

REPORT DOCUMENTATION PAGE			Form Approved OMB NO. 0704-0188		
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 08-04-2023		2. REPORT TYPE Final Report		3. DATES COVERED (From - To) 7-Aug-2017 - 6-Jan-2023	
4. TITLE AND SUBTITLE Final Report: Data Analytics for Cyber Security: Defeating the Active Adversaries			5a. CONTRACT NUMBER W911NF-17-1-0356		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBER 611102		
6. AUTHORS			5d. PROJECT NUMBER		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAMES AND ADDRESSES University of Texas at Dallas 800 West Campbell Road, AD15 Richardson, TX 75080 -3021			8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211			10. SPONSOR/MONITOR'S ACRONYM(S) ARO		
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) 69823-NC.18		
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Murat Kantarcioglu
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU			19b. TELEPHONE NUMBER 972-883-6616

RPPR Final Report

as of 11-Apr-2023

Agency Code: 21XD

Proposal Number: 69823NC

Agreement Number: W911NF-17-1-0356

INVESTIGATOR(S):

Name: Murat Kantarcioglu
Email: muratk@utdallas.edu
Phone Number: 9728836616
Principal: Y

Organization: **University of Texas at Dallas**

Address: 800 West Campbell Road, AD15, Richardson, TX 750803021

Country: USA

DUNS Number: 800188161

EIN: 751305566

Report Date: 06-Apr-2023

Date Received: 08-Apr-2023

Final Report for Period Beginning 07-Aug-2017 and Ending 06-Jan-2023

Title: Data Analytics for Cyber Security: Defeating the Active Adversaries

Begin Performance Period: 07-Aug-2017

End Performance Period: 06-Jan-2023

Report Term: 0-Other

Submitted By: Murat Kantarcioglu

Email: muratk@utdallas.edu

Phone: (972) 883-6616

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 4

STEM Participants:

Major Goals: The overall arching goal of this project is to develop techniques to increase the robustness of machine learning systems against adversarial attacks. In addition, we focus on understanding how the developed techniques could be leveraged in practice so that humans and ML tools could work together.

Accomplishments: Please see the attached report and slides.

Training Opportunities: We have organized a reading group on deep learning and adversarial machine learning topics allowing students to learn the recent advances in this area.

In addition, Dr. Kantarcioglu thought a course on adversarial machine learning.

The students who participated in this research had chance to learn cutting edge technologies with respect to deep learning, adversarial machine learning, IoT data, and explainable AI.

Results Dissemination: The results are disseminated via publishing our work in major conferences and venues. In addition, PI Kantarcioglu and co-PIs Thurasingham and Xi gave many talks on the topic.

Co-PI Thuraisingham participated in Women in Data Science event held Stanford University that was live streamed to around 100,000 people. She gave an overview of our Adversarial Machine Learning Research for ARO project.

RPPR Final Report as of 11-Apr-2023

Honors and Awards: Dr. Kantarcioglu received the IEEE Intelligence and Security Informatics Society Technical Achievement Award for his contributions to data security and privacy.

Dr. Xi's publication on pancreatic cancer biomarker detection is Editor Choice Articles in Current Metabolomics, 2018. A US patent application is filed.

Dr. Thuraisingham received 2017 CODASPY Lasting Research Award, presented at the Association of Computing Machinery's Conference on Data and Applications Security and Privacy (ACM CODASPY)

Dr. Thuraisingham received Dallas Business Journal 2017 Women in Technology Awards.

Dr. Kantarcioglu and Dr. Thuraisingham have received jointly: ACM SACMAT 2019 10 Year Best Paper /Test of Time Award for their work on social network access control.

Please see the press release below for more details: <https://cs.utdallas.edu/drs-kantarcioglu-and-thuraisingham-presented-with-the-sacmat-test-of-time-award/>

Dr. Thuraisingham has received: 2018 IEEE ICDM Service Award for Exceptional Service and Leadership to Community in promoting Data Mining to the Cyber Security Community

Dr. Thuraisingham has also received Top 25 Women in Cyber Security, Cyber Defense Magazine, 2019 <http://cyberdefenseawards.com/top-25-women-in-cybersecurity/>

Dr. Thuraisingham has become ACM Fellow.

Dr. Kantarcioglu has been elected to 2020 fellows of the prestigious American Association for the Advancement of Science (AAAS).

Dr. Thuraisingham named by the D Magazine's D CEO Dallas 500 to be one of the 500 most influential leaders in North Texas in 2021.

Dr. Kantarcioglu has been elected to become IEEE Fellow.

