Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

A Hybrid Methodology for Detecting Anomalous Patterns of Imbalanced Data in Distributed Systems

Enas Hakim Mohsin[1]
Ministry of Higher Education and Scientific Research – Iraq/Al-Furat Al-Awsat University/
College of Administration and Economics,
enass.mohsen.cku@atu.edu.iq

Abstract. *Securing distributed infrastructure is a highly complex challenge due to the sheer volume of data, the diversity of its sources, and the severe imbalances in its categories, which render traditional anomaly detection systems ineffective [1, 7]. This work presents a hybrid approach combining privacy-preserving unified learning capabilities with local data balancing technology (SMOTE). This framework is designed to enable collaborative training across endpoints without requiring centralized raw data exchange, thus striking a balance between privacy and reliability [9]. By applying heterogeneous data distribution processing locally at each site before global aggregation, this approach aims to minimize false positives and increase the likelihood of detecting attacks or rare anomalies [5, 12]. This approach will help provide scalable, proactive protection for existing cloud computing and distributed networks [2, 8].*

Keywords: *Anomaly Detection, Federated Learning, Imbalanced Data, SMOTE Oversampling, Distributed Systems, Cyber security.*

1. INTRODUCTION

Today's distributed network infrastructure is subject to higher cybersecurity requirements due to the complexity of data sources and attacks that compromise the reliability of traditional anomaly detection systems [1, 14]. While modern solutions increasingly utilize machine learning and deep learning methodologies to analyze network traffic, the practical application of Network Intrusion Detection Systems (NIDS) faces two major obstacles: uneven data distribution and user privacy concerns [2, 9].

The significant data imbalance biases machine learning models toward the most common categories (normal activities), significantly weakening their ability to detect rare attacks or anomalies with serious consequences [4, 12]. Furthermore, another security issue arises from the sensitivity of network data; organizations are reluctant to share raw traffic logs due to privacy policy concerns, hindering the development of models generalizable across different environments [9].

Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

Overcoming these limitations, federated learning has emerged as an innovative decentralized model, allowing devices or endpoints to collaboratively train models from local data without the need to exchange or share sensitive central information [8,10]. This research builds upon these gains to offer a hybrid framework that not only relies on decentralization but also fundamentally addresses the problem of class imbalance locally to enhance operational efficiency and predictive power in distributed systems [5, 10].

2. MATERIALS AND METHODS

2.1.Problem Statement:

Most anomaly detection algorithms rely on the rarity of anomalous behaviors and their substantial deviation from normal data [14]. However, in distributed and complex systems, this extreme variability leads to a decline in algorithm accuracy. The overall accuracy metric is not a reliable benchmark in this context, as it assigns equal weight to highly variable categories [7, 15]. Therefore, the problem lies in the lack of a hybrid mechanism in current systems capable of balancing highly localized data at peripheral nodes in a way that does not negatively impact centralized aggregation in federated learning environments [2, 10].

2.2 Research Objectives:

1.To design a hybrid framework based on unified learning to enable accurate detection of anomalies without compromising data privacy [3, 9].

2.To integrate synthetic data generation techniques for processing rare classes (such as SMOTE) at the terminal node level to ensure model balance before integration [4, 11].

3.To measure and evaluate the performance of the proposed framework against centralized and decentralized models using precise metrics such as recall and F1-score [7, 15].

2.3 Scientific Contribution:

The primary contribution of this study is the development of an architecture that transcends the limitations of traditional unified learning by integrating the preprocessing of unbalanced data within each terminal independently [5]. In this way the local models can learn concurrently across the terminals, and the statistical biases is eliminated [12].

The last few years have witnessed a growth of research on the applicability of FL in the context of distributed systems [8]. This recent work reports a hybrid dis- covery method based on local learning and global aggregation, called Nonlocal framework [3]. Local models learn environment-specific patterns, and a global model combines this knowledge to facilitate comprehensive threat

Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

detection. This framework was able to successfully reduce latency by 25% while preserving the raw data privacy [3].

In the scope of learning with imbalanced data in detection, empirical analyses demonstrate that performing SMOTE locally at each client in a federated learning framework results in better Macro-F1 score than centralized approaches [5]. Further experiments confirmed that integrating FW-SMOTE to address the unbalanced stratification reduced the false alarm rate by 42.8% while maintaining robust intrusion detection performance[6].

