

Ensemble Knowledge Distillation-based Federated Learning for Effective Intrusion Detection in Heterogeneous Networks

N. Shirisha¹, A. Prashanthi², N. Banupriya³, Syed Muqthadar Ali⁴, Dr. G. Amirthayogam⁵, Pavan Kumar Vadrevu⁶

¹Department of Computer Science and Engineering MLR Institute of Technology Dundigal, Hyderabad
Email: dr.nallashirisha@gmail.com

²Associate Professor, Department of CSE, Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad, Hyderabad, Telangana, India
Email: prashanthi.a@cse.nnrg.edu.in

³Assistant Professor Department of Computer and Communication Engineering Sri Eshwar College of Engineering
Email: banupriya.n@sece.ac.in

⁴Senior Assistant Professor Cvr College of Engineering CSE Hyderabad, Telangana
Email: sma.cvrce@gmail.com

⁵Assistant Professor (SG) Department of Computer Science and Engineering Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology Chennai
Email: amir.yogam@gmail.com

⁶Professor Department of Information Technology Shri Vishnu Engineering College for Women
Email: vadrevu.pavan@svecw.edu.in

Abstract—The rapid development of low-cost consumer electronics and cloud computing has led to the widespread adoption of Internet-of-Things (IoT) devices in various applications such as smart cities and industrial control systems. However, these devices are vulnerable to cyber attacks due to their open deployment environment and limited computing capabilities. Intrusion Detection Systems (IDS) have emerged as an effective way to secure IoT networks by monitoring and detecting abnormal activities. However, existing IDS approaches rely on centralized servers, which can cause high response time and operational costs. Additionally, sharing behaviour data in an open and distributed IoT network may violate privacy requirements. To address these challenges, the paper introduces Federated Learning (FL) as a collaborative training approach for developing a decentralized shared model of IDS. FL allows the training of a global IDS model without exposing training data to others. Local model parameters are transmitted to a central server, which aggregates them into a global model. The updated parameters are then distributed to all local clients. This approach reduces response time and communication overhead while preserving privacy.

Index Terms— Ensemble, IoT, IDS, Cloud, Federated Learning.

I. INTRODUCTION

In the era of digitalization, our daily life is gradually changed by ubiquitous Internet of Things (IoT) devices.

These devices record and upload our personal information from day to night, thus generating exponential data every day for every person. By integrating with big data and edge computing devices, a lot of promising applications have emerged, such as intelligent industrial control systems and smart cities. The document discusses the use of Federated Learning (FL) in Intrusion Detection Systems (IDS) for securing Internet-of-Things (IoT) networks. It highlights the challenges faced by traditional IDS approaches and proposes a method called Federated Learning Ensemble Knowledge Distillation (FLEKD) to address these challenges. FLEKD aims to train a decentralized shared model of IDS without compromising on-device privacy and to handle the heterogeneity of data captured by various IoT devices. Due to the openness of the IoT network deployment environment and the weak security control protocol operation ability caused by limited computing capabilities, IoT devices are very vulnerable to network attacks. Attackers may comprise edge devices or databases through the network to steal or tamper with data, causing great social and economic impact [1], [2]. Hence, security monitor detection measures should be considered to enhance IoT security. As one of the most important security posture monitoring tools, the Intrusion Detection System (IDS) has attracted much attention in the context of IoT [3]. IDS can notify an administrator of potential cyber-attacks by analysing network traffic packets or system logs. Traditional IDS systems rely on rule-based approaches [4] that are not effective to detect novel attacks. Furthermore, big data is another challenge for the rule-based IDSs. With the growth of IoT networks and increasing complexity of attacks, Machine Learning (ML) and Deep Learning (DL) methods have emerged as effective techniques for constructing automatic IDSs [5], [6]. These techniques enable the system to learn from vast amounts of behaviour data, establish specific models efficiently, and identify attacking patterns that are difficult to detect manually. However, the use of ML and DL in IDS for IoT networks comes with its own set of challenges. First, training models by local devices themselves will lead to poor performance because of the constrained computation ability and insufficient data. Second, the communication overhead is large and will cause privacy concerns to transferring raw data of all devices to a central server for collaborative training. Another challenge is handling the heterogeneous data collected by different devices [7]. The data distribution, number of samples, and collection time on each edge device may be different. In recent years, there are also research efforts investigating emerging ML or DL paradigms to improve specific performance of IDS models, such as improving model adaptability to handle heterogeneous data by collaborative learning [9] and reducing data annotation pressure by weakly supervised learning [10].