Protocol Activity Status:

Technology Transfer: PIs Kantarcioglu and Thuraisingham are actively working with ARL south researchers. Kantarcioglu already had a paper with Jonathan Bakdash from ARL South.

Co-PI Xi has started collaboration as a part of ARL Central.

In addition, Dr. Kantarcioglu is collaborating with Dr. Charles Kamhoua to conduct joint research.

PI Kantarcioglu gave a talk to summarize some of our research results to researchers led by Dr Nathaniel D. Bastion, Senior Research Scientist at Intelligent Cyber-Systems and Analytics Research Lab, Army Cyber Institute, U.S. Military Academy, West Point, NY.

PARTICIPANTS:

Participant Type: Co PD/PI

Participant: Bowei Xi

Person Months Worked: 4.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: PD/PI

Participant: Murat Kantarcioglu

RPPR Final Report
as of 11-Apr-2023

Person Months Worked: 5.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Co PD/PI
Participant: Bhavani Thuraisingham
Person Months Worked: 1.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)
Participant: Yan Zhou
Person Months Worked: 15.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Graduate Student (research assistant)
Participant: Lynda Nguyen
Person Months Worked: 9.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Graduate Student (research assistant)
Participant: Imrul Anindya
Person Months Worked: 2.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Graduate Student (research assistant)
Participant: Vibha Belavadi Chandramouli
Person Months Worked: 10.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Graduate Student (research assistant)
Participant: Kirsan Sullivan
Person Months Worked: 10.00
Project Contribution:
National Academy Member: N

Funding Support:

ARTICLES:

RPPR Final Report as of 11-Apr-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: EURASIP Journal on Image and Video Processing
Publication Identifier Type: DOI **Publication Identifier:** <https://doi.org/10.1186/s13640-018-0259-9>
Volume: **Issue:** **First Page #:**
Date Submitted: 10/10/18 12:00AM **Date Published:** 3/24/18 5:00AM
Publication Location:
Article Title: A neighborhood regression approach for removing multiple types of noises
Authors: Walid Sharabati, Bowei Xi
Keywords: Image denoising, Neighborhood L 2 regression filter, Poisson noise, Stripped noise Salt & pepper noise Gaussian noise
Abstract: Image denoising is an important first step to provide cleaned images for follow-up tasks such as image segmentation and object recognition. Many image denoising filters have been proposed, with most of the filters focusing on one particular type of additive or multiplicative noise. In this article, we propose a novel neighborhood regression approach. Using the neighboring pixels as predictors, our approach has superb performance over multiple types of noises, including Gaussian, Poisson, Gaussian and Poisson, salt & pepper, and stripped noise. Our L2 regression filter can be parallelized to significantly speed up the denoising process to process a large number of noisy images. Meanwhile, our regression approach does not need tuning parameters or any training images, and it does not need any prior knowledge of the variance of the noise. Instead, our regression filter can accurately estimate the variance of the added Gaussian noise.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: Y

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Computers & Security
Publication Identifier Type: DOI **Publication Identifier:** <https://doi.org/10.1016/j.cose.2017.11.018>
Volume: **Issue:** **First Page #:**
Date Submitted: 10/10/18 12:00AM **Date Published:** 8/1/18 5:00AM
Publication Location:
Article Title: Spatial counts under differential privacy mechanism on changing spatial scales
Authors: JunJianga, Bowei Xi, Murat Kantarcioglu
Keywords: Spatial countsDifferential privacyPoisson-Binomial distributionData generationTransformation invariance
Abstract: With a spatial statistical database covering a large region, how to publish differential privacy protected information is a challenge. In previous works, information was published using large fixed spatial cells. In this paper, we develop novel flexible methods to publish the spatial information, which allows the users to freely move around the large region, zoom in and zoom out at arbitrary locations, and obtain information over spatial areas both large and small. We develop two methods to publish the spatial information protected under differential privacy. First the region is divided into the smallest spatial cells, where each cell does not observe an event happening more than once. Given repeated measurements, such as multiple day data, the noise added Bernoulli probabilities are computed for all the smallest spatial cells. For larger spatial cells of high interests to users, the noisy Bernoulli probabilities are combined into noisy Poisson-Binomial distributions which also satisfy
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: Y

RPPR Final Report as of 11-Apr-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published

Journal: IEEE transaction on Dependable and Secure Computing

Publication Identifier Type: DOI

Publication Identifier: 10.1109/TDSC.2021.3127439

Volume: 20

Issue: 1

First Page #: 83

Date Submitted: 3/30/23 12:00AM

Date Published: 1/31/23 6:00AM

Publication Location:

Article Title: Exploring the Effect of Randomness on Transferability of Adversarial Samples Against Deep Neural Networks

Authors: Yan Zhou, Murat Kantarcioglu, Bowei Xi

Keywords: Robust DNN, Randomization

Abstract: We investigate the transferability of adversarial attacks against deep neural networks (DNNs)—the contagion effect of adversarial attacks that, once deceiving one DNN model, can easily deceive other DNN models built on similar data. We demonstrate that introducing randomness to DNN models can break the curse of the transferability of adversarial attacks, given that the adversary does not have an unlimited attack budget. Two randomization schemes are explored: 1.) a random selection—single or ensemble—from a set of DNNs is surprisingly more robust against the strongest form of complete-knowledge attacks (a.k.a, white box attacks); 2.) after a small Gaussian random noise is added to its learned weights, a DNN model can potentially increase its resilience to adversarial attacks by as much as 74.2%. We compare the two randomization techniques to the Ensemble Adversarial Training technique and show that our randomization techniques are superior under different attack budget constraints.

Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: Y

Publication Type: Journal Article

Peer Reviewed: N

Publication Status: 5-Submitted

Journal: arXiv.org

Publication Identifier Type: Other

Publication Identifier: <https://arxiv.org/abs/1804.04780>

Volume:

Issue:

First Page #:

Date Submitted: 10/10/18 12:00AM

Date Published: 4/13/18 5:00AM

Publication Location:

Article Title: Adversarial Clustering: A Grid Based Clustering Algorithm Against Active Adversaries

Authors: Wutao Wei, Bowei Xi, Murat Kantarcioglu

Keywords: Adversarial Clustering

Abstract: Nowadays more and more data are gathered for detecting and preventing cyber attacks. In cyber security applications, data analytics techniques have to deal with active adversaries that try to deceive the data analytics models and avoid being detected. The existence of such adversarial behavior motivates the development of robust and resilient adversarial learning techniques for various tasks. Most of the previous work focused on adversarial classification techniques, which assumed the existence of a reasonably large amount of carefully labeled data instances. However, in practice, labeling the data instances often requires costly and time-consuming human expertise and becomes a significant bottleneck. Meanwhile, a large number of unlabeled instances can also be used to understand the adversaries' behavior. To address the above mentioned challenges, in this paper, we develop a novel grid based adversarial clustering algorithm. Our adversarial clustering algorithm is able to identify the

Distribution Statement: 1-Approved for public release; distribution is unlimited.

Acknowledged Federal Support: Y

RPPR Final Report as of 11-Apr-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery
Publication Identifier Type: DOI **Publication Identifier:** 10.1002/widm.1259
Volume: 9 **Issue:** 3 **First Page #:**
Date Submitted: 10/23/19 12:00AM **Date Published:** 4/1/18 5:00AM
Publication Location:
Article Title: A survey of game theoretic approach for adversarial machine learning
Authors: Yan Zhou, Murat Kantarcioglu, Bowei Xi
Keywords: Adversarial Machine Learning
Abstract: In this work, we provided a survey of game theory inspired approaches for adversarial machine learning.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: Y

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: IEEE Transactions on Dependable and Secure Computing
Publication Identifier Type: DOI **Publication Identifier:** 10.1109/TDSC.2020.3019508
Volume: 18 **Issue:** 5 **First Page #:** 2061
Date Submitted: 3/30/23 12:00AM **Date Published:** 9/16/21 9:57PM
Publication Location:
Article Title: Robust Transparency Against Model Inversion Attacks
Authors: Yasmeen Alufaisan, Murat Kantarcioglu, Yan Zhou
Keywords: transparency , privacy-preserving , model-inversion attack
Abstract: Transparency has become a critical need in machine learning (ML) applications. Designing transparent ML models helps increase trust, ensure accountability, and scrutinize fairness. Some organizations may opt-out of transparency to protect individuals' privacy. Therefore, there is a great demand for transparency models that consider both privacy and security risks. Such transparency models can motivate organizations to improve their credibility by making the ML-based decision-making process comprehensible to end-users. Differential privacy (DP) provides an important technique to disclose information while protecting individual privacy. However, it has been shown that DP alone cannot prevent certain types of privacy attacks against disclosed ML models. DP with low values can provide high privacy guarantees, but may result in significantly weaker ML models in terms of accuracy. This raises the question whether we can disclose accurate transparent ML models while preserving privacy.
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: Y

RPPR Final Report as of 11-Apr-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published