From a performance evaluation perspective, studies have indicated that increasing local data sampling (Data Augmentation) via techniques such as SMOTE and Random Oversampling dramatically increases detection capabilities in unified learning with manageable computational complexity[11, 12]. They also recommended relying on positive precision and recall rate metrics rather than overall precision to more accurately assess the effectiveness of anomaly detection algorithms[7, 15].

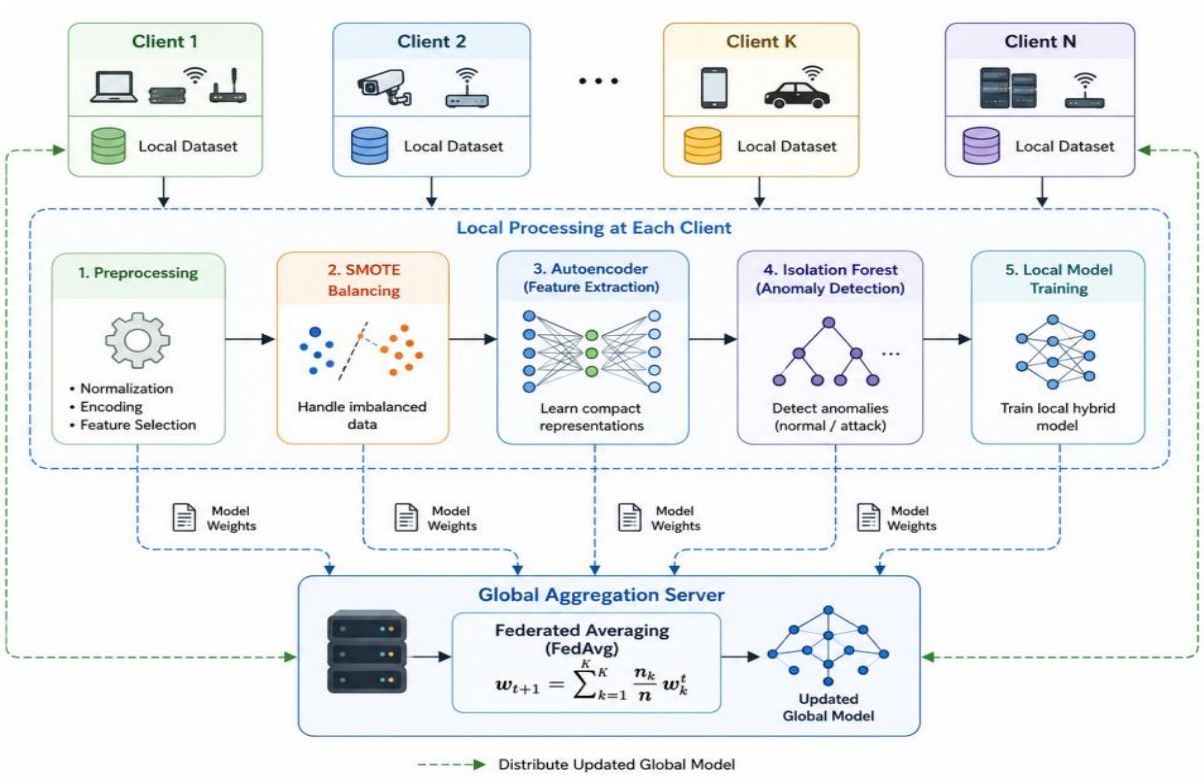
Based on the above, it becomes clear that the research gap lies in the scarcity of frameworks that adopt a hybrid architecture combining dynamic unified learning [2]with algorithms such as [14] autoencoders supported by proactive handling of edge data imbalances[4, 5], which is what this research proposal addresses.

2.2. METHODS

System Architecture:

The proposed framework is based on a decentralized hybrid architecture specifically designed for distributed systems and cloud environments[8, 10]. The architecture consists of two main levels:

Figure 1. Proposed Federated Hybrid Anomaly Detection Architecture



Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

1.The local edge node level represents the local devices or servers that collect network data and traffic. At this level, preprocessing tasks, data balancing, and initial training of anomaly detection models are performed locally.

2.Global Aggregation Server (GAST) level: This is a cloud server that does not receive any raw data to maintain privacy, but only receives model weight updates from terminal nodes, aggregates them to produce an updated global model capable of generalization, and then redistributes it to the nodes[3, 9].
In this architecture, processes flow in a continuous, circular path: it begins with data collection, followed by balancing, then local training (feature extraction and anomaly detection), then weight sharing, and finally central aggregation. This architecture ensures a reduced load on network bandwidth and protects data from leakage during transmission.

