IEEE GLOBECOM®

Taipei, Taiwan

8-12 December 2025

Greece 2.0 NATIONAL RECOVERY AND RESILIENCE PLAN

Funded by the European Union NextGenerationEU

Mitigating Model Replacement Attacks in Federated Learning-based Intrusion Detection Systems

Pavlos Bouzinis, Panagiotis Radoglou-Grammatikis, Nikos Mitsiou, Thrassos Oikonomou, Vasiliki Koutsioumpa, Georgios Th. Papadopoulos, Vasiliki Kotsirou, George Karagiannidis, and **Panagiotis Sarigiannidis**

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Authors & Contributors

Pavlos Bouzinis

Panagiotis Radoglou-Grammatikis

Nikos Mitsiou, Thrassos Oikonomou, Vasiliki Koutsioumpa, George Karagiannidis

Georgios Th. Papadopoulos

Panagiotis Radoglou-Grammatikis, Vailiki Kotsirou, Panagiotis Sarigiannidis

Funded

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Presentation Structure

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Introduction

1

- ❑ **Introduction**
- ❑ **Motivation & Contributions**

Main Part

2

- ❑ **Model Replacement Attacks in FL**
  - Attack Description & Formulation
  - Defense Mechanism
- ❑ **Evaluation & Results**

Conclusions

3

- ❑ **Key Findings & Future Work**

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Introduction, Motivation & Contributions

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. ΥΠ3ΤΑ-0560908.

Introduction

FL-based Network Intrusion Detections Systems (NIDS): enables multiple devices/clients to collaboratively develop AI-based IDS models without sharing their raw training data and keeping them localized.

Common attacks against FL-based NIDS: include poisoning attacks, where adversaries manipulate local training data to corrupt the global model, and backdoor attacks, where malicious clients embed hidden triggers that make the model misclassify specific inputs.

Model replacement attack in FL-based NIDS: malicious clients attempt to **directly overwrite the global FL model** with a compromised version, often aiming to insert backdoor tasks that bypass detection during inference or subtly corrupt the intended model behavior.

Motivation & Contributions

Motivation

- FL may increase the attack surface, while model replacement attacks can degrade global model performance without raising suspicion.
- The effects and impact of model replacement attack on FL-based NIDS should be understood. Appropriate defense mechanisms should be designed and evaluated.

Contributions

- Formulation of model replacement attack in the context of FL-based NIDS.
- Design of a weight clipping-based defense mechanism, to supress malicious updates. The techniques does not exclude clients from the FL process.
- Demonstration of defense's efficacy in supressing model replacement attacks, while maintaining main-task detection performance

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Model Replacement Attack

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

IEEE GlobeCom// Taipei, Taiwan 8-12 December, 2025

Model Replacement Attack

Attacker Assumptions & Objective

- **The attacker has full access to the local dataset of the compromised FL client:** It can modify the dataset by any means (e.g., change labels, modify feature values) and generate a malicious one.
- **The attacker has complete control over the local training process of the impersonated FL client:** This allows them transmit any modified weights to the FL server.

The **objective** of the attacker is to create a backdoor task, ensuring that the global FL model fails on specific targeted tasks while still performing successfully on others (e.g., bypass only SYN-flood DoS attacks without detection by the NIDS)

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Model Replacement Attack

Attack Formulation

- During FL round t, attacker sends the following update to the server:

$$\beta \left(\boldsymbol{w}_{\text{mal}}^{*} - \boldsymbol{w}^{t} \right)$$

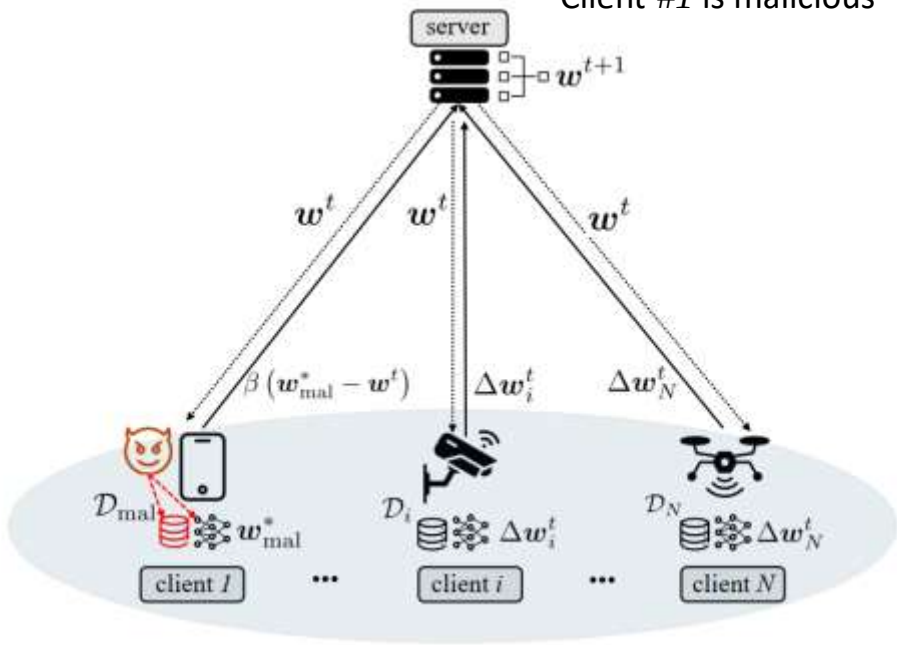
where

$\boldsymbol{w}_{\text{mal}}^{*}$: Pre-trained malicious model on a malformed dataset $\mathcal{D}_{\text{mal}}$

$\boldsymbol{w}^{t}$: Global model at round t

β : boost factor (usually equal to the number of clients N)

N clients,
Client #1 is malicious



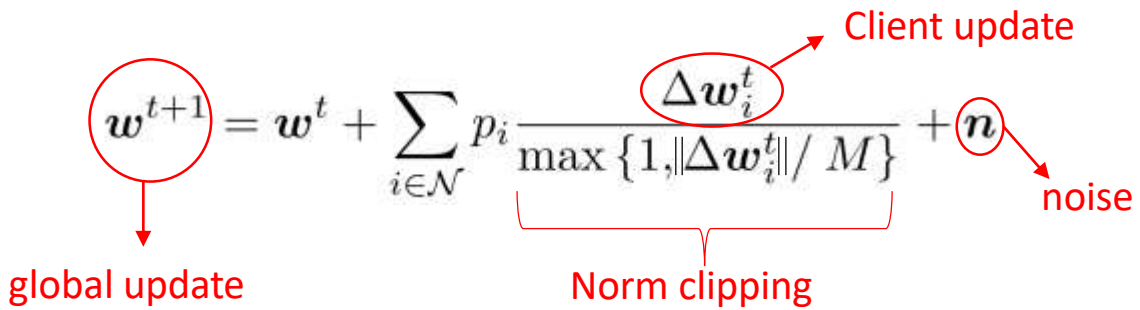
This update aims to replace all client model updates with the malicious one, $\boldsymbol{w}_{\text{mal}}^{*}$

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Model Replacement Attack

Defense Mechanism

- Since the weight re-scaling operation of the attacker introduces noticeable differences in magnitude between malicious and benign updates, **the server integrates norm clipping in the FL aggregation step** t, as follows:



where

$M \triangleq \alpha \cdot \mathbb{E}_{i \in \mathcal{N}} \left[\|\Delta \boldsymbol{w}_i^t\| \right]$: clipping threshold (with an order of the average model update norm)

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Model Replacement Attack

Overall Algorithm

Algorithm 1 Model replacement attack and server's defense

1: server **initializes** global model $\boldsymbol{w}^0$
2: attacker **defines** $\mathcal{P}$, $\mathcal{D}_{\text{mal}}$, β, and **trains** $\boldsymbol{w}^*_{\text{mal}}$ on $\mathcal{D}_{\text{mal}}$
3: **for** $t = 0, 1, 2, \ldots$ **do** → Iteration over FL rounds
4: server **sends** $\boldsymbol{w}^t$ to all clients
5: **for** each client $i \in \mathcal{N} \setminus \{1\}$ **do**
6: **train** local model $\boldsymbol{w}_i^t$ on $\mathcal{D}_i$
7: **send** $\Delta \boldsymbol{w}_i^t = \boldsymbol{w}_i^t - \boldsymbol{w}^t$ to the server
8: **if** $\mathcal{P}$ **then**
9: attacker **sends** $\beta\,(\boldsymbol{w}^*_{\text{mal}} - \boldsymbol{w}^t)$ to the server
10: **else**
11: attacker **trains** local model $\boldsymbol{w}_1^t$ on $\mathcal{D}_1$
12: attacker **sends** $\Delta \boldsymbol{w}_1^t = \boldsymbol{w}_1^t - \boldsymbol{w}^t$ to the server
13: server **computes** $\boldsymbol{w}^{t+1}$ according to (4) → Server applies defense mechanism during aggregation
14: **output**: global model $\boldsymbol{w}$

