2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS) | 979-8-3503-3071-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICPADS60453.2023.00348

Federated Anomaly Detection with Isolation Forest for IoT Network Traffics

Junyan Li
Macquarie University
junyan.li@hdr.mq.edu.au

Xuyun Zhang
Macquarie University
xuyun.zhang@mq.edu.au

Haolong Xiang
Macquarie University
haolong.xiang@hdr.mq.edu.au

Amin Beheshti
Macquarie University
amin.beheshti@mq.edu.au

***Abstract*—With the development of modern technology, the application of various types of devices in life has become more extensive, especially with the emergence of the Internet of Things (IoT), which makes a large number of devices to be connected to the network. However, while bringing convenience to life, attacks against IoT devices have also begun to appear and become one of the most concerned issues. These IoT devices produce and transmit data containing important user information every second. Attackers are targeting this characteristic to initiate further malicious attacks on these valuable data. In order to protect users from such attacks, it is important that threats can be detected and identified in time before they cause damage. To address this issue, this paper first reviewed the current state-of-the-art anomaly detection methods and based on the finding of uncovered areas of existing methods, this paper proposed a new anomaly detection framework leveraging the combination of Federated Learning and Isolation Forest. The framework performs tree construction on the clients' end and further uploads the encrypted data containing the nodes' information of the trees to the central server for the forest construction, after multiple interactions, the abnormal behaviour in the clients could be able to be identified more effectively and ultimately improved the accuracy of the detection results while protecting the privacy of the client data.**

***Index Terms*—Anomaly Detection, Federated Learning, IoT network, Cybersecurity**

I. Introduction

The rapid development of society in recent years cannot be separated from the use of modern technologies. The use of these technologies covers almost every aspect of people's lives, including industry, medicine, and other different professional fields, with the help of technology, the workload of repetitive work could be helped to reduce, and more efficient on complex work. Nowadays, devices that can connect to the network are no longer limited to traditional cable or cellular connection methods. Through the deployment of wireless networks and their connection with sensors, a large number of devices has now been designed with network connection-enabled sensors, including the most common ones such as mobile phones and laptops, as well as household items including video doorbells and refrigerators, are now all able to be connected to the internet and controlled remotely, these devices are now formed as the Internet of Things (IoT) network [1].

Along with the widespread adoption of IoT technology and further formation of smart environments such as Industrial IoT, Internet of Vehicles, etc., the world has become more connected than ever before, and the presence of IoT is everywhere now. But the rapid growth of IoT devices can be a double-edged sword. Although the adoption of IoT devices can help a lot with automating processes, optimizing resource usage, and improving life quality in a variety of contexts, they could also result in the generation of vast amounts of data during the process. Because these data contain important information that is closely related to the user, therefore, as the number of IoT devices increases, so does the risk of security breaches and cyber-attacks. Therefore, it is important to protect the privacy of users as well as the security of the network, to achieve this, an effective method for detecting and preventing abnormal behavior in the IoT network needs to be proposed.

Anomaly detection is one of the most used methods to identify deviations from a regular pattern of behavior. It has been widely applied in various sectors, including banking, medical monitoring, and especially security attack identification. And now, it has been found that it could provide great use to identify the anomalies in the IoT network. Anomalies in the IoT network could indicate the presence of threats, such as malicious intrusions in a network. There are many researchers and partitioners who have developed different anomaly detection techniques to protect the security of networks, including but not limited to the use of deep learning, statistical methods, machine learning algorithms, etc. However, while significant progress has been made in developing effective anomaly detection methods for network security purposes, there are still many uncovered areas that need to be addressed. For example, in the IoT scenarios, they tend to generate large-scale high dimensional data, which could include sensor reading, network traffic captured hardware information, etc. which could be relatively more difficult to use the traditional methods for conducting complex analysis to further identify the anomalies in time.

Despite these challenges, a method called Isolation Forest was proposed by Zhou, unlike the traditional methods such as Random Forest and Local Outlier Factor (LOF) etc., it outperformed them in the merits of the processing time and AUC score of large-scale datasets [2]. Therefore, it has become the priority choice to conduct anomaly detection for recent research. In IoT networks, it is important to avoid exposing data to parties other than the data owner due to the highly sensitive nature of the data generated and processed by IoT devices on a daily basis. But, when the manufacturers or organization need to perform further analysis or customization

2690-5965/23/$31.00 ©2023 IEEE
DOI 10.1109/ICPADS60453.2023.00348

would require a certain amount of data for training, even though a vast number of devices are already connected to the internet, they are mostly owned by individuals or service providers, which in most of the time would be grudging to share such information due to the potential possibility of information disclosure or data privacy agreements, etc. [3], [4].