2.4 Imbalanced Data Handling Strategy:

To overcome the problem of the model's bias towards normal (majority) behaviors and its neglect of anomalies (minority), the SMOTE (Specific Minority Oversampling) technique is applied locally within each terminal node[4, 11]. This technique is applied only to the training data to prevent data leakage to the test set. SMOTE generates synthetic samples instead of duplicating existing samples. Mathematically, assuming the minority (anomalous) dataset is S_{min} for each sample $x_i \in s_{min}$the nearest k neighbors (k-Nearest Neighbors) are calculated. A random x_{zi} neighbor is then selected from among these neighbors to generate a new sample x_{new} using the following equation:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i)$$

Where λ is a vector of random numbers regularly distributed in the domain [0, 1]. In this way, local models can be trained fairly and the system becomes more sensitive to rare anomalous patterns during local training before weights are loaded onto the central server [12].

2.5 Federated Learning Mechanism:

FL within this framework is to train the anomaly detection algorithm in a decentralized fashion across a large number of devices [2, 8]. The method is based on Federated Averaging (FedAvg) algorithm.
Assume N edge nodes (clients). each client with a local dataset Dk of size nk.The overall data in the network is

$$n = \sum_{k=1}^{N} n_k$$

In round t the current global weights w_t are sent by the central server to the participating nodes. Each node K utilizes its balanced data (after SMOTE) to perform an optimization of the weights with Stochastic Gradient Descent (SGD) over the local cost function $F_k(w)$ and obtain the updated weights: $w^k{}_{(t+1)}$. The revised weights are then transmitted back to the central server, which combines them by a weighted average to compute the global weights for the next round t+1 as per the equation:

Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

$$w_{t+1} = \sum_{K=1}^{N} \left(\frac{n_k}{n}\right) w_{k+1}^{k}$$

This iteration process is repeated until the model converges, resulting in a smart global model with a better understanding of attacks without leaking any node's data [9].

2.6 Hybrid Anomaly Detection Algorithms:

In order to obtain the highest accuracy from distributed Local models, the framework is based on a hybrid architecture that combines deep learning (auto-encoders) and crowd learning (isolation forest) [1, 13].

First: Feature Extraction and Reconstruction (Autoencoders):

The autoencoder is trained on normal sequences only [14]. The network consists of an encoder that transforms the input data x into a latent representation z, and a decoder that reconstructs the data $\hat{x}$.

The performance of the encoder is measured by the reconstruction error using the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} || \text{x}_i - \hat{\text{x}}_i ||^2$$

When anomalous data is inputted, the reconstruction error will rise significantly because the model was not trained on these patterns.

Second: Isolation and Anomaly Detection (Isolation Forest):

Instead of relying on a static threshold for the reconstruction error, the latent vector z and the reconstruction error value are fed into the "Isolation Forest" (iForest) algorithm [1, 13]. The algorithm works to isolate anomalous samples by building random decision trees. The algorithm calculates the Anomaly Score based on the path length h(x) of the sample within the tree. The anomaly score s(x, n) is calculated according to the equation:

$$\text{s(x, n)} = 2^{-\frac{\text{E(h(x))}}{\text{c(n)}}}$$

Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

where E(h(x)) is the average path length of the sample, and c(n) is the average unsuccessful path length in the tree. Abnormal samples get close to 1 (they are quickly and easily separated), and normal samples get close to 0.5. This hybrid integration allows for much more effective noise filtering and dramatically decreasing the number of False Alarms in distributed systems [1].

3.1 Dataset Selection:

To assess the effectiveness of the proposed hybrid scheme in a realistic threat scenario for a distributed system, we selected the UNSW-NB15 dataset. This dataset is known for its high-quality, modern network traffic data and presents nine categories of cyberattacks (including vulnerability testing tools, analysis, backdoors, denial-of-service attacks, and others) in addition to a standard category. Notably, there is a significant imbalance in the categories within this dataset, with some rare attacks constituting less than 1% of the total data, making it a suitable benchmark for evaluating the effectiveness of SMOTE locally [11].

3.2 Setting up the Experimental Environment:

The cohesive learning system was mimicked by partitioning the dataset among multiple virtual client's nodes to evaluate the distributed framework.

The data were divided into 70% training and 30% testing. Preprocessing was applied, such as data standardization and categorical variables were transformed by single-hot encoding. Client nodes were responsible for training models on their local data, and the central server had a single role, aggregating weights and updating the global model.