repeat

Benign clients train the local model and send the updates to the server

If attacker decides to attack, sends malicious update. Otherwise, acts as benign client.

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Evaluation & Results

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Experimental Setup

Attacker's Backdoor task

Replace the global model with a malicious one, with the aim to classify "`SYN Flood`" labeled instances as "`Normal`", instead of "`DoS`".

Objective: This approach increases the likelihood that the model remains vulnerable to future attacks that exploit SYN flooding

Dataset

CIC-IOT-2023

Classes: 8
Format: Tabular
Features: 48 features (network traffic statistics)
Domain: IoT

Settings

Number of clients: 10
Total FL rounds: 40
Model: Multi-layer perceptron (3 layers)
Number of attackers: 1
Attack frequency: Every 5 rounds

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Evaluation Metrics

Main Task Accuracy

Measures the classification accuracy across all attack categories and normal traffic.

$$\mathbf{ACC} = \frac{\textbf{correctly classified samples}}{\textbf{total test samples}}$$

Reflects the IDS's fundamental detection capability that must be preserved under attack scenarios

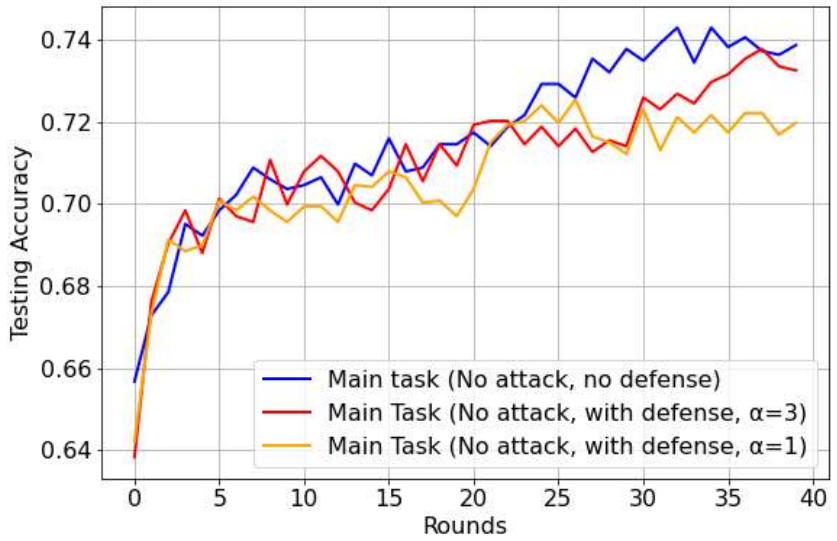
Backdoor Task Accuracy

It measures the attacker's objective by calculating the percentage of targeted "`SYN Flood`" attack samples that are misclassified as "`Normal`", specifically

$$\mathbf{ACC} = \frac{\textbf{"SYN Flood" samples classified as "Normal"}}{\textbf{total "SYN Flood" samples}}$$

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Experimental Results



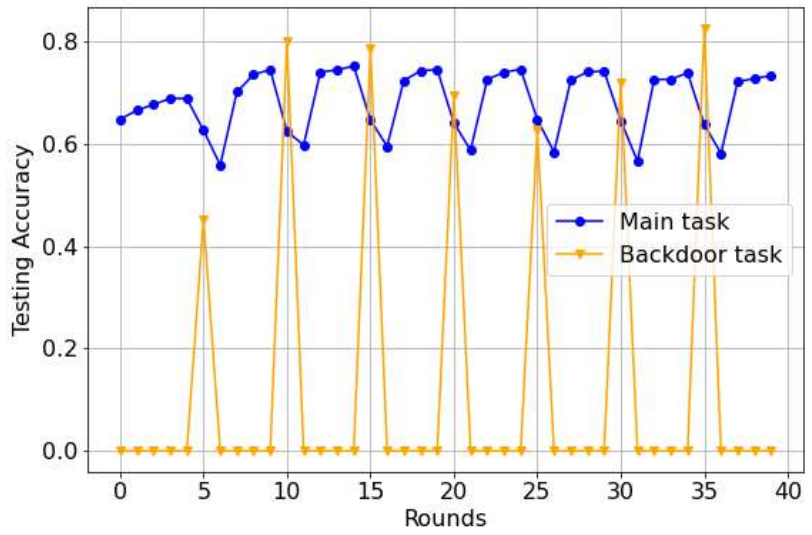
Testing accuracy vs. FL rounds in the absence of an attack