Journal: Proceedings of the AAAI Conference on Artificial Intelligence

Publication Identifier Type: DOI

Publication Identifier: 10.1609/aaai.v35i10.17118

Volume: 35

Issue: 10

First Page #: 9268

Date Submitted: 3/30/23 12:00AM

Date Published: 5/1/21 5:00AM

Publication Location:

Article Title: Defending against Backdoors in Federated Learning with Robust Learning Rate

Authors: Mustafa Safa Ozdayi, Murat Kantarcioglu, Yulia R. Gel

Keywords: Adversarial learning

Abstract: We investigate the transferability of adversarial attacks against deep neural networks (DNNs)—the contagion effect of adversarial attacks that, once deceiving one DNN model, can easily deceive other DNN models built on similar data. We demonstrate that introducing randomness to DNN models can break the curse of the transferability of adversarial attacks, given that the adversary does not have an unlimited attack budget. Two randomization schemes are explored: 1.) a random selection—single or ensemble—from a set of DNNs is surprisingly more robust against the strongest form of complete-knowledge attacks (a.k.a, white box attacks); 2.) after a small Gaussian random noise is added to its learned weights, a DNN model can potentially increase its resilience to adversarial attacks by as much as 74.2%. We compare the two randomization techniques to the Ensemble Adversarial Training technique and show that our randomization techniques are superior under different attack budget constraints.

Distribution Statement: 1-Approved for public release; distribution is unlimited.

Acknowledged Federal Support: Y

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published

Journal: Proceedings of the AAAI Conference on Artificial Intelligence

Publication Identifier Type: DOI

Publication Identifier: 10.1609/aaai.v35i8.16819

Volume: 35

Issue: 8

First Page #: 6618

Date Submitted: 3/30/23 12:00AM

Date Published: 5/1/21 5:00AM

Publication Location:

Article Title: Does Explainable Artificial Intelligence Improve Human Decision-Making?

Authors: Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, Murat Kantarcioglu

Keywords: Explainable AI, Human Decision Making

Abstract: We investigate the transferability of adversarial attacks against deep neural networks (DNNs)—the contagion effect of adversarial attacks that, once deceiving one DNN model, can easily deceive other DNN models built on similar data. We demonstrate that introducing randomness to DNN models can break the curse of the transferability of adversarial attacks, given that the adversary does not have an unlimited attack budget. Two randomization schemes are explored: 1.) a random selection—single or ensemble—from a set of DNNs is surprisingly more robust against the strongest form of complete-knowledge attacks (a.k.a, white box attacks); 2.) after a small Gaussian random noise is added to its learned weights, a DNN model can potentially increase its resilience to adversarial attacks by as much as 74.2%. We compare the two randomization techniques to the Ensemble Adversarial Training technique and show that our randomization techniques are superior under different attack budget constraints.

Distribution Statement: 1-Approved for public release; distribution is unlimited.

Acknowledged Federal Support: Y

BOOKS:

Publication Type: Book Peer Reviewed: Y **Publication Status:** 1-Published

Publication Identifier Type:

Publication Identifier: <https://doi.org/10.2200/S00861ED1V01Y201806AIM039>

Book Edition:

Volume:

Publication Year: 2018

Date Received: 10-Oct-2018

Publication Location: <https://doi.org/10.2200/S00861ED1V01Y201806AIM039>

Publisher: Morgan & Claypool Publishers

Book Title: Adversarial Machine Learning

Authors: Yevgeniy Vorobeychik, Murat Kantarcioglu

Editor:

Acknowledged Federal Support: Y

RPPR Final Report
as of 11-Apr-2023

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 2018 IEEE International Conference on Information Reuse and Integration (IRI)
Date Received: 29-Oct-2019 Conference Date: 06-Jul-2018 Date Published: 06-Aug-2018
Conference Location: Salt Lake City, UT, USA
Paper Title: Adversarial Anomaly Detection Using Centroid-Based Clustering
Authors: Imrul Chowdhury Anindya, Murat Kantarcioglu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 24th ACM SACMAT Conference
Date Received: 29-Oct-2019 Conference Date: 04-Jun-2019 Date Published: 09-Jun-2019
Conference Location: Toronto ON, Canada
Paper Title: Securing Big Data: New Access Control Challenges and Approaches
Authors: Murat Kantarcioglu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 2018 IEEE International Conference on Intelligence and Security Informatics (ISI)
Date Received: 29-Oct-2019 Conference Date: 10-Nov-2018 Date Published: 12-Nov-2018
Conference Location: Miami, FL
Paper Title: Attacklets: Modeling High Dimensionality in Real World Cyberattacks
Authors: Cuneyt G. Akcora, Jonathan Z. Bakdash, Yulia R. Gel, Murat Kantarcioglu, Laura R. Marusich, Bhavani
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Pacific-Asia Conference on Knowledge Discovery and Data Mining
Date Received: 09-Nov-2020 Conference Date: 14-Jun-2020 Date Published: 19-Apr-2020
Conference Location: Macau
Paper Title: Adversarial Active Learning in the Presence of Weak and Malicious Oracles
Authors: Yan Zhou, Murat Kantarcioglu, Bowei Xi
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Spie Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications
Date Received: 09-Nov-2020 Conference Date: 27-Apr-2020 Date Published: 01-Jun-2020
Conference Location: Los Angeles, CA
Paper Title: Efficacy of defending deep neural networks against adversarial attacks with randomization
Authors: Yan Zhou, Bowei Xi, Murat Kantarcioglu
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: In Network and Distributed System Security Symposium (NDSS)
Date Received: 09-Nov-2020 Conference Date: 23-Feb-2020 Date Published:
Conference Location: San Diego, CA
Paper Title: Adversarial Classification Under Differential Privacy
Authors: Jairo Giraldo, Alvaro A. Cardenas, Murat Kantarcioglu, Jonathan Katz
Acknowledged Federal Support: **Y**

RPPR Final Report

as of 11-Apr-2023

Publication Type: Conference Paper or Presentation

Publication Status: 1-Published

Conference Name: CODASPY '21: Eleventh ACM Conference on Data and Application Security and Privacy

Date Received: 30-Mar-2023

Conference Date: 21-Apr-2021

Date Published:

Conference Location: Virtual Event USA

Paper Title: BlockFLA: Accountable Federated Learning via Hybrid Blockchain Architecture

Authors: Harsh Desai, Mustafa Ozdayi, Murat Kantarcioglu

Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation

Publication Status: 1-Published

Conference Name: Proceedings of the Workshop on Artificial Intelligence Safety

Date Received:

Conference Date: 24-Jul-2022

Date Published:

Conference Location: Vienna, Austria

Paper Title: Understanding Adversarial Examples Through Deep Neural Network's Classification Boundary and Uncertainty Regions

Authors: Juan Shu, Bowei Xi, Charles A. Kamhoua

Acknowledged Federal Support: **Y**

Partners

,

I certify that the information in the report is complete and accurate:

Signature: Murat Kantarcioglu

Signature Date: 4/8/23 12:16PM

Our project accomplishments

Overview

In this project, we focused on different aspects of adversarial attacks in machine learning and how to mitigate these attacks using various approaches. Our main accomplishments are summarized as follows:

- **Adversarial Unsupervised Learning:** [2] Unlike semi-supervised learning, the proposed grid based adversarial clustering algorithm, a.k.a. ADClust, does not label all the objects. Instead, ADClust identifies normal regions, abnormal regions, and the overlapping areas in between them inside large mixed clusters. While ADClust is able to identify the different regions inside a mixed cluster, semi-supervised learning algorithms completely fail to classify the normal and abnormal sub-clusters. ADClust keeps the objects in the overlapping areas unlabeled. Since adversaries actively place attack objects to move into normal regions, trying to find the best classification boundary in the heavily mixed overlapping areas is meaningless. The unlabeled overlapping areas identified by ADClust are analogous to confidence bands covering point estimates (i.e., classification boundaries).
- **Adversarial Active Learning:** [6] The adversarial active learning algorithm is another approach to utilize information from large amount of unlabeled data. Strategic sampling technique is used to select the next object to label by an oracle (e.g., a domain expert). Labeled objects are used as seeds to cluster the unlabeled data. Data clusters are used to estimate which are the genuine oracles, and which are the malicious and noisy oracles. Only the estimated genuine oracles are used to label the next object and to add it to the training set. The proposed adversarial active learning algorithm requires a much smaller training set to achieve accurate results compared with supervised learning. When one third of the oracles are genuine oracles, the proposed adversarial active learning algorithm needs only 100 labeled data points to be as accurate as the ideal active learning algorithm, which has no malicious and noisy oracles. When only one sixth are genuine oracles, the adversarial active learning algorithm shows up to 50 percent improvement over the state-of-art techniques.
- **Robust DNN Learning Using Randomization:** [7] Instead of making a single DNN model robust to perfect knowledge attacks (a.k.a, white box attacks), introducing randomness to an ensemble of DNNs largely eliminates the possibility of perfect knowledge acquisition, resulting in a significantly more robust DNN ensemble against the strongest form of attacks. We show that by simply adding small random noises to the learned weights of a DNN model, we can significantly improve the classification accuracy by as much as 74.2%, on condition that the adversary's goal is to misclassify a sample with minimum perturbation cost. We also show that when the adversary has an unlimited budget of data perturbation, all defensive techniques would eventually break down as the budget increases. Therefore, it is important to understand the game saddle point where the adversary would not further pursue this endeavor. Essentially, transferability is the observable trait of the severity of adversarial data perturbation. We show that by randomly selecting a DNN model from a pool of DNNs at each query request, including probing request and classification request, we can effectively mitigate transferability based adversarial attacks. Furthermore, we show that the ensemble of a set of DNN models can achieve nearly perfect defense against adversarial attacks when the attack budget is low. Even when the attack budget is high, the ensemble is significantly more robust than the Ensemble Adversarial Training technique. Our contributions can be summarized as follows: a.) we demonstrate