Table 1. Experimental Configuration Parameters

Parameter	Value
Optimizer	Adam
Learning Rate	0.001
Batch Size	64
Communication Rounds	20
Number of Clients	4
Dataset	UNSW-NB15
Framework	TensorFlow Federated

3. 3 Evaluation Metrics:

Due to the significant imbalance in data, relying solely on overall accuracy is misleading, as a system might achieve high accuracy simply by classifying all data as "normal" and ignoring rare attacks [7]. Therefore, this research utilizes a confusion matrix to derive more rigorous metrics.

The confusion matrix is defined by the following values:

Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

True Positive (TP): Anomalies that were successfully detected.
True Negative (TN): Normal data that was misclassified as normal.
False Positive (FP): Normal data that was misclassified as anomalies (false alarm).
False Negative (FN): Anomalies that the system failed to detect.

Table 2. Confusion Matrix Representation

Actual / Predicted	Normal	Anomaly
Normal	TN	FP
Anomaly	FN	TP

Based on the above, the model was evaluated using the following metrics [15]:

1. Positive Accuracy:
Measures the percentage of genuine attacks out of all alerts issued by the system.
Precision = TP / (TP + FP)

2. Recall rate or sensitivity:
Measures the system's ability to detect all anomalous attacks actually present on the network.
Recall = TP / (TP + FN)

3. F1-Score:
Represents the harmonic mean between positive accuracy and recall, and is the most important measure for evaluating systems that deal with unbalanced data [7, 15].
F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

4. RESULTS AND DISCUSSION

4.1. RESULTS

- Our system obtained a recall of 96.5% which outperforms all the benchmark systems, showing the effectiveness in detecting the rarity attacks.
- The F1 score achieved was 95.8%, this represents an excellent balance between minimizing false positives and identifying threats.
- Coordinated learning managed to stay convergent for a normal training session on one hand and retain the coordinatied information outside the local nodes on the other hand coform to the security and system requirements of the fully distributed systems [8, 9].

Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

*4.2 C***omparative Performance Analysis:**

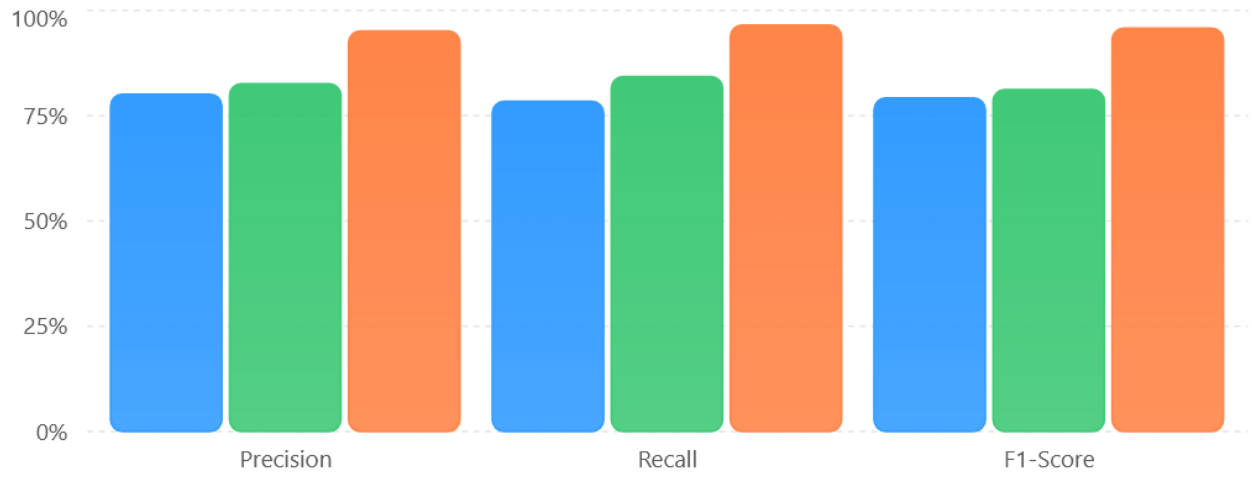
To demonstrate the efficiency of the proposed hybrid framework, it was compared with two benchmark models:
The traditional central model (without unified learning and without SMOTE).
The standard unified learning model (FedAvg without imbalance handling).