This has become one of the most important problems in the development of anomaly detection that needs to be tackled, namely, how to obtain data from multiple parties for better training and more accurate training results while protecting the privacy and security of local data. To address these issues, this work proposed a novel approach, in addition to the existing Isolation Forest method, we would also like to combine the use of Federated Learning. The proposed approach inherits the advantages of both methods to construct as an unsupervised tree structure anomalies detection model.

The main contribution of this work could be concluded into three parts:

- Firstly, we investigated the methodological gap for anomaly detection in IoT network due to lack of valid data and proposed a more comprehensive federated architecture by combining the two different algorithms and architectures. On top of it, with the aggregation of unsupervised learning algorithms of iForest and Federated Learning, our approach could greatly improve the accuracy of the detection result.
- Secondly, with the use of our model, each participant only has to share a portion of the local model parameters to other participants rather than sharing all the raw data, we could effectively protect the privacy of the participants in the processes of anomaly detection. In addition, we introduced differential privacy to enhance data privacy protection, ensuring that the model does not leak sensitive information during data transmission, and guaranteeing the privacy and security of the source data and data owners.
- Moreover, in our experiments, we used multiple IoT network datasets such as IoT23 and IIoTset to demonstrate the stable performance of our method across different data distributions and attack scenarios. While using AUC and F1 scores as evaluation metrics, we demonstrated that our method performed well in terms of both accuracy and sensitivity.

The rest of this paper is organized as follows. Section 2 provides a review of the relevant works on anomaly detection and the application of Federated Learning in different network, and section 3 and 4 describes our motivations and experimental methodology respectively, including datasets, evaluation metrics. Finally, section 5 concludes the paper and proposes future research directions in this field.

II. Literature Review

At current stage, anomaly detection in IoT network faces several challenges. Firstly, is the scarcity of the data, most data owners including individual and organization who owns the data, would prefer not to share with other third parties because of privacy concerns, therefore the acquisition of data, especially ones with labelled anomaly data are extremely difficult. Secondly, is about the definition of the anomaly, depends on the specific context, the anomaly is the variable that differs along with the changes of time, environment, and other factors as well. So, the detection method needs to be adaptable and generalizable [5]. To overcome these challenges, researchers have developed a variety of anomaly detection methods.

A. State-of-the-art Anomaly Detection Methods

In the recent research of anomaly detection methods proposed, they can mostly fell into the category of either traditional or deep learning-based algorithms. When referring to the traditional algorithm, we usually include methods based on statistical analysis, and pre-defined models to define and identify outliers. Such as analytical models using Gaussian distributions [6], and Z-score [7] etc., anomalies are considered to be different from most of the same occurrences or distributions. Although the traditional approach is still the most widely used and popular method today, it does have some drawbacks, such as limitations when detecting complex anomalies or identification in large amounts of high-dimensional data.

Machine learning based methods as a sub-type of the traditional anomaly detection methods, have recently become the primary preference when developing anomaly detection algorithms. They have the ability to find useful information and patterns in the process of learning from large amounts of data, and the modes of learning can be further divided into supervised, semi-supervised, and unsupervised learning [8]. This capability is one of the reasons why it has become such a popular choice.

Meanwhile, there is deep anomaly detection (DAD) techniques, are anomaly detection based on the deep learning. In the past few years, deep learning has achieved remarkable success in various fields, including computer vision and natural language processing, as well as in combination with other techniques [9]. The application of deep learning in anomaly detection has focused on processing complex time-series data and high-dimensional features. As previously introduced, traditional anomaly detection methods are often based on statistical methods and rules, but for complex time-series data and high-dimensional features, these methods may not be able to capture anomaly patterns properly [10]. The advantage of deep learning is that it can automatically learn complex feature representations from the data to better identify anomalies. Although deep learning performs well in anomaly detection, there are some limitations and challenges. For example, deep learning models are usually more complex than traditional methods, requiring more computational resources and time to train. Besides, deep learning models are less explanatory, making it difficult to explain why a particular sample is classified as anomalous.

It is noting that the highly regarded detection methods, the tree-based anomaly detection methods, are now being widely used and studied. Different anomaly detection methods based on tree structure have been proposed and the random forest-based detection method is one of the most popular choices. Random forest is an ensemble learning method [11] that performs anomaly detection by constructing multiple decision trees. Each tree is constructed by randomly selecting features and samples, and the final anomaly score can be obtained either by voting or averaging [12]. This ensemble technique increases the model's capacity to recognise intricate patterns in the data and strengthens its resistance to noise or outliers [13]. This is also a common advantage that most tree models have, the unique advantages of Random Forests include that they are not easily overfitted and have good scalability. On the other hand, its disadvantages would be the high requirements for computational resources and storage space [14].

