# Enhancing Privacy in Critical Agencies IDS: Decentralized AI with Federated Learning

Zeyad Alamoudi[1], Faisal Balamash[1], Anas Alzahrani[1], Abdulaziz Alsharif[1], Ammar Alshehri[1] and Mahmoud Ellouh[1]

***Abstract*— This paper presents an approach to improving privacy in Intrusion Detection Systems (IDS) for critical agencies using decentralized AI and Federated Learning (FL). Traditional centralized AI models for IDS require sharing sensitive raw datasets, posing significant privacy and security risks. The project uses FL to enable agencies to train AI models on their local, sensitive data without sharing raw datasets. By only sharing model parameters, agencies can jointly develop a powerful, shared AI model while ensuring data privacy and meeting regulatory requirements. The implemented system shows high accuracy in detecting various cyber threats, with F1-scores ranging from 0.96 to 1.00 for different attack types, while maintaining the privacy of sensitive data.**

## I. Introduction

Critical infrastructure systems form the backbone of healthcare, finance, and energy sectors. As these systems become increasingly linked, they also become more susceptible to cyber threats that could potentially disrupt essential services. This growing complexity presents both opportunities for efficiency and challenges for security.

Intrusion Detection Systems (IDS) serve as crucial protection, continuously monitoring for potential threats. Traditionally, these systems have been trained by aggregating data from various sources into a central location. While effective, this approach raises significant privacy concerns, particularly when dealing with sensitive data from critical agencies.

## II. Motivation

The motivation for investigating privacy-preserving IDS stems from the unique challenges faced by critical agencies. Consider a power grid operator possessing valuable data on cyber threats. This data, while potentially invaluable for training an IDS, is highly sensitive and cannot be shared in its raw form due to security and regulatory concerns.

This paper aims to develop an IDS that excels in threat detection while maintaining the utmost discretion in handling sensitive data. The goal is to provide agencies with a solution that harnesses the power of FL without compromising data confidentiality, effectively balancing security needs with privacy requirements.

[1]All authors are with King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. Email: {faisal.balamash, zeyad.alamoudi, ammar.alshehri, abdulaziz.alsharif, anas.alzahrani, mahmoud.ellouh}@kaust.edu.sa

## III. Related Work

Before discussing the approach, it's important to acknowledge the foundational work in related fields:

- **Federated Learning in Healthcare**: Researchers in the medical field have led the way in FL for analyzing patient data across hospitals, showing significant progress in preserving patient privacy while improving diagnostics [2].
- **Distributed IDS**: Previous studies have investigated distributed approaches to IDS, focusing on distributed detection mechanisms [3]. While these efforts have improved scalability, this work extends further into privacy preservation.
- **Privacy-Preserving Machine Learning**: Notable advancements have been made in homomorphic encryption for machine learning [4], allowing computations on encrypted data. The approach in this paper, while different, shares the common goal of privacy preservation.

This work builds upon these foundations, combining the privacy benefits of FL with the specific needs of critical infrastructure IDS.

## IV. Federated Learning: The Chosen Approach

Federated Learning forms the approach to privacy-preserving IDS.

### A. Why Federated Learning?

FL was selected for the IDS implementation due to several key advantages:

- **Privacy Preservation**: FL allows data to remain localized with its owner, significantly reducing privacy risks.
- **Federated Potential**: It enables multiple agencies to contribute insights without exposing sensitive information.
- **Adaptability**: FL can effectively handle diverse and changing cyber threats.
- **Regulatory Compliance**: In data protection laws, FL offers a path to compliance while promoting collaboration.

### B. How Federated Learning Works in the IDS

The FL process for IDS follows these steps:

1) The central server initializes an IDS model and distributes it to all participating agencies.
2) Each agency trains this model on their local, confidential data.

3) Agencies transmit only the model updates, not the raw data, back to the central server.
4) The central server aggregates these updates to create an improved global model.
5) This improved model is then redistributed to all agencies, and the cycle continues.

This approach enables federated learning while maintaining data privacy.

## V. Beyond Federated Learning

While the current focus is on FL, the potential for integration with other privacy-preserving algorithms is recognized.

### A. The Multi-Algorithm Potential

The decentralized AI approach is designed with versatility in mind, potentially accommodating various AI algorithms beyond FL. This flexibility positions the system as a comprehensive solution for privacy-preserving AI in cybersecurity.