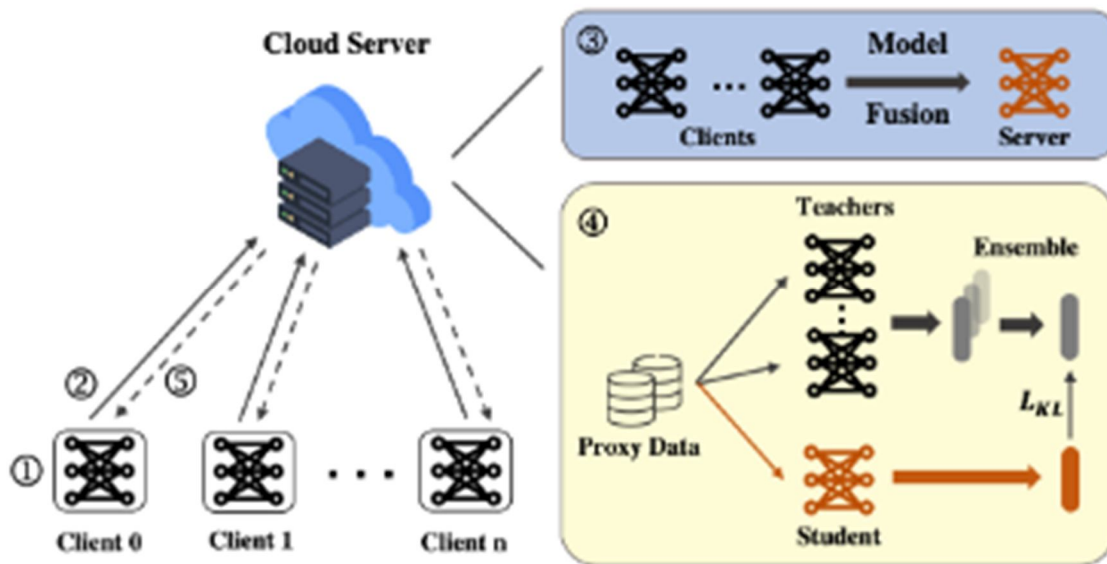


Figure-1. The overview of the FLEKD-IDS framework. ①: Train local model. ②: Transmit local models to server. ③: Fuse clients' models. ④: Ensemble knowledge distillation and fine-tune the global model. ⑤: Distribute the latest global model.

The document provides background information on IDS and its role in network security. It explains that traditional IDS systems are rule-based or signature-based, which are effective for detecting known threats but may not be as advanced as ML/DL-based IDSs in analyzing large volumes of data and detecting emerging attacks.

The document also mentions the research efforts in using ML and DL techniques to improve IDS performance.

II. BENEFITS OF FEDERATED LEARNING

Federated Learning (FL) is a distributed machine learning paradigm that allows collaborative training of a global model without sharing raw data. It offers several advantages over centralized ML and DL models, including improved privacy protection, reduced communication overhead, and the ability to handle heterogeneous data.

A. Ensemble Knowledge Distillation for Heterogeneous Data

The paper proposes an effective method called Federated Learning Ensemble Knowledge Distillation (FLEKD) to mitigate the heterogeneity problems across various clients. FLEKD leverages mutually exclusive and unlabeled data to aggregate and transfer knowledge from heterogeneous client models to a global model. This technique allows for flexible aggregation, reducing the impact.

III. AN EFFECTIVE HETEROGENEOUS INTRUSION DETECTION SYSTEM FRAMEWORK

The document proposes an effective framework called Federated Learning Ensemble Knowledge Distillation (FLEKD) for IDS in heterogeneous IoT networks. FLEKD utilizes Federated Learning (FL) to collaboratively train a decentralized shared model of IDS without exposing training data to others. It also introduces the concept of knowledge distillation, which allows transferring the knowledge of a large, complex model to a smaller, simpler model. FLEKD leverages ensemble knowledge distillation to mitigate the heterogeneity problems across various clients and improve the model convergence speed.