In addition to the previously mentioned tree structures, another type focused on this article is the Isolation Forest. The Isolation Forest utilizes the binary trees and randomization to provide an approach to perform a faster and more robustness detection for anomalies [15]. The state-of-the-art research are now developing the iForest based approaches to conduct detection for varies purposes. Such as in [16], it proposed a more efficient approach for detecting data anomalies based on Isolation Forests with the goal of reducing the calculation time of the current existing Extended Isolation Forest (EIF) [17]. In terms of real-world applications, the study in [18] utilized iForest for the detection of anomalies in data of power consumption. In addition, [19] proposed an ensemble-based approach to improve the accuracy of iForest in detecting anomalies using the set of nearest neighbours. In [20], the researcher focused on the enhancement of anomaly scores in the isolation forest. There are also many proven reliable methods that have been investigated in multiple aspects and extensions, such as [21], [22]. These methods solve the challenges of slow computation, and show good scalability and robustness when dealing with datasets of different sizes for different domains.

B. Federated Learning

Current research trends in Federated Learning on detection methods for anomalies in networks include the Internet of Things network environment [23]. In the research of [24], the authors focused on Federated Learning for anomaly threat detection in multi-task learning across different network traffic classes and sources. The experiments have proved to a significant reduction in training time expenses through the collaborative use of multi-tasking models with deep neural networks. The research in [25] is dedicated to the use of federated learning techniques to detect network traffic anomalies on the Internet of Things network environment. The authors believed that using the nature of IoT malware infections, where only a limited number of IoT networks contain infected devices, could mitigate the impact of anomalies on the overall data training. This paper focused on how to mitigate the contamination of anomaly samples and the effectiveness of the proposed approach is demonstrated by simulation evaluation with real network traffic data. Article [26], presents a self-learning anomaly detection system designed for IoT devices. This system utilize a novel anomaly detection method based on device type-specific communication profiles without the need for human intervention or labelling data. It employs a joint learning approach to effectively aggregate behavioural features, enabling it to detect emerging unknown attacks. In [27] the technology is based on Federated Stacked Long Short-Time Memory (LSTM) model for anomaly detection in smart buildings using IoT sensor data. The combined of Federated Learning method allows the proposed model outperforms centralised approaches, converging twice as fast and reducing communication costs. And in [28], the study explored the application of federated learning in the detection of vehicle trajectory anomalies.

In summary, the technology of Federated Learning when combined with different methods can be better used for effective anomaly detection within a number of fields including network intrusion, IoT network traffic etc. The combined use of its distributed data and privacy-preserving mechanisms makes it one of the most popular anomaly detection methods today, and through proper exploration, a more secure, and efficient detection method could be developed and applied in different applications.

III. Methodology

To develop a robust anomaly detection method for IoT networks, it is crucial to acknowledge that the current challenge in detecting anomalies stems primarily from the scarcity of both labeled and unlabeled data. Most available data is outdated or publicly accessible, impeding the development of protective measures. Even with this limited amount of data, only a small portion pertains to the IoT network concept, resulting in even fewer datasets sharing similar attributes as IoT networks. Additionally, due to their low computational power, most IoT devices cannot perform complex calculations or protection locally and are thus vulnerable to attacks that could compromise user privacy.

Obtaining more data requires acquiring from more sources, and the best approach is by resources sharing with trustworthy partners. But this brings us to another aspect of the problem, which is the risk of data sharing. The first is the privacy of data, which can reveal a lot of information that is closely related to the data owner, such as the personal information of patients in hospitals, the information of IoT device users, including voice prints and stored account information, etc.; and even the user's trajectory on the Internet of Vehicles (IoV). Once those raw data are exploited by attackers with malicious intentions, it can be extremely harmful to the data owner in numerous aspects. Furthermore, even if some data owners agree to share their data with trusted partners, the question of whether there are risks involved in the sharing process is another issue that needs to be addressed.

To address the aforementioned challenges and considerations, we introduce an ensemble model that integrates Isolation Forest [2] and Federated Learning [23]. This approach leverages both techniques to facilitate anomaly detection within dis-

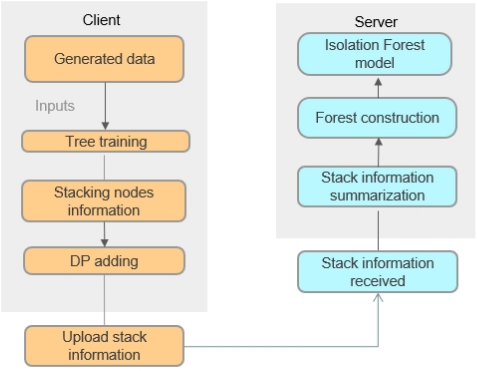

Fig. 1. Overview of Isolation Forest Federation

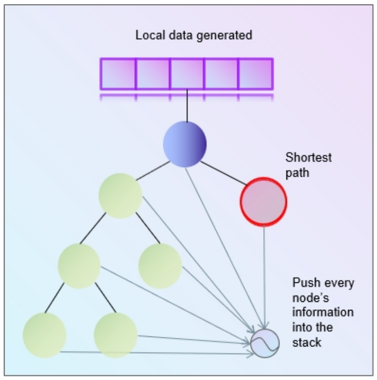

Fig. 2. Isolation Tree Construction

tributed environments. A robust and secure anomaly detection method is proposed using the inherent anomaly detection capability of Isolation Forest and the privacy-preserving properties of Federated Learning. The method provides an efficient and privacy-sensitive solution for anomaly detection in distributed environments.