### B. Split Learning: An Alternative Approach

Split Learning presents an alternative to FL. In this approach, the model itself is divided between the client (agencies) and the server. This method could potentially offer improved privacy guarantees, particularly during the model update transmission phase.

While not implemented in the current project, Split Learning remains a promising avenue for future investigation and potential integration into the privacy-preserving IDS framework.

## VI. System Model

The system model illustrates the practical implementation of FL in IDS.

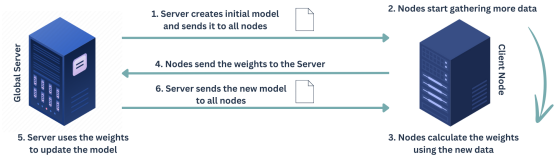



Fig. 1. Federated Learning Process in IDS

As depicted in Fig. 1, the FL process follows a cyclical pattern:

1) The central server initializes the IDS model and distributes it to all client nodes.
2) Each agency collects new data and trains the model locally.
3) Agencies compute weight updates based on their local training.
4) Only the weight updates, not the raw data, are transmitted back to the central server.
5) The central server aggregates these updates to refine the global model.
6) The improved model is then redistributed to all agencies, and the cycle repeats.

This approach ensures continuous improvement of the IDS model through diverse data sources while maintaining data confidentiality.

## VII. Implementation

The implementation phase involved several key steps to bring the FL IDS concept to fruition.

### A. Dataset and Data Preprocessing

The CSE-CIC-IDS2018 dataset was utilized, a comprehensive collection of network traffic data [1]. This dataset is particularly valuable due to its recency, diversity, and realistic representation of modern network threats. It contains benign traffic as well as the most up-to-date common attacks, making it ideal for training an IDS.

To manage the dataset's complexity and ensure a balanced representation, a strategic sampling approach was employed. The goal was to create a subset of 15,000 samples that maintained the dataset's rich diversity while being computationally manageable for the FL setup. This sample size was chosen as it provides a good balance between computational efficiency and maintaining the statistical significance of the results.

The sampling process was carefully designed to ensure a balanced representation across different types of network activity. This balance is needed for training an IDS that can effectively detect a wide range of threats without bias towards any particular type of activity. The approach included:

- Stratified sampling to ensure representation of all attack types
- Oversampling of minority classes to address class imbalance
- Undersampling of majority classes to prevent domination by common traffic types

This balanced sampling was vital to prevent model bias and ensure that the IDS could effectively detect both common and rare attack types.

Data quality was addressed by filling missing values with average values for each feature, ensuring data integrity for the learning algorithm. Feature scaling was also performed to normalize the range of independent variables, which is essential for many machine learning algorithms to perform well.

### B. Data Distribution Across Clients

To simulate a real-world scenario where different agencies possess varying types and amounts of data, the preprocessed dataset was distributed among five client 'agencies' using a Dirichlet distribution [5]. This approach creates intentionally imbalanced datasets across clients, reflecting the reality that different agencies may encounter varying types and frequencies of network activities.

As shown in Figure 2, the data distribution across clients is highly imbalanced. This intentional imbalance simulates real-world scenarios where different agencies have access to different types of data. In particular, Client 2 has a high concentration of Bot and DoS attacks, while Client 3 specializes in DDOS attack-HOIC data. This diverse spread was instrumental in thoroughly testing the FL approach across various scenarios, simulating the real-world challenge

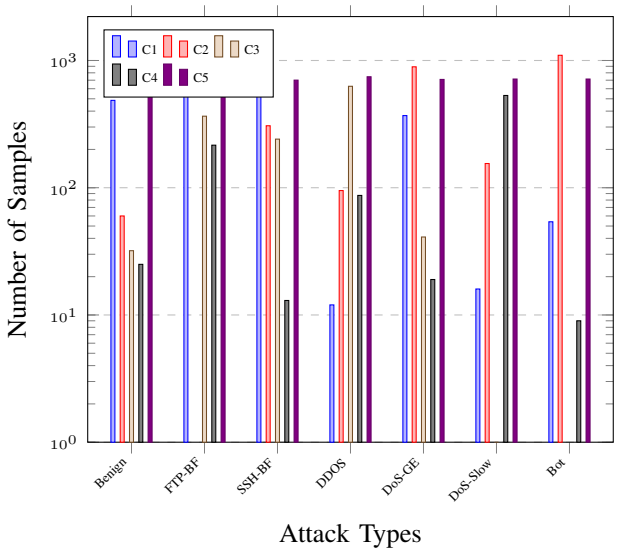



Fig. 2. Client Data Distribution

of dealing with non-IID (non-independent and identically distributed) data across different agencies.