A. Knowledge Distillation

Knowledge distillation is a technique used to train a student model to mimic the behavior of a teacher model. It involves transferring the knowledge from the teacher model to the student model using a proxy dataset as the medium. The class probabilities of the teacher model's last layer or the feature representations of its middle hidden layers are used as soft targets for knowledge transfer. The distillation loss term, often formulated as the Kullback-Leibler (KL) divergence, is used to measure the difference between the teacher and student's softmax output probabilities. In the context of federated learning, where there is no publicly labeled dataset, the distillation loss term is the only loss term used.

Kullback-Leibler (KL) divergence is a measure of the difference between two probability distributions. In the context of knowledge distillation, it is used to quantify the difference between the softmax output probabilities of the teacher model and the student model. It is calculated using the formula (i),

$$LKL(S||T) = \sum Si(\sigma(z)) \log(Si(\sigma(z))/Ti(\sigma(z))), \quad (i)$$

where S and T represent the student and teacher logits respectively, σ is the softmax function, and $Si(\sigma(z))$ and $Ti(\sigma(z))$ are the softmax output probabilities of the student and teacher models respectively.

B. FLEKD: Using Ensemble Knowledge Distillation for IDS in Heterogeneous IoT Networks

FLEKD (Federated Learning Ensemble Knowledge Distillation) is a method proposed in the document for training a decentralized shared model of Intrusion Detection Systems (IDS) in IoT networks. It utilizes Federated Learning (FL) to collaboratively train the model without exposing training data to others. FLEKD addresses the heterogeneity problem in IoT networks by using ensemble knowledge distillation, which allows for a flexible aggregation method that reduces the impact of clients' heterogeneity while maintaining user privacy. Experimental results demonstrate that FLEKD outperforms local training and traditional FL in terms of both speed and performance, improving the system's ability to identify unknown attacks compared to local training models and naive FL global models.

IV. EXPERIMENTAL RESULTS

The document mentions that extensive experiments were conducted on the CICIDS2019 dataset to evaluate the performance of the proposed FLEKD framework. The results demonstrate better detection performance and improved ability to identify unknown attacks compared to local training models and naive FL global models. The framework also exhibits clear advantages in handling diverse data features, sample quantity, and missing certain classes.

Algorithm 1: Federated Learning with Ensemble Knowledge Distillation

This algorithm describes the process of federated learning with ensemble knowledge distillation.

STEP-1: It takes as input a proxy IDS dataset (unlabeled) and private IDS datasets (labeled), and initializes client models.

STEP-2: The algorithm then proceeds with communication rounds, where a subset of clients is selected to participate.

STEP-3: For each client, the current global model is synchronized, and a local model is updated using the client's private dataset. The client models are then transmitted to the central server. STEP-4: The server computes an updated consensus by averaging the client models' parameters. STEP-5: Ensemble knowledge distillation is performed, where the server model's logit vectors are calculated based on the proxy dataset, and the client models' logit vectors are calculated based on the proxy dataset using an ensemble algorithm.

STEP-6: The global model is fine-tuned using Kullback-Leibler divergence. The process continues until the desired number of communication rounds is reached, and the trained server model is returned as the output.

A. Heterogeneous distribution of client data

In the context of the document, the term "heterogeneous distribution of client data" refers to the variation and diversity in the data collected from different devices or clients in a federated learning framework. This heterogeneity can arise due to differences in feature dimensions, attack types, and data distribution among the clients. It poses a challenge for centralized machine learning and deep learning models in the context of IoT security detection. To address this challenge, the document proposes an effective heterogeneous federated learning framework using ensemble knowledge distillation (FLEKD). This framework aims to utilize the knowledge from diverse client models to improve the generalization and performance of the server model after aggregation.

B. Parameters Setting

The document provides information about the parameters setting used in the federated learning process. The client number is set to 9, indicating that there are 9 clients participating in the training. According to the client number, every three clients have private data sets of the same dimension. The learning rate is set to 0.001, and a uniform decay strategy of five percent is applied. The temperature for knowledge distillation is set to 1.5, and the deterministic temperature for ensembling weights is set to 0.5. Adam optimization is used, and a decay strategy of five percent is applied for each epoch. These parameters are crucial for the training and optimization.

