Rochester Security Summit 2023

October 25th and 26th, 2023

FLAME:

Federated **L**earning against **M**alicious **E**ngineering.

Employing Trust and Reputation to Enhance Learning Security and Privacy

Sergei Chuprov

Dr. Leon Reznik

B. Thomas Golisano College of Computing and Information Sciences,

Rochester Institute of Technology

sc1723@rit.edu, leon.reznik@rit.edu

Quick self-introduction

Dr. Leon Reznik

Bio: Professor of Computer Science (primary affiliation) and Computing Security (secondary affiliation) at the Rochester Institute of Technology.

PhD Students:

- Igor Khokhlov, 2016-20, Asist. Professor of Cybersecurity at Sacred Heart University, CT
- Sergei Chuprov, 2021-24

Master's students:

- Graduated **58** students over last five years

https://www.cs.rit.edu/~lr/

Quick self-introduction

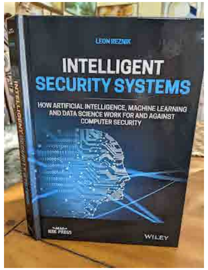
Dr. Leon Reznik

Current courses (taught at least once over last ten years):

- CSCI-532 Introduction to Intelligent Security Systems
- CSCI-630 Foundations of Intelligent Systems
- CSCI-735 Foundations of Intelligent Security Systems
- CSCI-736 Foundations of Neural Networks and Machine Learning
- CSCI-788 MS Project supervision

My new textbook:

L. Reznik Intelligent security systems: How artificial intelligence, machine learning and data science work for and against computer security, 2022, 384 pp, IEEE Press-Wiley&Sons, URL


LEON REZNIK
INTELLIGENT SECURITY SYSTEMS
HOW ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DATA SCIENCE WORK FOR AND AGAINST COMPUTER SECURITY
IEEE PRESS
WILEY

https://www.cs.rit.edu/~lr/

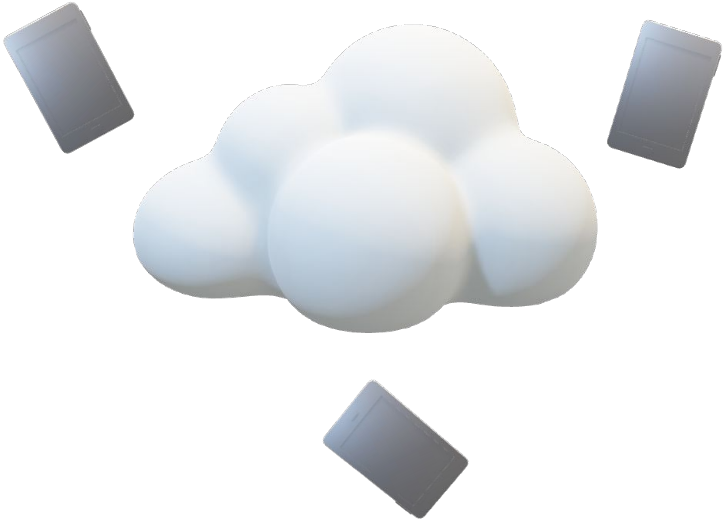
Lab of Data Quality and Intelligent Security

Major projects and products:

- **Current:**
  - 2023-26: NSF Collaborative research: IDEAS lab: ETAUS: Smarter Microbial Observatories for Realtime ExperimentS (SMORES) (award # 2321652) - joint with Harvard University, University of Georgia and Florida International University
- **Recently completed:**
  - 2020-22: US Military Academy/DoD “Self-learning capabilities for a mission oriented data quality and security assurance in military IoT systems” (award # W911NF2010337)
  - 2021-22: CRDF Global/DoState “Security evaluation and improvement of the personal infrastructure with new tools and education development” (award # G-202102-67515)
  - 2016-21: NSF “CICI: Data Provenance: Data quality and security evaluation framework for mobile devices platform” (award # ACI-1547301)
- **Software and data produced:**
  - Six Android security and data quality apps, 2018-22 – available on Google Play
  - Data collections of electronic sensors incorporated into mobile devices available on the market with their quality evaluations – available at http://www.dataqualitylabs.com/dataView

http://www.dataqualitylabs.com/home

Why Federated Learning?



Addressing the needs and challenges with FL

- FL was initially introduced by Google in 2016 to improve keyboard predictions on user smartphones. The **challenges** were:
  - Data had to be collected from billions of devices
  - Data on users’ mobile devices was private and could not be shared
  - Transmitting large amounts of data required excessive communication resources
  - Training of ML model on a large dataset in a centralized manner required excessive computational recourses
- FL allowed to **address all these challenges** as it:
  - Employs ML models training on the user mobile devices (aka local units)
  - Transmits only the ML models over communication channel, user’s data kept private
  - Employs ML models aggregation procedure, which allows to integrate multiple models into a single global model using a specialized aggregation function

Considered industrial Machine Learning (ML) applications and their needs and challenges

Case 1: Intelligent Transportation Systems (ITS)[1]

- Employ ML-based systems for events classification, obstacles detection, localization, path planning, etc.

Needs:

- The **need** to maintain high ML robustness to ensure road users safety

  Challenge: require to collect comprehensive and diverse data, produced by various data sources, on road objects/events and employ appropriate training procedures

- The **need** to ensure ITS users privacy

  Challenge: require to preserve privacy of the ITS users and avoid data leaks and confidentiality violations

[1] Chuprov, S., Bhatt, K.M., & Reznik, L. (2023). “Federated Learning for Robust Computer Vision in Intelligent Transportation Systems” in 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 26-27, doi: 10.1109/CAI54212.2023.00019.

Considered industrial ML applications and their needs and challenges

Case 2: Financial organizations (e.g., banks)[1]

- Employ ML-based systems for fraud and malicious activity detection, credit assessment, stock trading, etc.

Needs:

- The **need** to maintain high ML robustness to reduce financial and reputational losses and risks

  Challenge: require to collect comprehensive and diverse data, provided by multiple financial organizations, which might violate their confidentiality

- The **need** to ensure training data privacy

  Challenge: require to preserve privacy of the bank customers and avoid data leaks and confidentiality violations

$

[1] Chuprov, S., Memon, M., & Reznik, L. (2023). "Federated Learning with Trust Evaluation for Industrial Applications" in 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 347-348, doi: 10.1109/CAI54212.2023.00153.