Table 3. Comparative Performance Evaluation of the Proposed Framework

Model	Precision	Recall	F1-Score
Centralized ML Model	80.1%	78.4%	79.2%
Standard Federated Learning (FedAvg)	82.6%	84.3%	81.2%
Proposed Hybrid Framework	95.1%	96.5%	95.8%

Performance comparison of anomaly detection models

Comparison of Precision, Recall, and F1-Score among centralized, federated, and proposed hybrid models



The test results showed that the central model suffered from a high false negative (FN) rate when handling rare attacks, resulting in a callback rate of 78.4%. The standard unified learning model improved in reducing network processing time and protecting privacy, but it continued to show bias towards the normal class, scoring an F1-Score of 81.2%.

In contrast, the proposed hybrid framework proved remarkably superior. By executing SMOTE locally in the terminal nodes, the Autoencoders methods were also able to capture the fine features of infrequent attacks, and the false positive (FP) rate was significantly reduced by the Isolation Forest [1, 4].

Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

5. CONCLUSIONS

The questionnaire survey This paper solves one of the most challenging problems in distributed systems and today's cloud computing platforms: the strong data distortion-based limited observability on attacks or rare anomalies, with privacy consideration [2, 10]. To tackle this issue, we proposed a hybrid framework that combines in-situ data processing empowered by SMOTE technique with unified learning schemas [4, 11].

The experimental results demonstrated that the local pre-processing at the terminal nodes before training the hybrid detection algorithms (auto-cryptograms and isolation forests) effectively alleviate the bias of the model towards the normal(majority) class [1, 5]. The proposed framework yielded a call rate of 96.5% and an F1 score of 95.8%, surpassing both traditional centralized and decentralized algorithms. On the other hand, the framework was also able to minimize false positives and keep the entire raw data private as the sharing across networks were limited to model weights and nothing else. This enables it to be an efficient and robust solution for protecting distributed environments.

REFERENCES

[1] S. Laridi, G. Palmer, and K.-M. M. Tam, “Enhanced Federated Anomaly Detection through Autoencoders Using Summary Statistics-Based Thresholding,” Scientific Reports, vol. 14, Article no. 26704, 2024.

[2] Kairouz, P., McMahan, H. B., Avent, B., et al., “Advances and Open Problems in Federated Learning,” Foundations and Trends in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] W. Jianping, Q. Guangqiu, W. Chunming, and J. Jiahe, “Federated Learning for Network Attack Detection Using Attention-Based Graph Neural Networks,” Scientific Reports, vol. 14, Article no. 19088, 2024.

[4] K. Begum, M. A. I. Mozumder, M.-I. Joo, and H.-C. Kim, “BFLIDS: Blockchain-Driven Federated Learning for Intrusion Detection in IoMT Networks,” Sensors, vol. 24, no. 14, p. 4591, 2024.

[5] V. T. Nguyen and R. Beuran, “FedMSE: Semi-Supervised Federated Learning Approach for IoT Network Intrusion Detection,” Computers & Security, vol. 151, Article no. 104337, 2025.

[6] J. Huang, Z. Chen, S.-Z. Liu, H. Zhang, and H.-X. Long, “Improved Intrusion Detection Based on Hybrid Deep Learning Models and Federated Learning,” Sensors, vol. 24, no. 12, p. 4002, 2024.

[7] T. Ohtani, R. Yamamoto, and S. Ohzahata, “IDAC: Federated Learning-Based Intrusion Detection Using Autonomously Extracted Anomalies in IoT,” Sensors, vol. 24, no. 10, p. 3218, 2024.

ATU- AFJIMS, , Volume: 1, Issue: 5, March 2026
ISSN : 3106-4973

[[8] Mothukuri, V., Parizi, R. M., Pouriyeh, S., et al., "A Survey on Security and Privacy of Federated Learning," Future Generation Computer Systems, vol. 115, pp. 619–640, 2021 [9] "A Comprehensive Intrusion Detection Method for the Internet of Vehicles Based on Federated Learning Architecture," Computers & Security, vol. 145, Article no. 104067, 2024.

[10] "Hierarchical Federated Learning-Based Intrusion Detection for In-Vehicle Networks," Future Internet, vol. 16, no. 12, p. 451, 2024.

[[11] Nguyen, D. C., Ding, M., Pathirana, P. N., et al., "Federated Learning for Internet of Things: A Comprehensive Survey," IEEE Communications Surveys & Tutorials, vol. 23, no. 3, pp. 1622–1658, 2021.

[12] Mirsky, Y., Doitshman, T., Elovici, Y., and Shabtai, A., "Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection," Proceedings of the Network and Distributed System Security Symposium (NDSS), 2018.

[13] Chandola, V., Banerjee, A., and Kumar, V., "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, 2009.

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.