### C. Model Architecture and Training

The neural network design balances simplicity and effectiveness:

- An input layer matching the feature set
- Two dense layers (64 and 32 neurons) with ReLU activation
- An output layer with softmax activation for multi-class classification

The Stochastic Gradient Descent (SGD) optimizer was employed with an initial learning rate of 0.01 and a momentum of 0.9. A learning rate scheduler was implemented to adjust the rate every 5 epochs, ensuring optimal learning progression.

The training process was iterative, with each client training their local model for 5 epochs in each FL cycle. These models were aggregated to create a global model, which was then redistributed. This cycle repeated 15 times, allowing the global model to learn from all clients without direct access to their data.

The dataset was split into 80% for training and 20% for testing, ensuring that the model's performance was evaluated on unseen data. This split was maintained across all clients to ensure consistent evaluation.

This federated approach, combined with the carefully preprocessed and distributed dataset, allowed for high performance across various attack types without overfitting to any particular client's data. The iterative nature of the process, along with the diversity in client datasets, ensures that the model generalizes well across different types of network activities.

Figure 3 shows the quick improvement in model accuracy over the FL cycles. The model achieved over 90% accuracy by the 4th cycle, with continued improvement leading to a plateau near 100% accuracy. This rapid improvement demonstrates the effectiveness of the FL approach in leveraging diverse data sources.

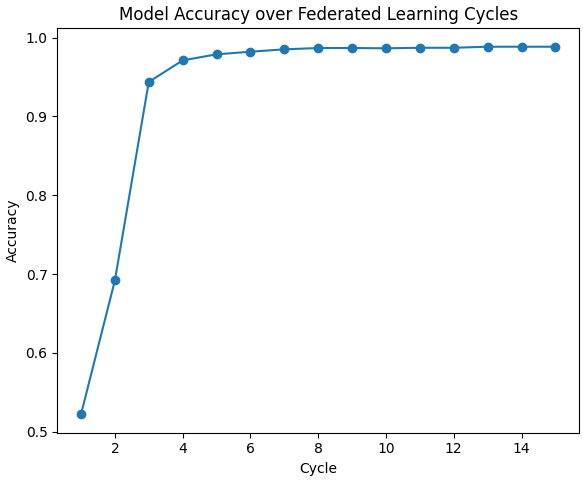



Fig. 3. Model Accuracy over Federated Learning Cycles

Figure 4 shows a corresponding sharp decline in loss during initial cycles, followed by a more gradual decrease as training progressed. This pattern is typical of deep learning model training and indicates that the model is learning effectively from the federated data. The consistent decrease in loss suggests that the model is successfully generalizing across the diverse datasets of different clients.

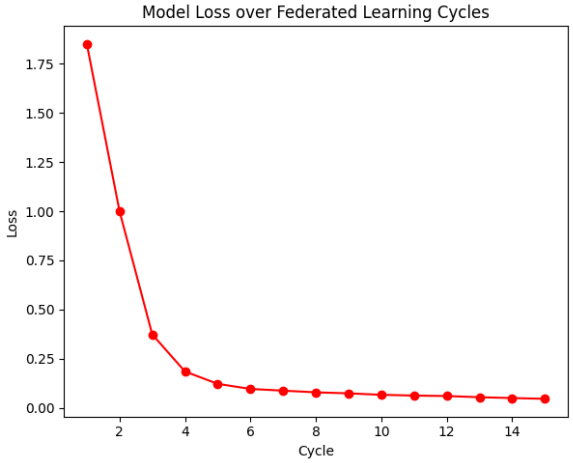



Fig. 4. Model Loss over Federated Learning Cycles

These figures illustrate the effectiveness of the FL approach. The quick increase in accuracy and decrease in loss over the first few cycles show that the model quickly learns to distinguish between different types of network activities. The subsequent plateauing indicates that the model has reached a stable state of high performance.

Importantly, these results are based on the 20% test data that the model has not seen during training. This suggests that the model is generalizing well to new, unseen data, rather than overfitting to the training set. The consistent high

performance across multiple cycles also indicates that the model is stable, capable of maintaining its accuracy even as it continues to learn from diverse data sources.