Considered industrial ML applications and their needs and challenges

Case 3: Medical and healthcare organizations (e.g., hospitals)[1]

- Employ ML-based systems for patients diagnosing, drug prescription, planning personalized treatments, etc.

Needs:

- The **need** to maintain high ML robustness to reduce the risk of false diagnosing and provide better treatment

  Challenge: require to collect comprehensive and diverse data, provided by multiple medical organizations, which might violate their confidentiality

- The **need** to ensure training data privacy

  Challenge: require to preserve privacy of the patients and avoid data leaks and confidentiality violations

[1] Chuprov, S., Satam, A. N., & Reznik, L. (2022). ``Are ML Image Classifiers Robust to Medical Image Quality Degradation?'' in 2022 IEEE Western New York Image and Signal Processing Workshop (WNYISPW), 2022, pp. 1-4, doi: 10.1109/WNYISPW57858.2022.9983488.

Quick self-introduction

Bio: PhD student in Computing and Information Sciences at the Department of Computer Science at RIT (**expecting to graduate in Spring 2024**)

Advisor: Dr. Leon Reznik, Professor of Computer Science (RIT)

Research Interests: security of AI- and ML-based systems, Data Quality and security assurance, robustness of ML-end systems with integrated network facilities

Sergei Chuprov

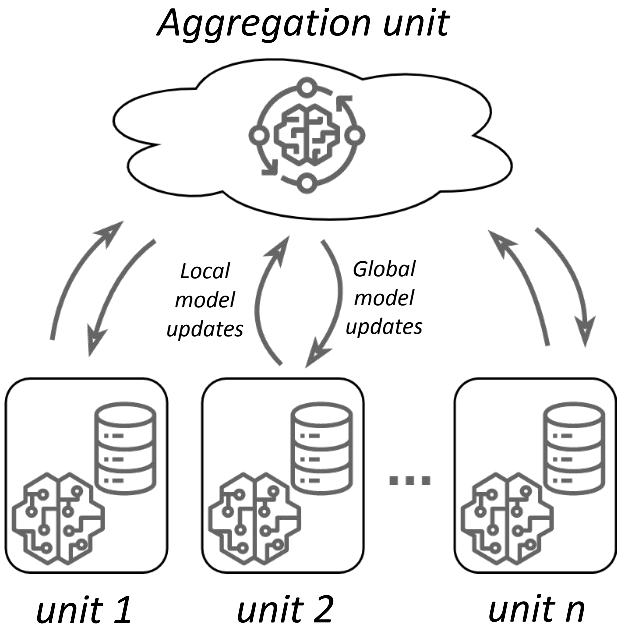
https://people.rit.edu/sc1723

What is FL?

Major stages of the Aggregation round

1. Local training on each local unit
2. Local updates transmission to the aggregation unit
3. Aggregation to produce global model
4. Global updates distribution back to the local units

Aggregation rounds are repeated until the termination criteria is met



FL Summary

- **FL Objectives:**
  - Privacy enhancement by keeping the data safe on local units
  - Required training time and computational resources decrease by distributing training procedure between local units
- **FL Challenges:**
  - Maintaining secure and reliable communication
  - Dealing with non-i.i.d. data
  - Dealing with malicious local units

Aspects we cover in our talk

- We investigate the **feasibility of FL employment** in multiple industrial scenarios
- We provide our recommendations for **training ML models more robust to Data Quality (DQ) variations in real scenarios**
- We investigate the conventional FL **methodological vulnerability** that might be easily exploited by a compromised local units and jeopardizes FL privacy
- We develop **FLAME**, which includes methods and software solutions to enhance ML robustness and to mitigate the FL methodological vulnerability by detecting compromised local units and excluding them from FL
- Our developments and solutions are reflected in our patent application[1], and we are **actively looking for a partnership and collaboration**

[1]: Chuprov, S., & Reznik, L. ``Federated Learning with A Compromised Unit Exclusion from Receiving Global Model Updates''. Provisional application filed on January 13, 2023

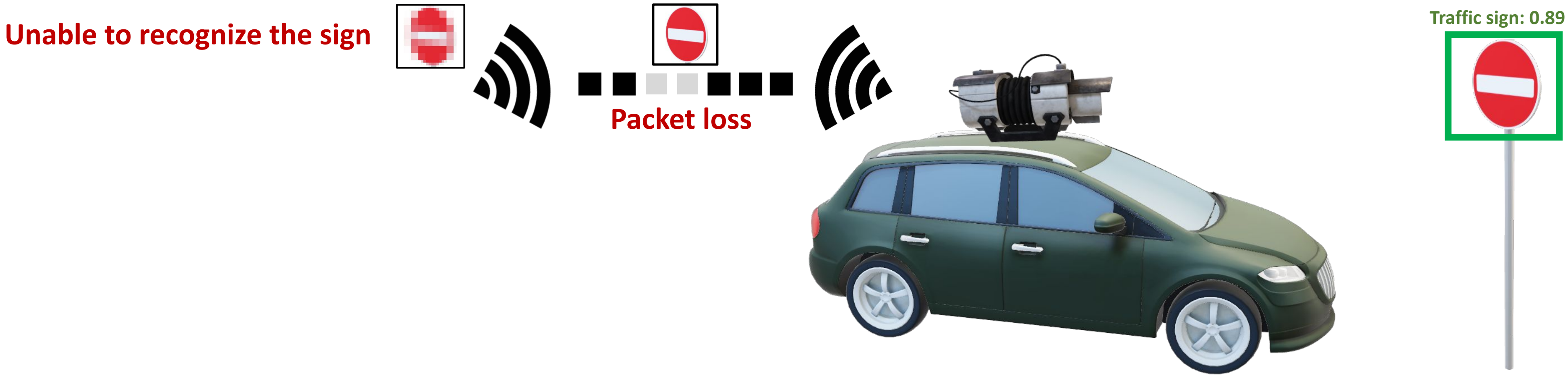
Outline

FLAME for enhancing ML robustness in industrial applications.

ITS use case

Why ML robustness matters?

- Robustness to Data Quality (DQ) variations is a severe challenge for real-time ML applications, especially those implemented in industry



How to enhance ML robustness in execution?