C. Performance of Proposed Framework

The performance of the proposed framework is evaluated using different sample parameters for the Dirichlet distribution. The table shows the performance of the model using three different values of α : 10, 1, and 0.5. The first part of the table displays the performance of the local training process conducted by each client on their private dataset. As the data becomes more heterogeneous, there is a significant decrease in the F1-score of each client. However, by introducing distributed training, the predictive performance can be improved significantly. For example, in the case of UDPLag, a class with very few examples, an F1-score of 95.40% can be achieved with α set to 0.5. This is a large improvement compared to the results obtained from local training, where some clients achieved less than 50% accuracy. Furthermore, F1-score improves in various other intrusion attack scenarios. Our proposed EKD fine-tuning further enhances the performance of the central model and accelerates the convergence rate.



Figure-2. Error rate of the server-side model on the test set

D. Main Results of Local Training

The main results of the local training conducted by clients 0-8 are presented in Table I. The scores are calculated using the F1-score metric. Each client's performance is evaluated for different intrusion attack scenarios, such as Portmap, LDAP, MSSQL, NetBIOS, Syn, UDP, UDPLag.

TABLE I
MAIN RESULTS. CLIENTS 0-8 CONDUCT LOCAL TRAINING AND ALL THE SCORES ARE CALCULATED BY F1-SCORE.

Local	Portmap			LDAP			MSSQL			NetBIOS			Syn			UDP			UDPLag		
Training	$\alpha=10$	1	0.5	10	1	0.5	10	1	0.5	10	1	0.5	10	1	0.5	10	1	0.5	10	1	0.5
Client 0	99.89%	99.74%	70.08%	99.83%	99.78%	0.00%	99.96%	99.85%	99.97%	99.84%	99.69%	99.54%	99.91%	99.90%	99.92%	99.98%	99.96%	92.73%	97.27%	96.77%	97.99%
Client 1	99.88%	99.89%	99.51%	99.86%	99.80%	96.59%	99.96%	99.97%	95.78%	99.88%	99.83%	99.72%	99.93%	99.94%	98.78%	99.98%	99.92%	99.96%	97.52%	98.42%	0.00%
Client 2	99.92%	99.82%	99.66%	99.94%	99.92%	99.83%	99.93%	99.93%	99.71%	99.94%	99.87%	99.81%	99.93%	99.90%	99.87%	99.99%	100.00%	99.46%	97.30%	97.03%	76.48%
Client 3	99.81%	99.60%	98.96%	99.17%	99.76%	96.76%	99.89%	99.95%	99.92%	99.80%	99.51%	98.70%	99.92%	99.89%	99.93%	99.29%	99.90%	99.94%	97.01%	97.11%	56.63%
Client 4	99.84%	2.30%	64.93%	99.92%	99.86%	99.77%	99.90%	66.47%	94.97%	99.89%	77.55%	0.00%	99.92%	99.90%	99.11%	99.99%	99.86%	99.73%	97.40%	94.12%	67.13%
Client 5	99.83%	99.97%	98.91%	99.90%	99.90%	83.64%	99.93%	99.98%	98.77%	99.89%	99.90%	99.52%	99.87%	99.92%	99.86%	99.96%	99.99%	75.67%	97.03%	97.63%	89.07%
Client 6	90.65%	95.13%	8.94%	99.78%	99.84%	99.29%	80.23%	67.99%	2.42%	93.57%	39.35%	52.28%	99.83%	99.84%	98.95%	99.99%	99.99%	99.91%	93.55%	93.28%	46.65%
Client 7	94.86%	14.50%	32.67%	99.84%	99.89%	99.71%	89.36%	21.53%	13.57%	95.19%	70.78%	53.99%	99.81%	99.90%	99.05%	99.91%	99.99%	99.99%	94.22%	93.05%	54.84%
Client 8	97.19%	91.88%	96.53%	99.82%	99.58%	94.26%	93.11%	84.77%	0.00%	95.94%	95.06%	68.50%	99.89%	99.92%	99.82%	99.94%	99.90%	97.11%	93.31%	95.59%	49.36%
Distributed	Portmap			LDAP			MSSQL			NetBIOS			Syn			UDP			UDPLag		
Training	$\alpha=10$	1	0.5	10	1	0.5	10	1	0.5	10	1	0.5	10	1	0.5	10	1	0.5	10	1	0.5
FL	99.00%	99.45%	99.54%	99.67%	99.64%	99.83%	99.98%	99.86%	99.62%	98.97%	99.57%	99.73%	99.93%	99.93%	99.88%	99.87%	99.78%	99.99%	97.55%	97.21%	95.40%
FLEKD	99.80%	99.73%	99.77%	99.84%	99.70%	99.91%	99.99%	99.98%	99.90%	99.82%	99.73%	99.86%	99.94%	99.93%	99.92%	99.91%	99.88%	99.99%	98.34%	97.64%	96.62%