that transferability should not be automatically assumed for a learning task, and by holding on to a subset F of DNN models, we can defeat the adversary who has the perfect knowledge of a DNN model built with the same training data used to build F , but with different initialization points for stochastic gradient descent. Furthermore, we demonstrate that the ensemble of F can provide the strongest defense against severe attacks; b.) we demonstrate that by adding a small Gaussian random noise to the weights of a pre-trained DNN, we can significantly and efficiently increase the robustness of the DNN model; c.) we compare our techniques to the existing ensemble adversarial training technique and show that our proposed randomization approaches are more robust against the state-of-the-art attacks. d.) in addition, we attack our own randomization ensemble technique with white box attacks to demonstrate its robustness.

- **Adversarial Classification in Noisy IoBT Environments:** [4] The last decade has seen a growing interest in adversarial classification, where an attacker tries to mislead a classifier meant to detect anomalies. In this years research efforts, we study this problem in an Internet of Battle Things (IoBT) setting where anomaly detection is being used in a noisy environment. We showed that a strategic attacker can leverage the additional noise present to mislead the classifier beyond what the attacker could do otherwise. We also proposed countermeasures against such attacks. We then evaluate the impact of our attacks and defenses in real-world IoT settings. To illustrate the problem, we begin with an intuitive overview illustrating how additional noise can affect anomaly detection. Suppose we have a central entity monitoring a sequence of real values (e.g., sensor readings from an IoBT device) $(y(0), y(1), y(2), y(n))$. Assume further that, under normal conditions, it will always be the case that $|y(k-1) - y(k)| < c$ for some constant c , and so a simple form of anomaly detection is to raise an alarm if this condition is ever violated. An attacker who is able to modify the IoBT data sent to the monitor will be able to bias those values and make them, say, artificially larger; however, if the attacker wishes to remain undetected it will be limited to biasing each value by at most c relative to the previous value, and in particular can only achieve $y(N) < y(0) + c.N$. Now assume that original sensor readings are modified due to some background noise. The attacker can leverage this change in at least two ways. First, it can now bias the values by a larger amount without being detected. Furthermore, the anomaly-detection algorithm now has a nonzero false-positive rate, and the attacker can try to exploit that as well, e.g., by modifying data in a more extreme way that raises the alarm but only a small percentage of the time. To address this challenge, we have initiated the study of how attackers can affect the utility and integrity of an IoBT system by leveraging the background noise. In this setting, the attacker takes advantage of the fact that noise is somehow added to measurements and then strategically uses this noise to hide false data that affects the utility of the system.

We showed how attacks for systems that operate in noisy environments (this noise also could be added to protect the privacy and secrecy as well) have a much higher impact than attacks against systems without such noise. Furthermore, we assumed a strong adversary that knows the data distribution under normal conditions and our exact attack-detection algorithm. Even though this is a strong assumption, it allows us to analyze the worst case scenarios. We then proposed a better detection strategy based on solving a game theory problem in the functional space of all possible distributions. Our experiments show the value of our defenses by comparing our solution to state of the art adversarial attack detection tools currently used in practice for traffic estimation. Our main contribution shows how the optimal attack probability distribution is related to the probability distribution of the added noise (we focused on Laplace and Gaussian noise settings). In addition, that same result shows how our optimal proposed algorithm is a Nash equilibrium between an attacker trying to find an optimal attack distribution, and the defender trying to design an optimal bad data detection algorithm. Therefore there is no other bad data detection algorithm that is better than our proposed design for an IoBT sensor setting. Furthermore, our saddle point guarantees that less powerful adversaries will perform less effectively against our defenses, even when we do not know which attack distribution they will use.