Two major types of approaches:

- **Reactive approach**
  - Aimed at monitoring ML performance during the system execution and apply measures post-factum (e.g., data cleaning, data restoration, model re-training)
- **Proactive approach**
  - Aimed at applying initial measures at the ML training phase to enhance robustness during ML execution (e.g., adding data of lower quality into the training set, adversarial training)

Our investigated approach

Proactive approach:

- Transfer Learning (TL)
- FL

Industrial domain:

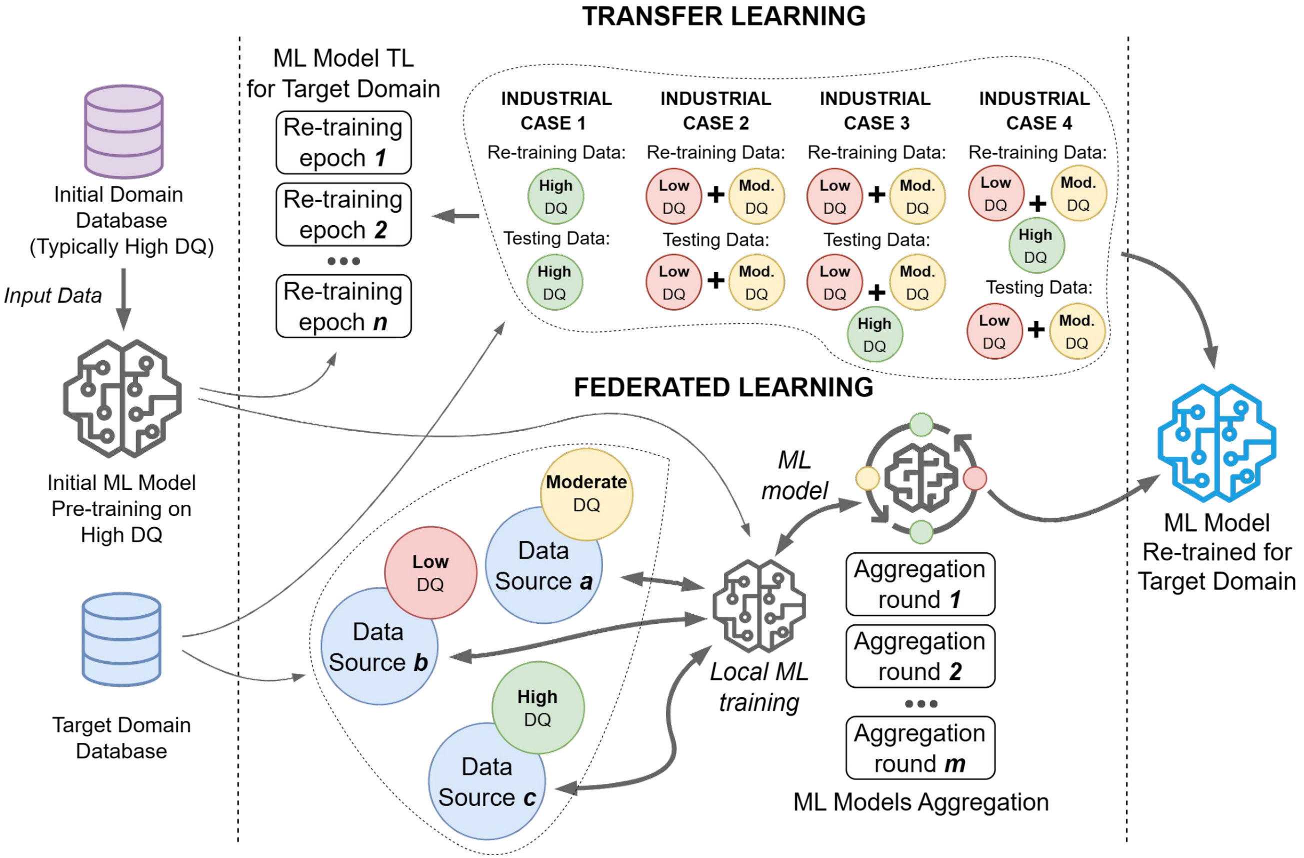
- Intelligent Transportation System (ITS)

Data:

- Stop and traffic sign images of varied DQ

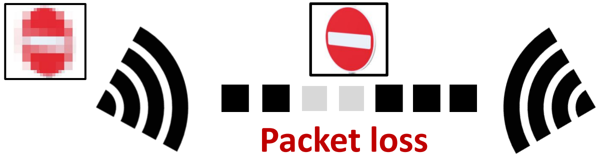
ML model:

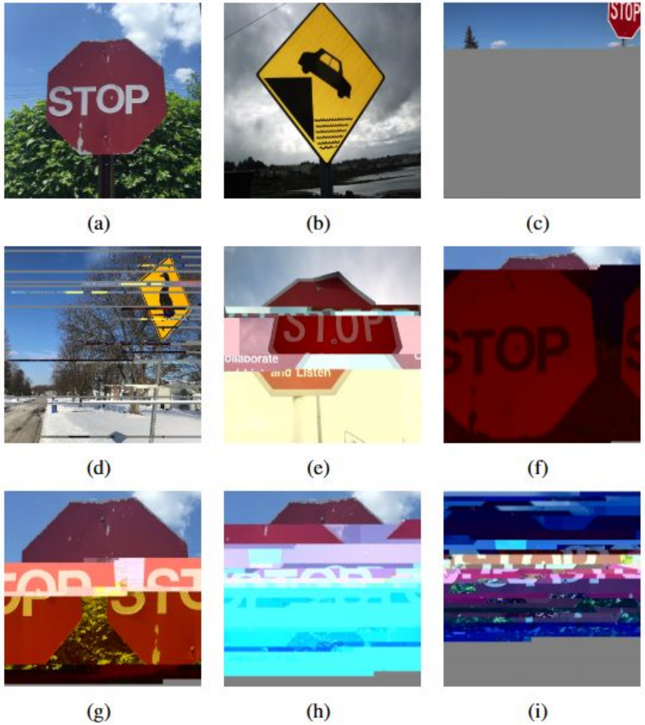
- Pre-trained VGG16



Collecting data with real corruptions

- To obtain data with corruptions caused be real network QoS degradation, we employed **POWDER** platform
- We established wireless network topology between two nodes and transmitted images with various **packet loss** rates and **buffer size** resources

