UNIVERSITÀ DI BOLOGNA

Defense... But at what cost?

CYBERSECURITY A.Y. 24/25

Görkem - 1900126677
José Ribeiro - 1900130105
Silas Pohl - 1900124387
Sofia Costa - 1900129396

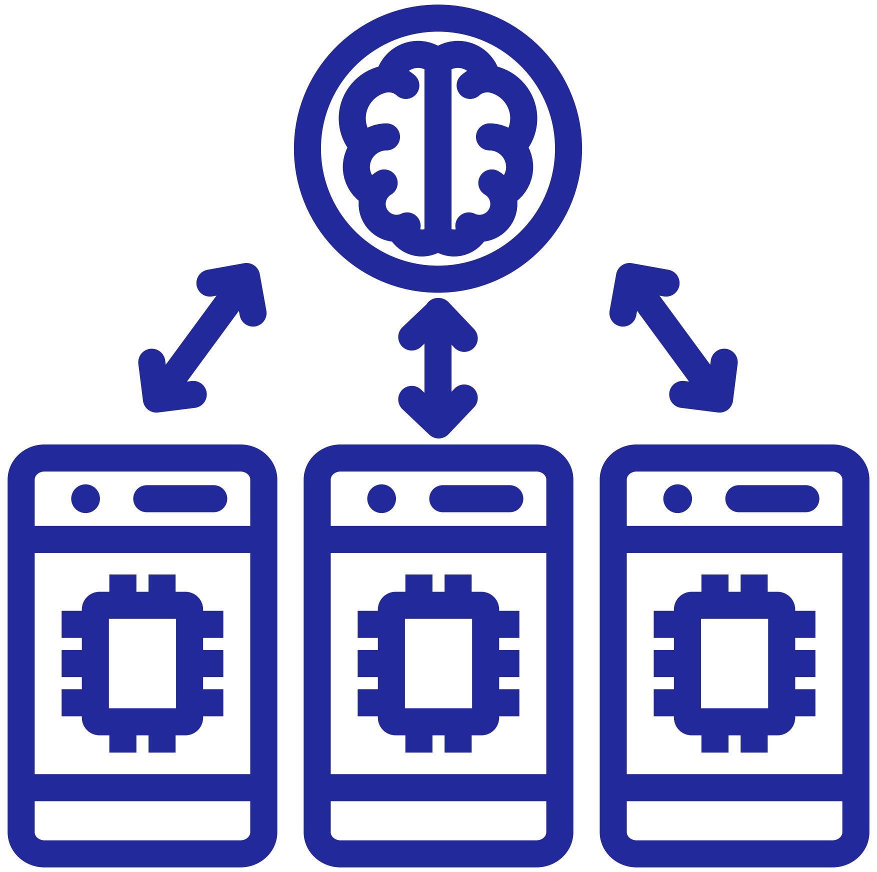
In this Presentation

Table of Contents

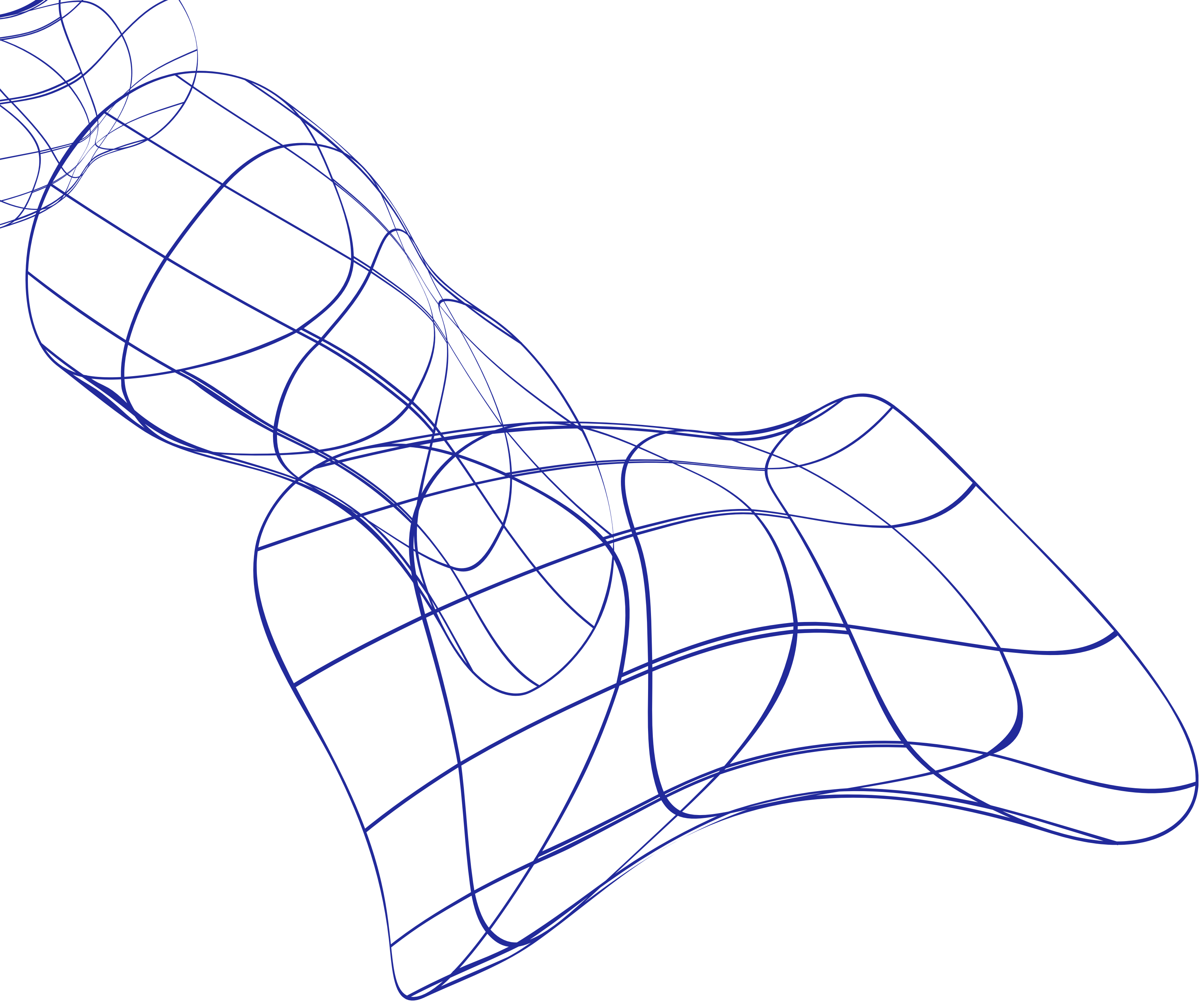
1. What is Federated Learning?

Federated Learning

- Federated Learning (FL) is a **decentralized** approach to training machine learning models, unlike machine learning settings (centralized).
- It leverages local data distributed across multiple clients, enabling **collaborative model training** without exposing sensitive information.
- However, its decentralized nature exposes it to adversarial threats, including **model poisoning**, **data poisoning**, and **backdoor attacks**.
- What can mitigate these attacks?
    - **Anomaly detection techniques.**



2. What is Anomaly Detection?



Anomaly Detection

- Anomaly Detection identifies suspicious activity that falls outside of established normal patterns of behavior.
- It is considered a more proactive type of defense that explicitly detects malicious updates and prevents their impact on the system.
- In FL environments, attacks such as **data poisoning** and **model poisoning** can be discovered using anomaly detection techniques.
- Some interesting types of anomaly detectors are **TRIM** and **RONI** (Reject On Negative Impact).

3. Objectives

Objective of the Study

A State-of-the-Art analysis of anomaly detection techniques within Federated Learning: **TRIM** and **RONI**

- **Examine Anomaly Detection Methods**