- **Accountable Federated Learning via Hybrid Blockchain Architecture:** [3] Federated Learning (FL) is a distributed, and decentralized machine learning protocol. By executing FL, a set of

agents can jointly train a model without sharing their datasets with each other, or a third-party. This makes FL particularly suitable for settings where data privacy is desired. At the same time, concealing training data gives attackers an opportunity to inject backdoors into the trained model. It has been shown that an attacker can inject backdoors to the trained model during FL, and then can leverage the backdoor to make the model misclassify later. Several works tried to alleviate this threat by designing robust aggregation functions. However, given more sophisticated attacks are developed over time, which by-pass the existing defenses, we approach this problem from a complementary angle in this work. Particularly, we aim to discourage backdoor attacks by detecting, and punishing the attackers, possibly after the end of the training phase. To this end, we develop a hybrid blockchain-based FL framework that uses smart contracts to automatically detect, and punish the attackers via monetary penalties. Our framework is general in the sense that any aggregation function, and any attacker detection algorithm can be plugged into it. We conduct experiments to demonstrate that our framework preserves the communication-efficient nature of FL, and provide empirical results to illustrate that it can successfully penalize attackers by leveraging our novel attacker detection algorithm.

- **Defending against Backdoors in Federated Learning with Robust Learning Rate:** [5]

Federated learning (FL) allows a set of agents to collaboratively train a model without sharing their potentially sensitive data. This makes FL suitable for privacy-preserving applications. At the same time, FL is susceptible to adversarial attacks due to decentralized and unvetted data. One important line of attacks against FL is the backdoor attacks. In a backdoor attack, an adversary tries to embed a backdoor functionality to the model during training that can later be activated to cause a desired misclassification. To prevent backdoor attacks, we propose a lightweight defense that requires minimal change to the FL protocol. At a high level, our defense is based on carefully adjusting the aggregation server’s learning rate, per dimension and per round, based on the sign information of agents’ updates. We first conjecture the necessary steps to carry a successful backdoor attack in FL setting, and then, explicitly formulate the defense based on our conjecture. Through experiments, we provide empirical evidence that supports our conjecture, and we test our defense against backdoor attacks under different settings. We observe that either backdoor is completely eliminated, or its accuracy is significantly reduced. Overall, our experiments suggest that our defense significantly outperforms some of the recently proposed defenses in the literature. We achieve this by having minimal influence over the accuracy of the trained models. In addition, we also provide convergence rate analysis for our proposed scheme.

- **Does Explainable Artificial Intelligence Improve Human Decision-Making?:** [1] Explainable AI provides insights to users into the why for model predictions, offering potential for users to better understand and trust a model, and to recognize and correct AI predictions that are incorrect. Prior research on human and explainable AI interactions has focused on measures such as interpretability, trust, and usability of the explanation. There are mixed findings whether explainable AI can improve actual human decision-making and the ability to identify the problems with the underlying model. Using real datasets, we compare objective human decision accuracy without AI (control), with an AI prediction (no explanation), and AI prediction with explanation. We find providing any kind of AI prediction tends to improve user decision accuracy, but no conclusive evidence that explainable AI has a meaningful impact. Moreover, we observed the strongest predictor for human decision accuracy was AI accuracy and that users were somewhat able to detect when the AI was correct vs. incorrect, but this was not significantly affected by including an explanation. Our results indicate that, at least in some situations, the why information provided in explainable AI may not enhance user decision-making, and further research may be needed to understand how to integrate explainable AI into real systems.