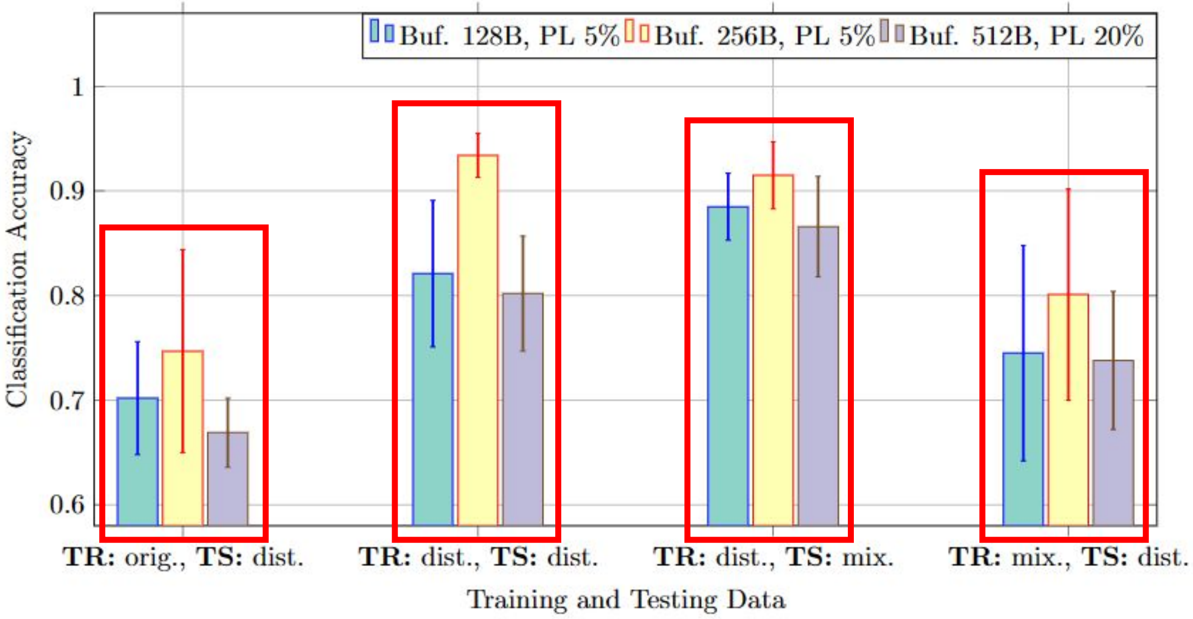




(a) - original stop sign image;
(b) - original traffic sign image;
(c) - the resulting image after the buffer overflow on the receiving end.
The image transmitted with:
(d) - 1024B buffer and 2% packet loss;
(e) - 2048B buffer and 5% packet loss.
Stop sign image transmitted with 512B buffer and:
(f) - 1% packet loss;
(g) - 5% packet loss;
(h) - 10% packet loss;
(i) - 20% packet loss

TL results[1]

- **Baseline:**
  - Re-trained on **original** images, tested on **distorted** images
- **Case 1:**
  - Re-trained on **distorted** images, tested on **distorted** images
- **Case 2:**
  - Re-trained on **distorted** images, tested on **mixed** images
- **Case 3:**
  - Re-trained on **mixed** images, tested on **distorted** images

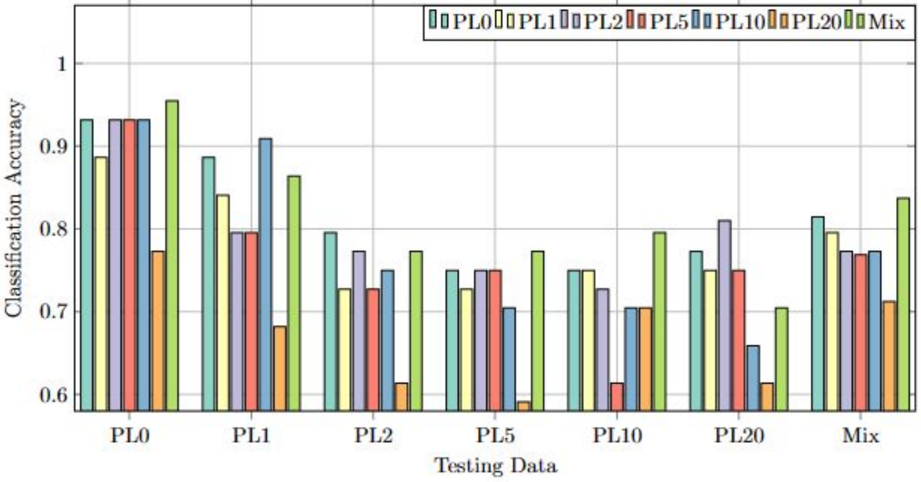


[1] Chuprov, S., Bhatt, K.M., & Reznik, L. (2023). "Federated Learning for Robust Computer Vision in Intelligent Transportation Systems" in 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 26-27, doi: 10.1109/CAI54212.2023.00019.

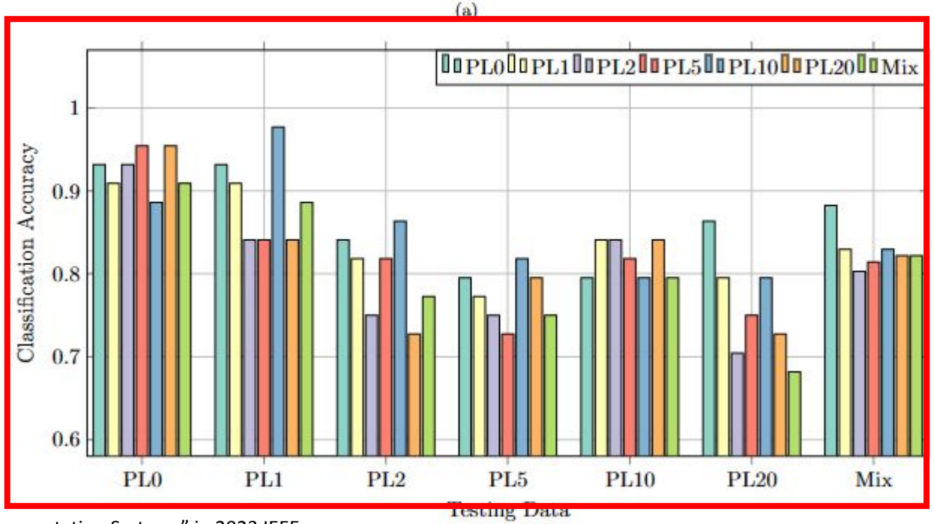
[2] Chuprov, S., Khokhlov, I., Reznik, L., & Shetty, S. (2022). ``Influence of Transfer Learning on Machine Learning Systems Robustness to Data Quality Degradation'' in 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9892247.