- This figure showcases the evolution of testing accuracy across FL rounds in the absence of an attack.
- It demonstrates the incorporation of the defense mechanism on the server side.
- For suitable values of the hyperparameter α (e.g., α=3), the performance is equivalent to the benchmark, indicating that the defense mechanism does not undermine the model in normal operational scenarios.

α is a hyperparameter that adjusts the weight clipping threshold applied by the server

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Experimental Results

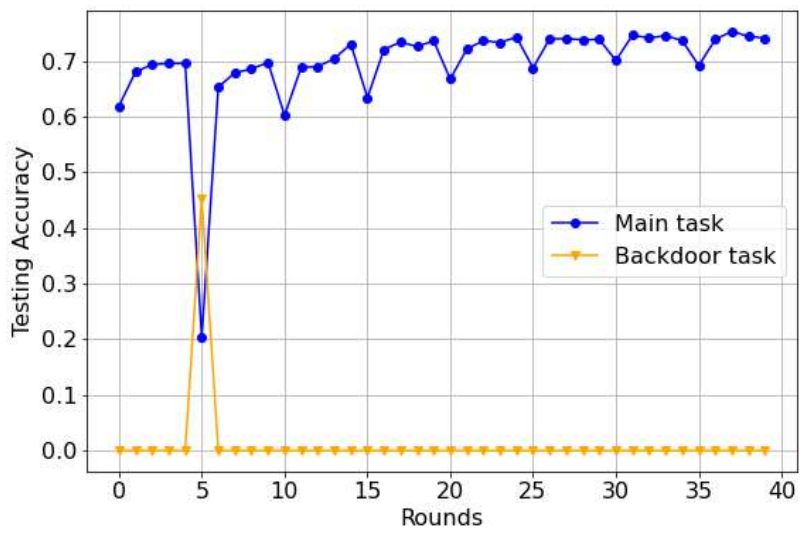


Impact of the model replacement attack without defense mechanism

- This figure shows the effect of the attack when no defense mechanism is applied on the server's side
- It is evident that the accuracy of the backdoor task is high during the particular rounds in which the attack was launched (every 5 rounds)
- On the other hand, although main task accuracy drops during malicious updates, it doesn't collapse entirely and begins to recover once normal training resumes, indicating the attack affects only the target class while leaving most model functionality intact.

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Experimental Results



Impact of the defense mechanism against the model replacement attack

- This figure demonstrates the main task and backdoor task accuracy of the global model, when applying the defense mechanism
- It can be observed that the defense mechanism completely suppresses the attack, as the backdoor accuracy reaches 0%
- Additionally, while the main task accuracy experiences a slight decline during the attack rounds, this decrease is minimal and diminishes as the rounds progress

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

Conclusions and Future Work

Key Findings

- Model replacement attacks in FL-based NIDS can be successful, introducing backdoors that **corrupt the intended model behavior**.
- The **backdoors are difficult to detect** because they cause only a slight reduction in the model's overall performance.
- Weight clipping defense approach **neutralized the attacker's ability to induce misclassifications** without compromising the NIDS's overall performance.

Future Work

- Explore adaptive weight clipping thresholds that are effective in a wide range of scenarios.
- Examine model replacement attacks in which multiple attackers collaborate, making the threat significantly more difficult to defend against.

IEEE GLOBECOM®

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. ΥΠ3ΤΑ-0560908.

Thank you for your attention!

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework (ΑΣΚΟΣ - Grant Agreement no. ΥΠ3ΤΑ-0560864)

IEEE GlobeCom// Taipei, Taiwan 8-12 December, 2025