  Investigate the methodologies of TRIM and RONI, their architectures and mechanisms for identifying and mitigating poisoned data points.

- **Evaluate Performance**

  Assess the effectiveness of TRIM and RONI in reducing the impact of poisoning attacks using quantitative metrics such as Mean Squared Error .

- **Comparison of Techniques**

  Identify the strengths, weaknesses, and operational differences between TRIM and RONI, and where one outperforms the other.

- **Provide Insights for Future Research**

  Discuss the findings and offer practical recommendations for improving Anomaly Detection in Federated Learning.

4. Datasets and Evaluation Setup

"Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning"
by Matthew Jagielski et al.

Datasets

Evaluate TRIM and RONI using real-world datasets.

Healthcare Dataset

- Predict weekly Warfarin dosage.
- 5,700 samples, 167 features.

Loan Dataset

- Predict loan interest rates.
- 5,000 samples (subset), 89 features.

House Pricing Dataset

- Predict house sale prices.
- 1,460 samples, 275 features.

Simulated Poisoning Attacks

Experimental setup to evaluate the performance

Types of Poisoning Attacks

- **Statistical:** Mimic legitimate data to avoid detection.
- **Optimization-Based:** Maximize disruption to model performance.
- with **poisoning rates** of **4%-20%** of the training data

Datasplits

- **Training**
- **Validation**
- **Testing**

Scenarios Evaluated

- **Baseline MSE:** Clean training data
- **Poisoned MSE:** Training with adversarial data
- **Mitigated MSE:** Using TRIM or RONI

5. Anomaly Detector: TRIM

TRIM: Approach and Architecture

Iterative algorithm to exclude anomalous data points.