FL results[1]

- We evaluated two distinct FL aggregation strategies:
  - **FedAvg** (fig. a);
  - **Geometric Median (GM)** (fig. b)
- We employed 10 local units, which possessed training data of various quality, affected by packet loss rates of 1, 2, 5, 10, and 20%.
- We cross-evaluated the performance of multiple training and testing cohorts:
  - Trained on **original** images, tested on **1, 2, 5, 10, 20% packet losses, and Mixed** image quality
  - Trained on **1% packet losses**, tested on **original, 2, 5, 10, 20, and Mixed** image quality



(a)



(b)

[1] Chuprov, S., Bhatt, K.M., & Reznik, L. (2023). "Federated Learning for Robust Computer Vision in Intelligent Transportation Systems" in 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 26-27, doi: 10.1109/CAI54212.2023.00019.

Advantages of employing TL and FL to enhance ML industrial applications robustness

- TL allowed to enhance ML performance demonstrated on the data of varied quality, which makes it feasible to employ
- FL demonstrated comparable ML performance to TL but allowed to enhance security and privacy of ITS users
- TL can be employed in a combination with FL, as the initial training procedure instead of training ML model from scratch

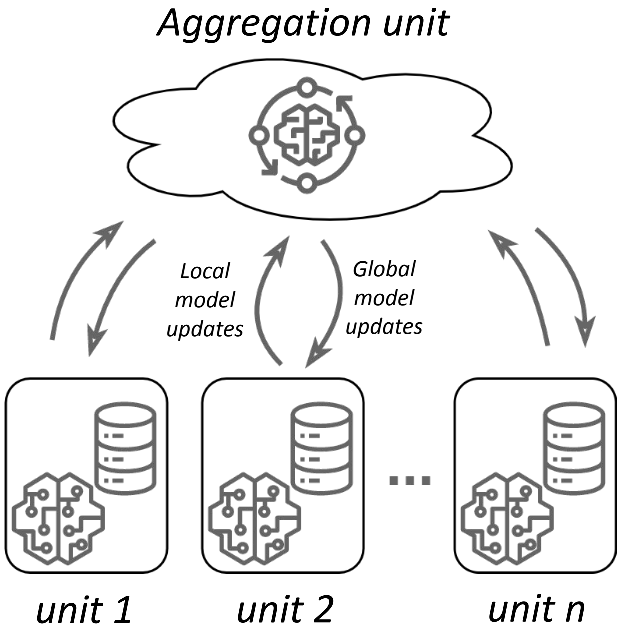
Outline

Existing vulnerabilities in the conventional FL architecture

Conventional FL vulnerability

Major stages of the aggregation round

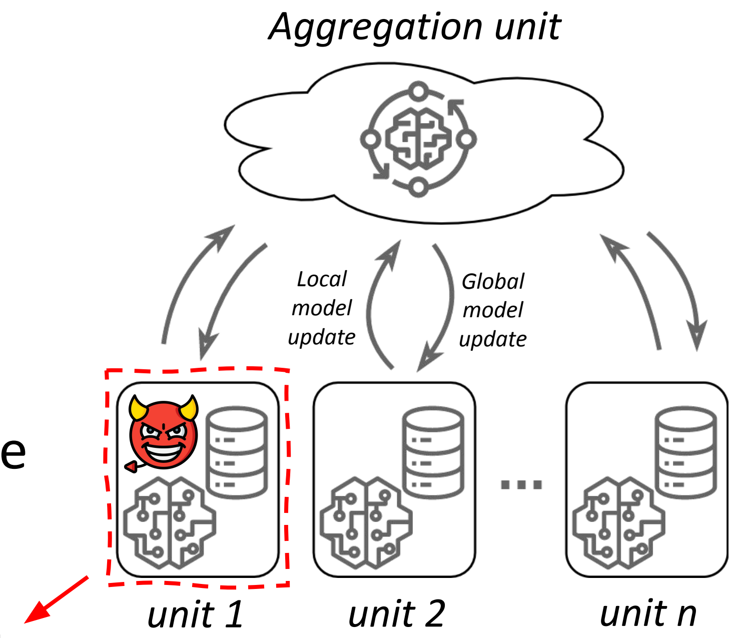
1. Local training on each local unit
2. Local updates transmission to the aggregation unit
3. Aggregation to produce global model
4. Global updates distribution back to the local units



Conventional FL vulnerability

Major stages of the aggregation round

1. Local training on each local unit
2. Local updates transmission to the aggregation unit
3. Aggregation to produce global model
4. Global updates distribution back to the local units

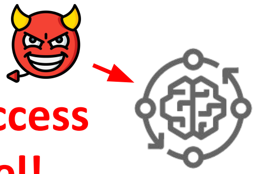


Compromised local unit

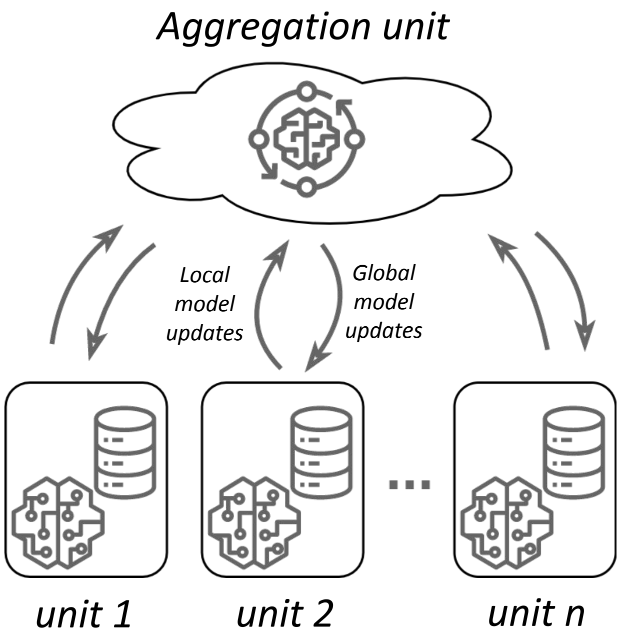
Conventional FL vulnerability

Major stages of the aggregation round

1. Local training on each local unit
2. Local updates transmission to the aggregation unit
3. Aggregation to produce global model
4. **Global updates distribution back to the local units**