The figure 1 is showing the overview of the model, the process starts at the client side, this model is composed of three main parts, the isolation forest mainly contributes in the structure of anomaly detection and federated learning provides distributed privacy protection on top of it. The first part is the individual training on the client side, that is, the construction of the tree in the Isolation Forest; Then, the encryption with differential privacy is carried out after the tree construction is finished, followed by the construction of the forest on the server side by collecting the encrypted data from multiple clients to collect the node data from different sources. We will describe each part in more detail in the following sections.

A. *Isolation Tree Construction*

Firstly, we consider that in an IoT network, each client (e.g., c_1, c_2 to c_n) is connected to a series of devices that generate or transmit a large amount of data related to the network. These data may come from a variety of sensing, measuring, and monitoring devices, including temperature, pictures, videos, and various other types. In order to perform anomaly detection without sacrificing data privacy, we first emphasise the importance of local storage. Specifically, each client stores the network data it generates locally, forming a distributed data collection D. This local storage ensures that the raw data is not compromised, setting the foundation for future processing.

And then, based on the distributed data collection D, we introduce the Isolation Forest algorithm as our anomaly detection tool. Different from traditional anomaly detection methods, Isolation Forest fully considers the multidimensional information of the data when constructing the isolation tree, which is compatible with the complex and diverse data characteristics in IoT networks. For the local data of each client, we carry out the construction of Isolation tree separately. In this process, we not only focus on the attributes of the data, but also record the detailed information of each node.

These include the node ID, number of dimensions i, maximum tree depth d, as well as the minimum and maximum values of the node. The recording of this information provides a rich representation of the nodes of the isolation tree and also helps to ensure that the generated isolation tree is able to identify anomalous data points more accurately in the future. The information about all the nodes in this tree is then stored through the use of stacks, which are used to allow the data to be transferred to the server in a systematic manner for better classification and computation using the node IDs as shown in Figure 2 .

The main goal of this process is to generate isolation tree's structure more accurately. Due to the diversity and complexity of data in IoT networks, it is often difficult for traditional methods to effectively capture anomalies in the data. While our approach improves the performance of anomaly detection by recording rich node information, which enables the isolation tree to better provide the characteristics of different data distributions.

We then introduce a differential privacy mechanism in our approach to further enhance the layer of data privacy protection while ensuring privacy. In our method, we use a specific implementation of differential privacy that achieves privacy protection by adding Laplace noise to the node information. This approach provides a balance between privacy and data utility, allowing us to protect the privacy of the data while still being able to extract useful information from the data.

B. *Data Federation and Forest Construction*

After the local node information within the stack encrypted with differential privacy is transmitted to the central server. The server in the model act as the core of the computation process, and in a Federated Learning framework, parameter aggregation is carried out after multiple parameters are uploaded to the server; in our model, due to the unique characteristics of the tree structure, the client, instead of acting as a mere

Algorithm 1 Client-side Federated Isolation Tree with Differential Privacy

```
Require: Local data $D$, number of dimensions $i$, maximum tree depth $d$, privacy parameter $\varepsilon$
Ensure: Local node information stack $\text{Stack}_i$
  Initialize local node information stack $\text{Stack}_i \leftarrow \emptyset$
  Build initial node $N_i \leftarrow \text{Node}(D_i)$
  Push initial node onto stack $\text{Stack}_i \leftarrow \text{Stack}_i \cup N_i$
  for round $= 1$ to $T$                    Number of rounds do
    for each node $N_i$ in $\text{Stack}_i$ do
      if Depth of $N_i$ $d_{N_i} < d$ then
        Randomly choose dimension $i$ ($i \in [1, m]$) to split in $N_i$
        Generate child nodes $N_{\text{left}}$ and $N_{\text{right}}$
        Add the split threshold of $N_i$ and child nodes
        Push $N_{\text{left}}$ and $N_{\text{right}}$ onto stack $\text{Stack}_i \leftarrow \text{Stack}_i \cup \{N_{\text{left}}, N_{\text{right}}\}$
      end if
    end for
  end for
  Apply differential privacy parameter $\varepsilon$ to $\text{Stack}_i$
  Upload $\text{Stack}_i$ to server
  for each new data instance $x$ do
    Partition $x$ using the updated local model
    Determine if $x$ is an anomaly based on partition and local model
  end for
  Return Local node information stack $\text{Stack}_i$
```

