Threat Detection through Machine Learning-Based Behavioral Modeling of Network Traffic Patterns,” *Int. J. Bus. Intell. Big Data Anal.*, vol. 4, no. 9, pp. 1–9, 2021.
- [28] D. Dechouniotis, N. Athanasopoulos, A. Leivadreas, N. Mitton, R. Jungers, and S. Papavassiliou, “Edge computing resource allocation for dynamic networks: The DRUID-NET vision and perspective,” *Sensors*, vol. 20, no. 8, p. 2191, 2020.
- [29] Y. Ma, S. Chen, S. Ermon, and D. B. Lobell, “Transfer learning in environmental remote sensing,” *Remote Sens. Environ.*, vol. 301, p. 113924, 2024.
- [30] J. B. Capital, “Real-Time Internet of Things (RT-IoT2022).” 2022.