E. Analysis

In this paper, the effectiveness of the proposed method in the context of an intrusion detection system in IoT security is evaluated. The performance of the method is investigated under three key scenarios: varying dimensions of the data, different sample sizes, and different attack category distributions in various clients. The results demonstrate the superiority of the approach in these scenarios. The impact of different dimensions is also considered, showing that when the available feature vectors are reduced to 24 dimensions, there is a significant decrease in performance. However, when federated learning and the proposed method of ensemble knowledge distillation are applied, there is a significant improvement.

F. The impact of different dimensions

When evaluating the effectiveness of the proposed method in the context of an intrusion detection system in IoT security, the impact of different dimensions of the data was considered. The experiments compared the results of training with datasets of different dimensions locally and with federated learning. It was found that when the available feature vectors were reduced to 24 dimensions, there was a significant decrease in performance. However, when federated learning and the proposed method of ensemble knowledge distillation were applied, there was a significant improvement in accuracy. This indicates that the proposed method can effectively solve the problem of dimensionality differences in IoT device data.

TABLE II
THE IMPACT OF DIFFERENT DIMENSIONS.

	Precision	Recall	F1-score
82 dim	95.93%	96.39%	96.02%
79 dim	86.31%	86.75%	86.17%
24 dim	84.17%	85.48%	84.71%
FL	99.37%	99.36%	99.37%
FLEKD	99.80%	99.80%	99.80%

G. The impact of different sample sizes

In the given document, it is mentioned that there may be significant differences in the number of samples between IoT devices due to the characteristics of network attacks. The dataset was divided into three groups with different sample sizes, represented as 'Base', '10Base', and '100Base'. When using local training only, there was a negative correlation between the number of samples and the detection performance. The best performance of 99.50% was achieved on 100*Base. However, when using FLEKD to aggregate and optimize the models of

different clients, the final F1-score reached 99.80%, which is higher than all results obtained by local training and ordinary FL. This demonstrates that FLEKD can effectively address the sample size differences between different IoT devices and significantly improve intrusion detection performance.

TABLE III
THE IMPACT OF DIFFERENT SAMPLE SIZES.

	Precision	Recall	F1-score
Base	84.17%	85.48%	84.71%
10*Base	92.97%	88.55%	87.25%
100*Base	99.51%	99.50%	99.50%
FL	99.37%	99.36%	99.37%
FLEKD	99.80%	99.80%	99.80%

H. The impact of different attack category distributions

In the experiments conducted using the CICIDS2019 dataset, it was observed that different IoT devices usually have heterogeneous data distributions when it comes to attack categories. Traditional centralized intrusion detection systems (IDSs) often struggle to effectively identify novel attacks. However, the FL-based and FLEKD-based IDS models showed high detection abilities for unknown attacks. By adopting FLEKD, even the most difficult-to-detect attack category, 'UDPLag', achieved a detection rate of 80.86%, while other attack categories saw improvements to around 99%. This demonstrates that FLEKD-based IDS can effectively address the problem of category heterogeneity, enabling different devices to identify attacks.

TABLE IV
THE IMPACT OF DIFFERENT ATTACK CATEGORY DISTRIBUTIONS.

	0: Portmap	1: LDAP	2:MSSQL	3:NetBIOS	4: Syn	5: UDP	6: UDPLag
Drop Label 0	0.00%	98.72%	59.09%	95.62%	99.67%	95.10%	0.00%
Drop Label 1	99.30%	0.00%	66.45%	99.62%	99.91%	99.98%	94.95%
Drop Label 2	66.77%	99.96%	0.00%	99.75%	99.95%	99.99%	96.81%
Drop Label 3	91.06%	97.03%	54.03%	0.00%	99.68%	95.84%	0.00%
Drop Label 4	75.76%	99.69%	94.54%	95.72%	0.00%	99.90%	11.05%
Drop Label 5	99.74%	66.62%	99.78%	99.83%	99.92%	0.00%	97.47%
Drop Label 6	91.21%	97.93%	0.00%	72.35%	99.65%	94.46%	0.00%
FL	99.05%	99.80%	94.37%	97.31%	99.92%	98.85%	75.28%
FLEKD	99.66%	99.88%	98.97%	99.80%	99.94%	99.97%	80.86%