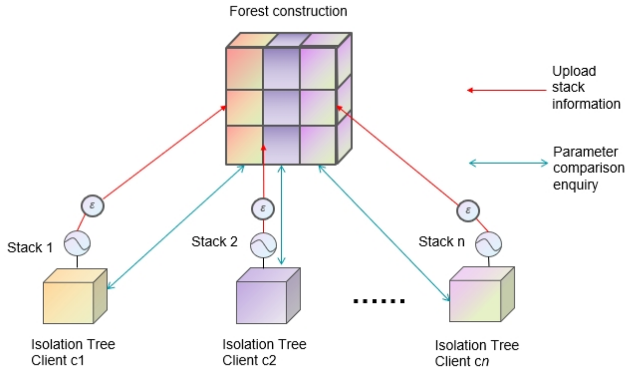

Fig. 3. Isolation Forest Federation

data transmitter, act as a tree structure that compose the forest construction. We perform the aggregation of information with the same node IDs in each tree and compute their maximum and minimum ranges. The purpose of this step is to extend the representation of the tree nodes by mapping nodes that would otherwise have only local information to a more global range of features, thus enhancing the generalisation of the model. In this way, we introduce a layer of globalised feature extraction into the scheme of the method's flow, integrating the information dispersed in different clients' nodes into a more comprehensive perspective. This global feature extraction allows the model to better capture anomaly patterns across nodes in subsequent anomaly detection tasks, thus improving the overall detection performance.

On this stage it can be seen as an efficient way of data co-operation, where each client contributes to the progress of the overall model without having to expose its individual data as shown in Figure 3. This approach not only offers advantages in terms of data privacy protection, but also provides some savings in terms of computational and communication costs. Through Federated Learning, we achieve collaborative modelling with multiple participants, which guarantees data privacy and takes full advantage of distributed data. The Federated Learning paradigm plays a key role in our approach. It enables collaboration between clients while protecting data privacy. In this phase, the data from each client will securely reach the central server for subsequent training. Importantly, the raw data never leaves the client environment. This decentralised approach ensures that sensitive information remains localised, while the global model benefits from the collective insights of the different datasets. Advanced aggregation techniques such as secure aggregation and homomorphic encryption ensure that the global model accurately reflects anomaly detection across the network.

In the final phase, the extended range of trees' nodes representation information would be passed back to individual clients, and this information will play a crucial role in future tasks, especially when it comes to operations such as data partitioning and segmentation. This phase marks the completion of a complete round of isolation forest construction. By continuously updating the tree nodes' range information at the local client, we provide a solid foundation for future actions. These tasks may include partitioning, segmenting, or further distributed analysis of the data. The retention of this information means that we are able to respond to different data distributions and characteristics with more flexibility and precision in future data manipulations.

IV. Experiments

This experimental section covers the dataset we have used, evaluation metrics, experiment design, and a comparison of our results with baseline models. Our experiments demonstrate outstanding results, proving the effectiveness and robustness of our approach.

A. Datasets

We are using following publicly available real-world datasets, which are all closely related and most used for the IoT network and network anomaly detections.

- KDD99[1]: The dataset is mainly used for network intrusion detection, which contains labelled network traffic data, including the four types of attacks as well as normal network traffic.

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Algorithm 2 Server-side Aggregation of Federated Isolation Forest

```
Require: Node information stacks $\text{Stack}_{\text{agg}}$ from clients
Ensure: Range information $\text{Range}_{\text{agg}}$ for each node ID
  Initialize range information $\text{Range}_{\text{agg}} \leftarrow \emptyset$
  for each client's stack $S$ in $\text{Stack}_{\text{agg}}$ do
    for each node information $(\text{ID}, \min, \max)$ in $S$ do
      if ID not in $\text{Range}_{\text{agg}}$ then
        $\text{Range}_{\text{agg}}[\text{ID}] \leftarrow (\min, \max)$
      else
        Update range information for ID with $(\min, \max)$
      end if
    end for
  end for
  for each client's stack $S$ in $\text{Stack}_{\text{agg}}$ do
    for each node information $(\text{ID}, \min, \max)$ in $S$ do
      Send $\text{Range}_{\text{agg}}[\text{ID}]$ to client
    end for
  end for
```

- IoT23[2]: In the IoT23 dataset contains normal and abnormal data from different types of IoT devices (e.g., IP cameras, smart sockets etc.). In the IoT23 dataset, we need to classify the normal and abnormal device behaviour by the characteristics of the devices such as sensor readings, network connection status, packet size, etc.
- UNSW NB15[3]: This is a widely used network intrusion detection dataset containing a variety of network traffic data as well as multiple attack types, which is suitable for intrusion detection research in IoT environments.
- Edge IIoTset[4]: Edge-IIoTset is an edge and industrial Internet of Things (IIoT) device-specific real-world dataset, it includes sensor data, log data, and network data from various IIoT devices and applications.