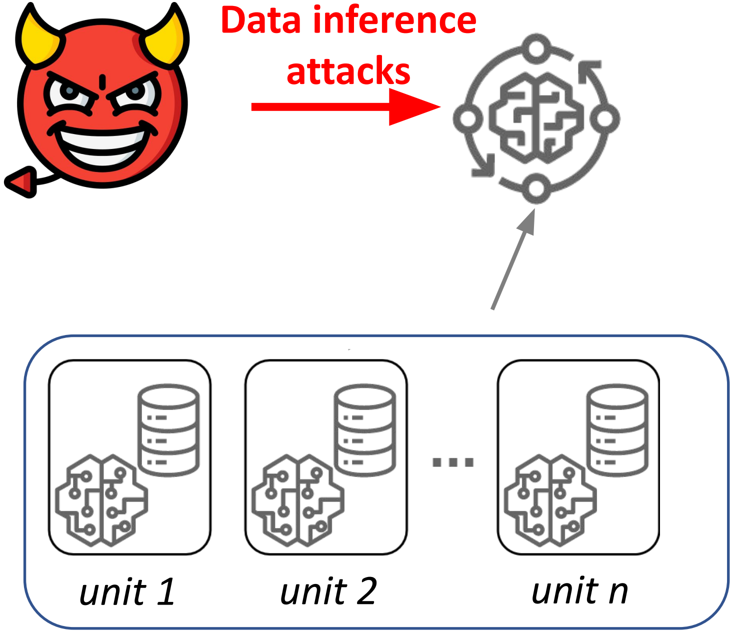


Adversary gains access to the global model!



Conventional FL vulnerability

- Global model is aggregated based on the local models from all the units participated in the aggregation round
- Gaining access to the global model, the adversary can implement **data inference attacks** on the global model:
  - Membership inference[1]
  - Training data properties inference[2]
  - Training samples and labels inference[3]
- **Result: privacy violation of the local units participated in the aggregation round!**



[1] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In SP, pages 739–753, 2019

[2] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In SP, pages 691–706, 2019

[3] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In NeurIPS, pages 14747–14756, 2019; Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. CoRR, arXiv:2001.02610, 2020

Outline

FLAME for enhancing FL security by embedding trust evaluation indicators.

Financial use case

How to Enhance FL security?

Conventional FL **possesses** certain vulnerabilities that may jeopardize user’s security and privacy

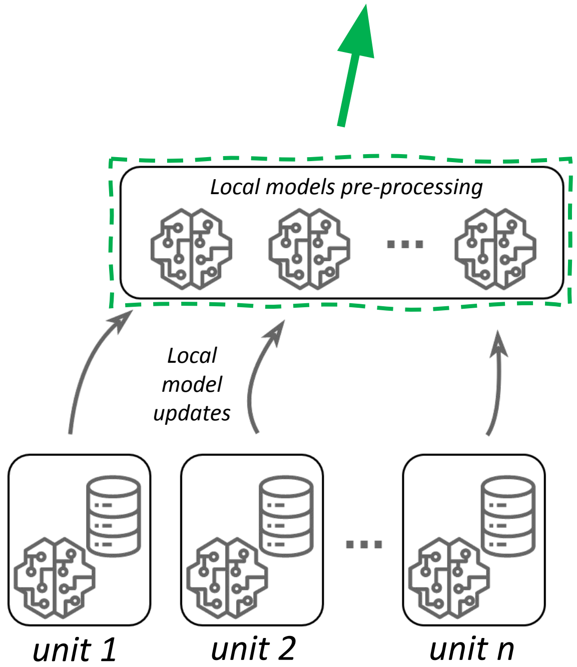
Our solution: to employ **Reputation and Trust**-based mechanisms[1,2] to accumulate the extracted knowledge on the quality of local units’ models and employ it to drive the aggregation procedure

[1] Chuprov, S., Viksnin, I., Kim, I., Marinenkov, E., Usova, M., Lazarev, E., ... & Zakoldaev, D. (2019). Reputation and trust approach for security and safety assurance in intersection management system. Energies, 12(23), 4527.

[2] Chuprov, S., Viksnin, I., Kim, I., Reznikand, L., & Khokhlov, I. (2020, July). Reputation and trust models with data quality metrics for improving autonomous vehicles traffic security and safety. In 2020 IEEE systems security symposium (SSS) (pp. 1-8). IEEE.

How to evaluate Reputation and Trust indicators?

1. Local updates pre-processing before the aggregation procedure



Problem with clusterinh: too strict and leads to high False Positive rates

Solution: accumulate the data on models provided with **Reputation and Trust**

Local model’s parameters clustering

If the model’s parameters lie out of the major cluster – the unit might be **potentially compromised**

Exclude potentially compromised local units from global model’s distribution

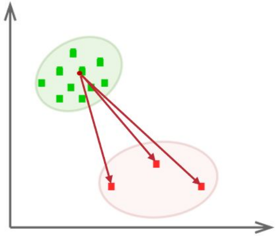
How to evaluate Reputation and Trust indicators?

We introduce two basic indicators:

1. **Reputation ($\boldsymbol{R}$)** – an indicator, which accumulates the *quality of the models* provided by the local units from the FL system initialization moment. Better quality → better R, and vice versa
2. **Trust** – an indicator, which regulates *how sensitive* is the trust to the changes in the local units' R

How to evaluate Reputation and Trust indicators?

1. Calculate normalized $L2$ the distance (d) from the center of the major ML models parameters' cluster to all the models
2. Calculate the initial local unit's Reputation (R) as $R = 1 - d$
3. In the next aggregation round, calculate R as:



$$R = \begin{cases} \left(R_i^{t-1} + \ d_i\right) - \left(\dfrac{R_i^{t-1}}{t}\right), \text{if } d \geq \alpha, R \in [0,1] \\ \left(R_i^{t-1} + \ d_i\right) - e^{-\left(1-d\left(\frac{R_i^{t-1}}{t}\right)\right)}, \ \text{if } d < \alpha, R \in [0,1] \end{cases}$$

How to evaluate Reputation and Trust indicators?