## VIII. Results and Discussion

The performance of the FL model proved to be highly promising. The results are discussed in terms of individual client performance and overall model evaluation.

### A. Client Performance

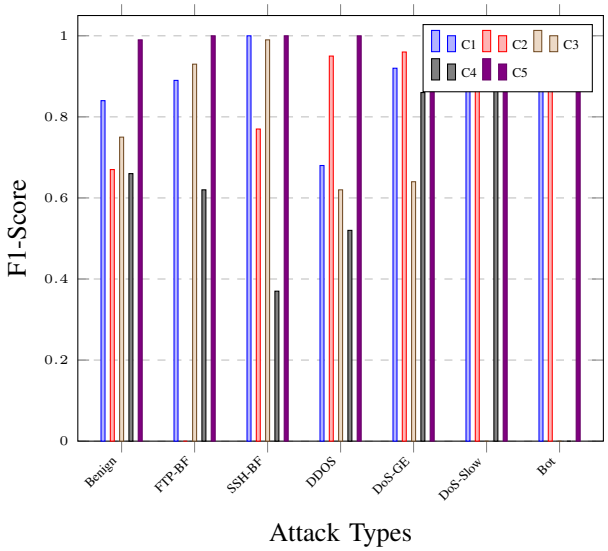



Fig. 5. Client Performance (F1-scores) Across Attack Types

As evident from Figure 5, Client 5 consistently outperforms other clients, likely due to its well-rounded dataset distribution. This highlights the importance of diverse, balanced datasets in training effective IDS models. Other clients show varying performance across different attack types, reflecting their specialized datasets. For instance, Client 2 excels at detecting DoS attacks but struggles with FTP-BruteForce, aligning with its data distribution. These results underscore the effectiveness of FL in leveraging diverse data sources.

### B. Model Evaluation Metrics

The overall performance of the FL model is summarized in Figure 6:

Figure 6 presents the overall performance metrics of the FL model. The model achieves high precision, recall, and F1-scores across all categories, with most scores at or near 1.0. This demonstrates the strength of the approach in detecting various types of network intrusions.

### C. Discussion of Results

The results of the FL approach to IDS are highly promising:

- **High Overall Performance**: The model achieves F1-scores ranging from 0.96 to 1.00 across different attack types, indicating excellent detection capabilities for both common and rare attack types.

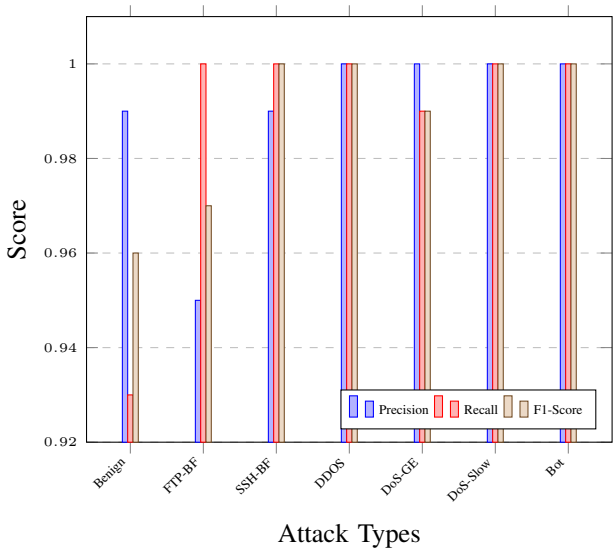



Fig. 6. Model Evaluation Metrics

- **Effectiveness of Federated Learning**: Despite the non-IID nature of the distributed dataset, the FL approach successfully leveraged the diverse data from all clients to create a strong global model.
- **Privacy Preservation**: By keeping raw data local and only sharing model updates, data privacy was successfully maintained while achieving high detection accuracy.
- **Adaptability**: The model's high performance across various attack types demonstrates its ability to adapt to different threat scenarios, an important feature for real-world IDS applications.

The variation in individual client performance (as seen in Figure 5) highlights the importance of data diversity in training effective IDS models. While some clients excelled in detecting specific attack types aligned with their local data distribution, the federated model managed to achieve high performance across all categories. This underscores the power of FL in overcoming individual data limitations.

## IX. Ethical Considerations: The Human Impact of Federated Learning in IDS

As FL advances in intrusion detection systems, it is important to consider the human dimensions and ethical implications of this work. This section examines the ethical lens that balances technological advancements with societal responsibility.