TABLE I
DATASET DESCRIPTION

Dataset	***Nodes***	***Records***	***Features***	***Type of data***
KDD99	4940	5,00,000	41	4
IoT23	23,710	Unknown	30	20
UNSW-NB15	255,573	700,000	49	9
Edge-IIoTset	Huge	Huge	1176	14

B. Evaluation Metrics

In classification problems, the confusion matrix [29] is one of the most important measurements. The determination of an event or item based on a certain criterion, by getting a series of positive or negative responses, can provide us with a clear and correct understanding of the definition of the item or event. In the process of determining a binary classification problem, we are inevitably getting one of the four results in the confusion matrix, including True Positive (TP), False Positive (FP), False Positive (FP), True Positive (TP) as shown in the Table II.

By further calculating these values, we can determine whether a node is abnormal or not at a certain point in the tree structure generated from our data. The following metrics are calculated: Accuracy, Precision, Recall, etc.

TABLE II
CONFUSION MATRIX

		Actual class	
		Positive class	Negative Class
Predicted class	Positive class	True Positive(TP)	False Positive(FP)
	Negative class	False Positive(FP)	True Positive(TP)

The accuracy is representing the correctly predicted rate among all data sample. Indicates the level of reliability that the prediction is correct

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \tag{1}$$

Which in our model:

$$Accuracy = \frac{Number of Correctly Predicted Anomalies}{Number of Predictions} \tag{2}$$

Where a True Positive Rate (TPR) is the proportion of positive classes predicted correctly over the actual positive classes, and a False Positive Rate (FPR) the proportion of the positive class with the wrong prediction to the actual negative class.

$$TPR = \frac{|TP|}{|TP| + |FN|} \tag{3}$$

$$FPR = \frac{|FP|}{|TN| + |FP|} \tag{4}$$

Where the Precision is the result of correctly predicted positive cases are also true positive.

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{5}$$

In the experiments AUC and F1 scores will also be used as a criterion for anomaly detection. Because AUC is a measure of how good the model is at identifying anomalies. It is the area under the ROC curve, and the ROC curve is a plot of TP and TN rate ratio. The higher the AUC value, the better the model's ability to distinguish between normal and abnormal instances.

$$AUC = \frac{1}{2} \times \Sigma_{i=1}^{m-1}(x'_{i+1} - x'_i)(y'_i + y'_{i+1}) \tag{6}$$

And, the F1 score is a measure of how accurate the model is in anomalies identification, and it is a weighted average of precision and recall. The higher the F1 score, the more accurate the model is.

$$F1score = 2 \times \frac{Precision \times Recall}{Presion + Recall} \tag{7}$$

[2] https://www.stratosphereips.org/datasets-iot23

[3] https://research.unsw.edu.au/projects/unsw-nb15-dataset

[4] https://www.kaggle.com/datasets/mohamedamineferrag/edgeiiotset-cyber-security-dataset-of-iot-iiot

C. Experiments Evaluation

In this section, we present the experimental results of applying the proposed methodology, which combines Isolation Forest and Federated Learning to compare with the traditional anomaly detection method Local Outlier Factor (LOF) and Isolation Forest it self on different dataset described in the previous sections. The primary goal of our experiments was to evaluate the effectiveness and performance of our approach in comparison to traditional anomaly detection methods.

The use of all datasets, they would required to be pre-processed with standardization, all of the attack types produced anomalies would be labelled with '-1', and the rest normal data instances would be labelled with '1'. And for some datasets, due to the huge number of irrelevant data, we manually selected specific attack types for experiments, therefore the size of dataset and results could be varied. We also evaluated the computation times required for training and testing of each method.

Firstly, we used datasets most related with the IoT network, the Edge IIoTset Cyber Security Dataset of IoT and industrial IoT (IIoT), which contains data generated by more than ten IoT and IIoT devices, e.g., heart-rate sensors, digital sensors, etc., under 14 cyber-attacks . As well as IoT23, all of 23 scenarios which containing real-world generated data from devices such as a Philips smart lamp, an Amazon Echo and smart doorlock etc.