4. Evaluate $Trust$ based on R as

$$Trust \ = \ \sqrt{\{(R^t)^2 + \ d^2\}} - \ \sqrt{\{(1 \ - \ R^t)^2 + \ (1 \ - \ d)^2\}},$$

$$Trust = \begin{cases} 1, \text{if } Trust \geq 1, \\ 0, \text{if } Trust \ \leq 0 \end{cases}$$

5. Compare $Trust$ of each local unit, participating in the aggregation round, to the established threshold β, and discard the units whose $Trust < \beta$ from the aggregation, global model distribution, and further communication

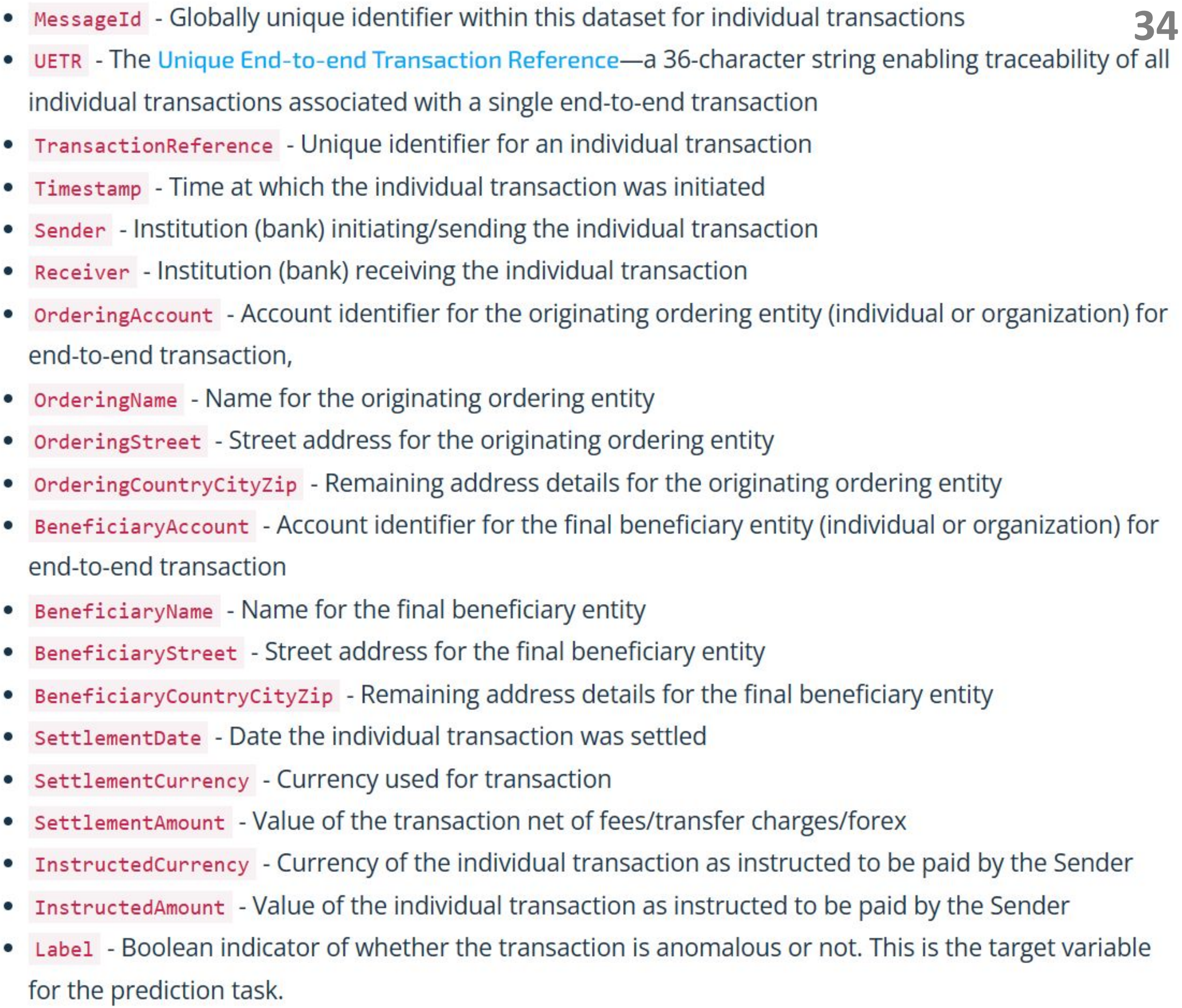
Financial use case study design

- **Use case:** fraud and suspicious transactions detection in SWIFT records
- We employed **industrial data provided by SWIFT**[1] on synthetic financial transactions, which was used in terms of **U.S. PETs Prize Challenge**[2] hosted by **NIST** and **NSF**
- Real industrial data with **tangled structural relationships** that are **hard to comprehend** by the state-of-the-art ML models (around 70% baseline performance)
- **4 million records** on financial transactions between various banks performed over a **30 days interval**
- More than 20 various attributes
- Classification target: **anomalous transactions**

1: https://www.swift.com/
2: https://www.drivendata.org/competitions/98/nist-federated-learning-1/page/522/

As the dataset is highly imbalanced, we employed the following pre-processing methods:

- Feature engineered two additional attributes:
  - *sender_currency_frequency*
  - *sender_currency_amount_average*
- Employed SMOTE library to avoid oversampling