V. CONCLUSION

The document concludes that the proposed FLEKD framework offers a promising solution for IoT networks' security. It highlights the advantages of FL in maintaining data privacy, mitigating data scarcity, and training with heterogeneous data. The proposed FLEKD framework for Intrusion Detection System (IDS) in IoT security outperforms local training and traditional Federated Learning (FL) in terms of speed and performance. It addresses the challenge of data heterogeneity by using knowledge distillation and ensemble methods to improve the generalization and accuracy of the global model. FLEKD demonstrates clear advantages in handling heterogeneous data and shows promising results in real-world scenarios. Overall, FLEKD offers a promising solution for enhancing the security of IoT.

REFERENCES

- [1] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network intrusion detection," IEEE network, vol. 8, no. 3, pp. 26–41, 1994.
- [2] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," Journal of Network and Computer Applications, vol. 36, no. 1, pp. 16–24, 2013.

- [3] B. B. Zarpel^{ao}, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in internet of things," *Journal of Network and Computer Applications*, vol. 84, pp. 25–37, 2017.
- [4] S.-J. Han and S.-B. Cho, "Detecting intrusion with rule-based integration of multiple models," *Computers & Security*, vol. 22, no. 7, pp. 613–623, 2003.
- [5] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," *IEEE communications surveys & tutorials*, vol. 21, no. 1, pp. 686–728, 2018.
- [6] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, p. e4150, 2021.
- [7] S. Agrawal, S. Sarkar, O. Aouedi, G. Yenduri, K. Piamrat, M. Alazab, S. Bhattacharya, P. K. R. Maddikunta, and T. R. Gadekallu, "Federated learning for intrusion detection system: Concepts, challenges and future directions," *Computer Communications*, 2022.
- [8] J. Fan, W. Yang, Z. Liu, J. Kang, D. Niyato, K.-Y. Lam, and H. Du, "Understanding security in smart city domains from the ant-centric perspective," *IEEE Internet of Things Journal*, 2023.
- [9] Z. Ma, L. Liu, W. Meng, X. Luo, L. Wang, and W. Li, "Adcl: Towards an adaptive network intrusion detection system using collaborative learning in iot networks," *IEEE Internet of Things Journal*, 2023.
- [10] W. Yang and K.-Y. Lam, "Effective anomaly detection model training with only unlabeled data by weakly supervised learning techniques," in *ICICS 2021*, 2021, pp. 402–425.
- [11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
- [12] S. I. Popoola, R. Ande, B. Adebisi, G. Gui, M. Hammoudeh, and O. Jo- gunola, "Federated deep learning for zero-day botnet attack detection in iot-edge devices," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3930–3944, 2021.
- [13] Y. Liu, N. Kumar, Z. Xiong, W. Y. B. Lim, J. Kang, and D. Niyato, "Communication-efficient federated learning for anomaly detection in industrial internet of things," in *IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.
- [14] V. Mothukuri, P. Khare, R. M. Parizi, S. Pouriyeh, A. Dehghantanha, and G. Srivastava, "Federated-learning-based anomaly detection for iot security attacks," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2545–2554, 2021.
- [15] B. Li, Y. Wu, J. Song, R. Lu, T. Li, and L. Zhao, "Deepfed: Federated deep learning for intrusion detection in industrial cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5615–5624, 2020.
- [16] P. Ruzafa-Alc^{azar}, P. Fern^{andez}-Saura, E. M^{armol}-Campos, A. Gonz^{alez}-Vidal, J. L. Hern^{andez}-Ramos, J. Bernal-Bernabe, and A. F. Skarmeta, "Intrusion detection based on privacy-preserving federated learning for the industrial iot," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1145–1154, 2021.
- [17] E. M. Campos, P. F. Saura, A. Gonz^{alez}-Vidal, J. L. Hern^{andez}-Ramos, J. B. Bernab^e, G. Baldini, and A. Skarmeta, "Evaluating federated learning for intrusion detection in internet of things: Review and challenges," *Computer Networks*, vol. 203, p. 108661, 2022.
- [18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [19] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.
- [20] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.
- [21] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2351–2363, 2020.
- [22] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni, "Bayesian nonparametric federated learning of neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 7252–7261.