From the test results in these two datasets, LOF performs poorly in the IoT scenarios, compared to the other two methods it scores only 51-56% of other two AUC scores, and its training time is much longer; meanwhile, with the EdgeIIoT datasets, iForest received great result on AUC with 0.96864, and our method scores only 0.06 lower compared to iForest, and the lengths of the training and testing are very similar, differing only by 0.3 in IoT23 :

TABLE III
EDGEIIOT DATASET

	AUC	Training Time	Testing Time
Isolation Forest	0.9864	0.1245	0.3048
LOF	0.5179	0.0513	0.0119
Federated iForest	0.9185	0.2139	0.3212

TABLE IV
IOT23

	AUC	Training Time	Testing Time
Isolation Forest	0.8400	0.3953	2.0116
LOF	0.4544	6.43606	1.625
Federated iFores	0.8122	0.4855	2.2349

While in the experiments of UNSW-NB15, we selected network data generated under DoS and Worms attacks, two of the most popular attack methods in current IoT network attacks. Similar to the results of the previous two sets of tests, LOF still only gets a low AUC score in these two experiments, but has the shortest training and testing time.

TABLE V
NB15 DOS

	AUC	Training Time	Testing Time
Isolation Forest	0.9937	0.1155	0.3768
LOF	0.4910	0.0788	0.0200
Federated iForest	0.9672	0.1200	0.3612

TABLE VI
NB15 WORMS

	AUC	Training Time	Testing Time
Isolation Forest	0.999	0.0621	0.0327
LOF	0.3288	0.0044	0.0010
Federated iForest	0.9801	0.0700	0.0331

The last two sets of experiments, one was used data generated under the smurf attack from KDD99 dataset. One used data named glass[5], which is a labelled dataset containing the physical properties of 70 different types of glass, which is widely used in machine learning studies for classification, clustering, and regression, The purpose of using this dataset is to explore whether our approach is also applicable in other domains. The results of the experiments have the same conclusions as the previous experiments, in terms of scores, iForest still has the best AUC scores, LOF scores are only 52-54% of the other two methods, and even in the experiment with dataset glass' test only reached 36%.

TABLE VII
KDD99 SMURF

	AUC	Training Time	Testing Time
Isolation Forest	0.9463	0.6344	1.3001
LOF	0.4988	0.7793	0.3214
Federated iForest	0.9102	0.7521	1.823

TABLE VIII
GLASS DATASET

	AUC	Training Time	Testing Time
Isolation Forest	0.7029	0.0600	0.0209
LOF	0.2500	0.0010	0.0010
Federated iForest	0.6995	0.0710	0.0301

The AUC score can effectively measure the model's ability of anomaly instances identify.Our method obtained high AUC scores in multiple datasets, second only to iForest, whereas the experimental results of LOF show that it is not an ideally suited method for anomaly detection in IoT networks. Compared to the iForest and LOF methods, we have achieved both maintaining efficient detection performance, and at the same time can obtain data from more collaborators while protecting the privacy of users and the security of data. It is an ideal approach for IoT network anomaly detection.

V. CONCLUSION AND FUTURE WORKS

In this paper, we proposed an ensemble method that combined Isolation Forest and Federated Learning for efficient

[5] https://archive.ics.uci.edu/dataset/42/glass+identification

and privacy-preserving anomaly detection in distributed IoT network environments.

Our approach leverages the powerful anomaly detection ability of Isolation Forest while ensuring data privacy through the use of Federated Learning. By conducting a comprehensive review of existing literature, we identified gaps and proposed a novel anomaly detection method for enhancing IoT network security. Detailed methodology descriptions and experimental validations demonstrated the effectiveness and feasibility of our approach across multiple datasets. In our experiments, we used IoT network datasets to demonstrate the stable performance of our method across different data distributions and attack scenarios. While using AUC and F1 scores as evaluation metrics, we demonstrated that our method performed well in terms of both accuracy and sensitivity.

Although this paper has made significant progress in combining Isolation Forest and Federated Learning, there are still some potential areas for further exploring. We can explore different parameter settings, model architectures and algorithms to further optimize our approach and improve its performance. In particular, we can experiment with more complex models and more sophisticated algorithmic uses. For example, replacing with more innovative techniques, such as LSHiForest [21], some structural enhancements including parameter uploading on each individual node, it could allow for a more flexible management of information sharing and model updating on each node by using an iterative approach, which may be able to progressively optimize the global model by integrating information from nodes in each iteration, and gradually improving the performance and robustness of the model. Improvement of privacy-preserving methods: although we introduced Laplace to protect data privacy, some other method substitutions such as Exponential may provide a better privacy-preserving technique to improve the model's level of protection in terms of data privacy [30].

In summary, our paper presented a novel solution for privacy-preserving anomaly detection in distributed environments. Future research can build upon these findings to meet growing cybersecurity needs.

References

[1] Andrew A. Cook, Göksel Mısırlı, and Zhong Fan. Anomaly detection for iot time-series data: A survey. *IEEE Internet of Things Journal*, 7:6481–6494, 07 2020.

[2] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 12 2008.