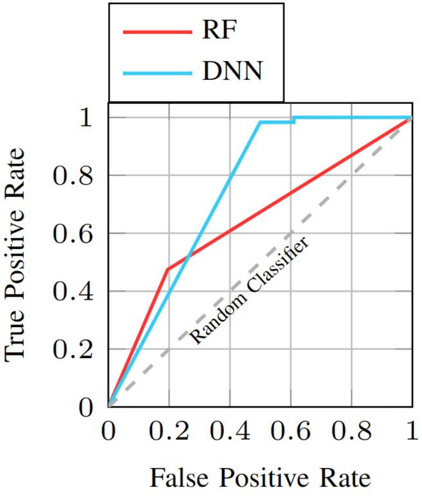
- `MessageId` - Globally unique identifier within this dataset for individual transactions
- `UETR` - The Unique End-to-end Transaction Reference—a 36-character string enabling traceability of all individual transactions associated with a single end-to-end transaction
- `TransactionReference` - Unique identifier for an individual transaction
- `Timestamp` - Time at which the individual transaction was initiated
- `Sender` - Institution (bank) initiating/sending the individual transaction
- `Receiver` - Institution (bank) receiving the individual transaction
- `OrderingAccount` - Account identifier for the originating ordering entity (individual or organization) for end-to-end transaction,
- `OrderingName` - Name for the originating ordering entity
- `OrderingStreet` - Street address for the originating ordering entity
- `OrderingCountryCityZip` - Remaining address details for the originating ordering entity
- `BeneficiaryAccount` - Account identifier for the final beneficiary entity (individual or organization) for end-to-end transaction
- `BeneficiaryName` - Name for the final beneficiary entity
- `BeneficiaryStreet` - Street address for the final beneficiary entity
- `BeneficiaryCountryCityZip` - Remaining address details for the final beneficiary entity
- `SettlementDate` - Date the individual transaction was settled
- `SettlementCurrency` - Currency used for transaction
- `SettlementAmount` - Value of the transaction net of fees/transfer charges/forex
- `InstructedCurrency` - Currency of the individual transaction as instructed to be paid by the Sender
- `InstructedAmount` - Value of the individual transaction as instructed to be paid by the Sender
- `Label` - Boolean indicator of whether the transaction is anomalous or not. This is the target variable for the prediction task.

Financial use case study results[1]

1. Evaluating the models trained in a centralized manner:

Random Forest (RF): ~63% AUC

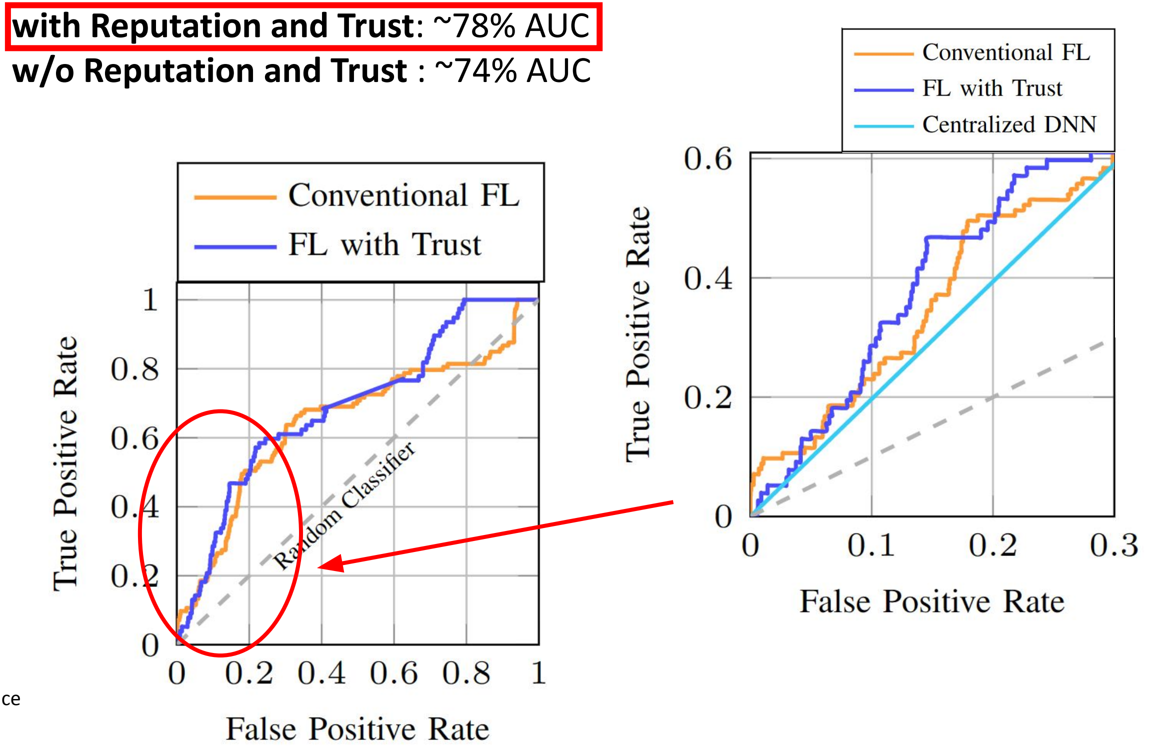
Deep Neural Network (DNN): ~74% AUC



2. Evaluating the **DNN** trained in a FL manner:

with Reputation and Trust: ~78% AUC

w/o Reputation and Trust : ~74% AUC



[1] Chuprov, S., Memon, M., & Reznik, L. (2023). "Federated Learning with Trust Evaluation for Industrial Applications" in 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 347-348, doi: 10.1109/CAI54212.2023.00153.

Advantages and limitations

- **Advantages** of the Reputation and Trust-based approach:
  - Allows to track how the quality of models provided by the local units changes throughout the FL process
  - Allows to detect intelligent adversaries who implement non-obvious attack strategies (e.g., on-off attacks)
  - Accumulating of knowledge is more reliable than one-time local model's evaluation
- **Assumptions and limitations:** requires active adversary that performs data manipulations, which can be detected by our Reputation and Trust-based approach

Our major products developed

- We developed Reputation and Trust methods that allow enhancing ML application security and robustness to DQ variations
- We implemented our methods as software prototypes, which we effectively verified in multiple industrial use cases
- Our developments found their realization in the following products:
    - **2 research papers:**
        - Chuprov, S., Bhatt, K.M., & Reznik, L. (2023). “Federated Learning for Robust Computer Vision in Intelligent Transportation Systems” in 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 26-27, doi: 10.1109/CAI54212.2023.00019.
        - Chuprov, S., Memon, M., & Reznik, L. (2023). “Federated Learning with Trust Evaluation for Industrial Applications” in 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 347-348, doi: 10.1109/CAI54212.2023.00153.
    - **1 patent application:**
        - Chuprov, S., & Reznik, L. ``Federated Learning with A Compromised Unit Exclusion from Receiving Global Model Updates''. Provisional application filed on January 13, 2023
    - **Software prototypes that realize our solutions**
- We are **open for collaboration, please, reach out to us if you have industrial ML applications that require to be more robust and secure!**

Rochester Security Summit 2023

October 25th and 26th, 2023

Thank you!

Sergei Chuprov

Dr. Leon Reznik

B. Thomas Golisano College of Computing and Information Sciences,

Rochester Institute of Technology

sc1723@rit.edu, leon.reznik@rit.edu