01

Initialization:
Train on all available data, including potentially poisoned points to provide baseline parameter estimates

02

Residual Calculation:
For each data point, the residual (the error between the model's prediction and the actual data) is calculated

03

Subset Selection:
TRIM selects a subset of training points with the smallest residuals, (most likely to represent legitimate, unpoisoned data)

04

Model Update:
The model is retrained using only the selected subset, refining its parameter estimates while ignoring high-residual (potentially poisoned) points.

05

Iteration
Steps 2-4 are repeated until convergence (model parameters stabilize,
and the loss function converge)

No Assumptions on Data Distribution:
TRIM does not rely on predefined assumptions about the underlying data distribution, making it suitable for diverse and heterogeneous datasets.

TRIM: Performance

Robust, efficient, and adaptable for anomaly detection in Federated Learning.

in numbers

Median MSE Increase: 6.1 %

Demonstrates TRIM's resilience to adversarial influence, ensuring the model remains close to baseline accuracy.

Maximum MSE Increase: 27.2%

Only 20% of adversarial scenarios caused deviations above this level, showing strong defense even under intense attacks.

Computational Efficiency: 0.02s

Processes datasets rapidly, even for high-dimensional data like the 275-feature house pricing dataset.

Consistency Across Datasets

Robust performance on diverse datasets (healthcare, loan, housing) ensures applicability to varied domains.

Adversarial Mitigation

Neutralizes poisoned data effectively, safeguarding model accuracy in adversarial environments.

Scalability for FL

Rapid convergence and low computational demand make TRIM practical for large-scale systems with many clients.

6. Anomaly Detector: RONI

Introduction and Context

What is RONI?

- A method to identify and exclude harmful data

Why is it needed?

- FL is vulnerable to adversarial threats like model and data poisoning

What are the key goals?

- Enhance robustness by filtering malicious or suboptimal updates.

Approach and Architecture

1- Baseline Model Training → **2- Impact Assessment** → **3- Impact Evaluation** → **4- Data Rejection**

The **RONI method** systematically evaluates the impact of new data or client updates on the model's performance. It establishes a baseline performance metric, assesses the impact of each update using a validation set, and excludes data that degrades the model's performance beyond a predefined threshold. This approach ensures the global model remains robust against malicious or suboptimal updates.

Performance

Effectiveness Factors:

- Validation set quality
- Performance thresholds

Computational Overhead:

- The RONI method requires retraining the model for each new data point or client update. This process is resource-intensive and can significantly increase the time and computational resources needed, especially in large-scale Federated Learning systems.

Threshold Sensitivity:

- Setting the performance degradation threshold is a critical challenge.

  - Strict thresholds: May result in rejecting beneficial updates, reducing the model's learning capability.
  - Lenient thresholds: May allow harmful updates to pass through, compromising the model's robustness.
  - Properly balancing this trade-off is essential for RONI's effectiveness.

Key Takeaways

- RONI is a robust anomaly detection method
- Balances model integrity with computational cost.
- Requires careful tuning of thresholds for optimal performance.
- Plays a pivotal role in securing FL systems.

7. Comparative Analysis

Overview of Performance Comparison

Evaluating TRIM and RONI on 3 datasets (Healthcare, Loan, House Price)

TRIM

Outperforms RONI in all cases

- Improves the MSE of RONI by a factor of **20.28.**
- Consistent performance across all 3 datasets