[3] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37:50–60, 05 2020.

[4] Christian Koetsier, Jelena Fiosina, Jan N. Gremmel, Jörg P. Müller, David M. Woisetschläger, and Monika Sester. Detection of anomalous vehicle trajectories using federated learning. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 4:100013, 04 2022.

[5] Tao Wang, Bo Zhao, and Liming Fang. Flforest: Byzantine-robust federated learning through isolated forest. *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*, 01 2023.

[6] Peter J. Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1:73–79, 01 2011.

[7] S Haider, Abul K Abbas, and Aijaz A Zaidi. A multi-technique approach for user identification through keystroke dynamics. *Systems, Man and Cybernetics*, 10 2000.

[8] Abir Smiti. A critical overview of outlier detection methods. *Computer Science Review*, 38:100306, 11 2020.

[9] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv:1901.03407 [cs, stat]*, 01 2019.

[10] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection. *ACM Computing Surveys*, 54:1–38, 03 2021.

[11] Steven J. Rigatti. Random forest. *Journal of Insurance Medicine*, 47:31–39, 01 2017.

[12] Jiong Zhang, M Zulkernine, and A Haque. Random-forests-based network intrusion detection systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38:649–659, 09 2008.

[13] Zhiruo Zhao, Kishan G Mehrotra, and Chilukuri K Mohan. Online anomaly detection using random forest. pages 135–147, 01 2018.

[14] Priyajit Biswas and Tuhina Samanta. Anomaly detection using ensemble random forest in wireless sensor network. 13:2043–2052, 06 2021.

[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 6:1–39, 03 2012.

[16] Julien Lesouple, Cédric Baudoin, Marc Spigai, and Jean-Yves Tourneret. Generalized isolation forest for anomaly detection. *Pattern Recognition Letters*, 06 2021.

[17] Sahand Hariri, Matias Carrasco Kind, and Robert J. Brunner. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, 33:1479–1489, 04 2021.

[18] Wei Mao, Xiu Cao, Qinhua Zhou, Tong Yan, and Yong-Kang Zhang. Anomaly detection for power consumption data based on isolated forest. *2018 International Conference on Power System Technology (POWERCON)*, 11 2018.

[19] Dong Xu, Yanjun Wang, Yulong Meng, and Ziying Zhang. An improved data anomaly detection method based on isolation forest, 12 2017.

[20] Antonella Mensi and Manuele Bicego. Enhanced anomaly scores for isolation forests. *Pattern Recognition*, 120:108115, 12 2021.

[21] Xuyun Zhang, Wanchun Dou, Qiang He, Rui Zhou, Christopher Leckie, Ramamohanarao Kotagiri, and Zoran Salcic. Lshiforest: A generic framework for fast tree isolation based ensemble anomaly analysis, 04 2017.

[22] Haolong Xiang, Zoran Salcic, Wanchun Dou, Xiaolong Xu, Lianyong Qi, and Xuyun Zhang. Ophiforest: Order preserving hashing based isolation forest for robust and scalable anomaly detection. 10 2020.

[23] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning. *ACM Transactions on Intelligent Systems and Technology*, 10:1–19, 02 2019.

[24] Ying Zhao, Junjun Chen, Di Wu, Jian Teng, and Shui Yu. Multi-task network anomaly detection using federated learning. *Proceedings of the Tenth International Symposium on Information and Communication Technology - SoICT 2019*, 2019.

[25] Takayuki Nishio, Masataka Nakahara, Norihiro Okui, Ayumu Kubota, Yasuaki Kobayashi, Keizo Sugiyama, and Ryoichi Shinkuma. Anomaly traffic detection with federated learning toward network-based malware detection in iot. 12 2022.

[26] Thien Huu Nguyen, Samuel Marchal, Markus Miettinen, Hossein Fereidooni, Nadarajah Asokan, and Ahmad-Reza Sadeghi. Dïot: A federated self-learning anomaly detection system for iot. 04 2018.

[27] Raed Abdel Sater and A. Ben Hamza. A federated learning approach to anomaly detection in smart buildings. *ACM Transactions on Internet of Things*, 2:1–23, 11 2021.

[28] C Koetsier, Jelena Fiosina, Jan N Gremmel, Monika Sester, Jörg Müller, and David M Woisetschläger. Federated cooperative detection of anomalous vehicle trajectories at intersections. 11 2021.

[29] Nilesh Kumar Sahu and Indrajit Mukherjee. Machine learning based anomaly detection for iot network: (anomaly detection in iot network), 06 2020.

[30] Victor , Felipe T Brito, Cheryl J Flynn, Javam C Machado, Subhabrata Majumdar, and Divesh Srivastava. Local dampening: differential privacy for non-numeric queries via local sensitivity. *The Vldb Journal*, 01 2023.