### A. Privacy: Beyond Technical Protection

While FL improves data privacy, considerations must extend beyond technical solutions to fully address privacy concerns:

- **Residual Privacy Risks**: Consider a scenario where an agency believes its sensitive network data is fully protected, only to later discover that model updates inadvertently revealed information about their security

posture. Systems that not only protect raw data but also guard against subtle information leakage must be built.
- **Informed Consent in Practice**: In a world of complex technologies, ensuring that all participating agencies, from large government institutions to smaller local utilities, truly understand the implications of their involvement is crucial. Transparency that goes beyond legal jargon to foster genuine informed consent must be strived for.
- **Right to be Forgotten in a Federated World**: If an agency decides to withdraw from the FL system, honoring their right to be forgotten when their contributions are intertwined with a globally distributed model becomes challenging. This requires rethinking data ownership and model attribution in federated systems.

### *B. Accountability in a Federated Framework*

The distributed nature of FL raises important questions about responsibility and oversight:

- **Tracing Errors in Federated Models**: When a critical threat goes undetected, tracing the root cause in a model built from diverse, private datasets becomes complex. Frameworks for accountability that respect the federated nature of FL while ensuring system reliability need to be developed.
- **Auditing Without Compromising Privacy**: Traditional security audits often require access to systems and data. Adapting these processes for federated systems, maintaining rigorous oversight without undermining the privacy benefits, is a challenge that needs to be addressed.

### *C. Charting an Ethical Path Forward*

To address these multifaceted challenges, the following are proposed:

- Developing an ethical framework specifically designed for FL in cybersecurity, involving diverse stakeholders from technologists to ethicists and policymakers.
- Establishing governance models that balance responsibility, ensuring that ethical considerations are at the forefront of development and deployment decisions.
- Committing to regular ethical impact assessments throughout the lifecycle of the FL IDS, adapting the approach as new challenges and insights are uncovered.

By proactively engaging with these ethical considerations, the aim is to develop FL IDS solutions that not only advance technological capabilities but also align with societal values and ethical principles. The goal is to create systems that people can trust to protect critical infrastructure while respecting fundamental rights and values.

## X. Conclusion

This paper demonstrates the efficacy of an FL approach in improving privacy for Intrusion Detection Systems in critical agencies. It has been shown that it's possible to develop strong IDS models through inter-agency collaboration without compromising sensitive data.

The performance of the model, with F1-scores ranging from 0.96 to 1.00 across various attack types, highlights the potential of this privacy-preserving approach in cybersecurity. This work opens up several avenues for future research and application:

- **Expanding to Other Domains**: This approach could be applied to other areas of cybersecurity or even fields like healthcare and financial fraud detection.
- **Further Optimization**: Future work could focus on optimizing the FL process for even greater efficiency, particularly in scenarios with more clients or larger datasets.
- **Exploring Hybrid Approaches**: Combining FL with other privacy-preserving techniques, such as differential privacy or homomorphic encryption, could lead to even stronger privacy guarantees.
- **Real-World Deployment**: The next logical step would be to deploy this system in real-world critical infrastructure settings, allowing for further refinement based on actual operational data.
- **Addressing Non-IID Challenges**: While the model performed well despite non-IID data, further research into techniques for handling extremely imbalanced or heterogeneous data distributions could improve the strength of FL in diverse real-world scenarios.

In conclusion, this research demonstrates that with machine learning techniques, stronger, smarter, and more private Intrusion Detection Systems can be built. As cybersecurity evolves, privacy-preserving techniques like FL are poised to play a crucial role in protecting critical infrastructure while respecting data privacy. This approach not only improves security but also promotes collaboration between agencies, potentially leading to more comprehensive and adaptive cybersecurity strategies in the future.

## References


[1] IPS/IDS dataset on AWS (CSE-CIC-IDS2018). [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html

[2] J. Xu et al., "Federated Learning for Healthcare Informatics," Journal of Healthcare Informatics Research, vol. 5, no. 1, pp. 1-19, 2021.

[3] A. Patel et al., "A Distributed Approach to Intrusion Detection Systems," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1988-2001, 2021.

[4] C. Zhang et al., "Privacy-Preserving Machine Learning through Data Obfuscation," arXiv preprint arXiv:2005.12763, 2020.

[5] T. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," The Annals of Statistics, vol. 1, no. 2, pp. 209-230, 1973.