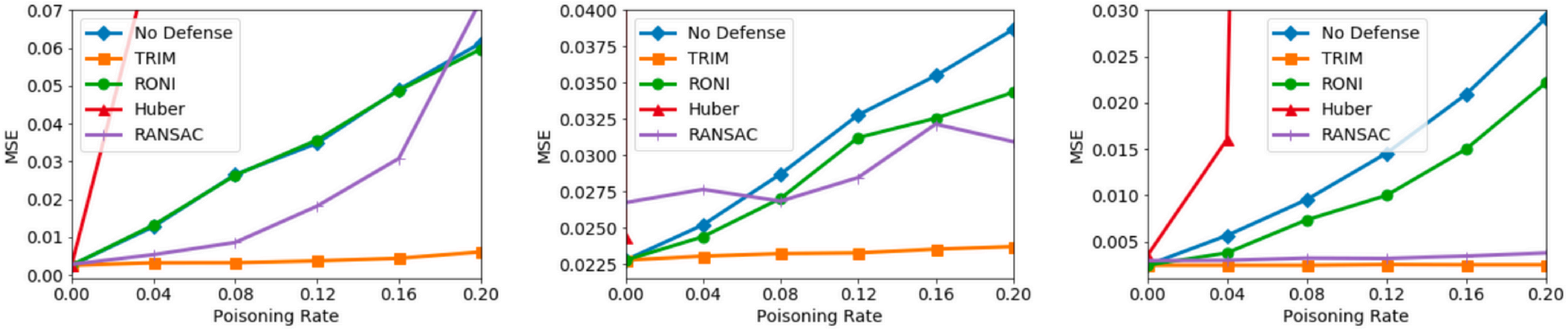
RONI

Inconsistent performance

- MSE increase: Up to **8.06%** above undefended models in several scenarios.
- Highly variable results, sometimes worse than undefended models.

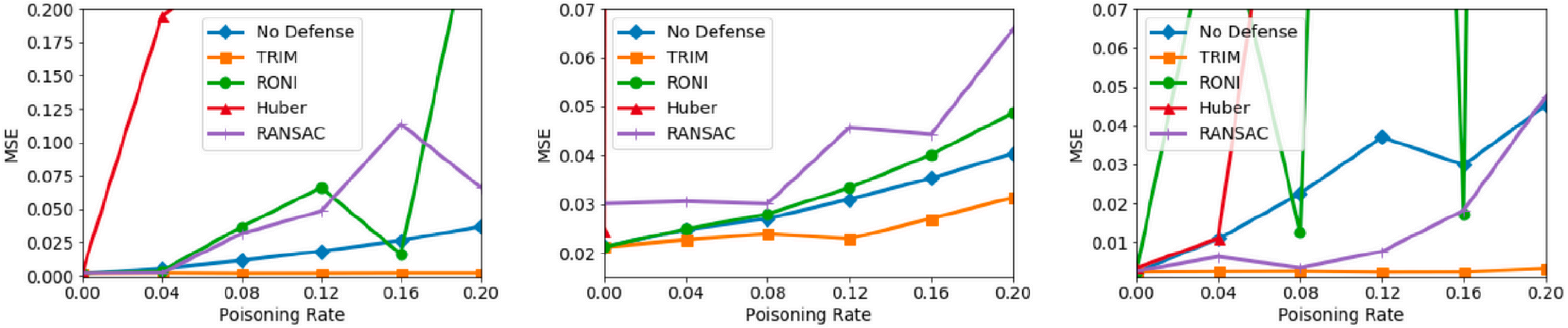
Performance Graphs

TRIM and RONI on ridge[1] and LASSO[2] regression.



(a) Health Care Dataset (b) Loan Dataset (c) House Price Dataset

[1]



(a) Health Care Dataset (b) Loan Dataset (c) House Price Dataset

[2]

Comparative Analysis

Robustness Comparison

TRIM

- Maintains low mean squared error (MSE), with the increase of poisoning rate
- Median MSE increase: 6.1% under high poisoning rates.

>

RONI

- Fails to maintain MSE levels with the increase of poisoning rate
- Highly inconsistent values in some instances.

Comparative Analysis

Efficiency Comparison

TRIM

- Iterative trimming converges quickly.
- Averages 0.02 seconds on the Healthcare and House Price datasets

>

RONI

- Computationally expensive, requires retraining for every point.
- Avg. runtime of 14.8 seconds on Healthcare & 15.69 seconds on the House price datasets

Pros and Cons

TRIM

+ Pros

- High robustness to poisoning attacks.
- Computational efficiency.
- Consistent performance across datasets.

— Cons

- Limited applicability to non-linear models.
- May erroneously exclude legitimate data.
- Scalability challenges with very large datasets.

RONI

+ Pros

- Filters data that negatively impacts performance.

- Threshold sensitivity for outlier detection.

— Cons

- Significant computational overhead.
- Prone to false positives/ negatives.

8. Conclusion

Conclusion

Main Insights:

- **Robustness**: TRIM consistently outperformed RONI across all datasets, maintaining low MSE even at high poisoning rates.
- **Efficiency**: TRIM's iterative approach converged quickly with minimal computational overhead, making it ideal for Federated Learning.
- **Scalability**: TRIM adapted well to diverse datasets, demonstrating versatility in adversarial scenarios.
- **Limitations**: RONI showed inconsistent results and significant computational overhead, limiting its scalability.

Insights for Future Research:

- Extend TRIM to non-linear models for broader applicability in Federated Learning.
- Optimize RONI's computational efficiency for large-scale deployments.
- Explore hybrid solutions combining the strengths of TRIM and RONI for enhanced anomaly detection.

Do you have any questions?

Let us know! We hope you learned something new.