References

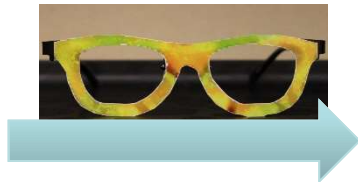
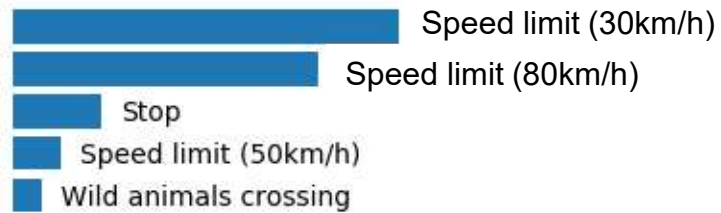
- [1] Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. Does explainable artificial intelligence improve human decision-making? In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6618–6626. AAAI Press, 2021.
- [2] Imrul Chowdhury Anindya and Murat Kantarcioglu. Adversarial anomaly detection using centroid-based clustering. In *2018 IEEE International Conference on Information Reuse and Integration, IRI 2018, Salt Lake City, UT, USA, July 6-9, 2018*, pages 1–8. IEEE, 2018.
- [3] Harsh Bimal Desai, Mustafa Safa Özdayi, and Murat Kantarcioglu. Blockfla: Accountable federated learning via hybrid blockchain architecture. In Anupam Joshi, Barbara Carminati, and Rakesh M. Verma, editors, *CODASPY '21: Eleventh ACM Conference on Data and Application Security and Privacy, Virtual Event, USA, April 26-28, 2021*, pages 101–112. ACM, 2021.
- [4] Jairo Giraldo, Alvaro A. Cárdenas, Murat Kantarcioglu, and Jonathan Katz. Adversarial classification under differential privacy. In *27th Annual Network and Distributed System Security Symposium, NDSS 2020, San Diego, California, USA, February 23-26, 2020*. The Internet Society, 2020.
- [5] Mustafa Safa Özdayi, Murat Kantarcioglu, and Yulia R. Gel. Defending against backdoors in federated learning with robust learning rate. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 9268–9276. AAAI Press, 2021.
- [6] Yan Zhou, Murat Kantarcioglu, and Bowei Xi. Adversarial active learning in the presence of weak and malicious oracles. In Leong Hou U and Hady W. Lauw, editors, *Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2019 Workshops, BDM, DLKT, LDRC, PAISI, WeL, Macau, China, April 14-17, 2019, Revised Selected Papers*, volume 11607 of *Lecture Notes in Computer Science*, pages 77–89. Springer, 2019.
- [7] Yan Zhou, Murat Kantarcioglu, and Bowei Xi. Exploring the effect of randomness on transferability of adversarial samples against deep neural networks. *IEEE Trans. Dependable Secur. Comput.*, 20(1):83–99, 2023.

Adversarial Machine Learning: Our solution highlights

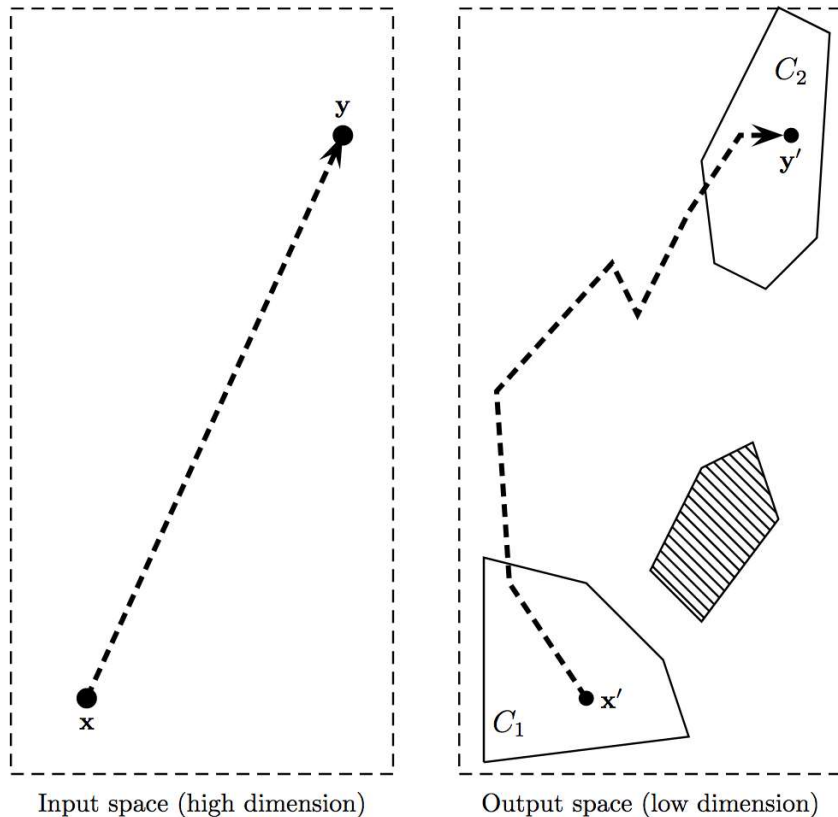
Exploring The Effect Of Randomness On Transferability Of Adversarial Samples Against Deep Neural Networks *

“Exploring The Effect Of Randomness On Transferability Of Adversarial Samples Against Deep Neural Networks”, IEEE Transactions On Dependable And Secure Computing, Vol. 20, No. 1, January/February 2023 83,

Motivation: Adversarial Attacks



Conjecture for Adversarial Attack Mitigation



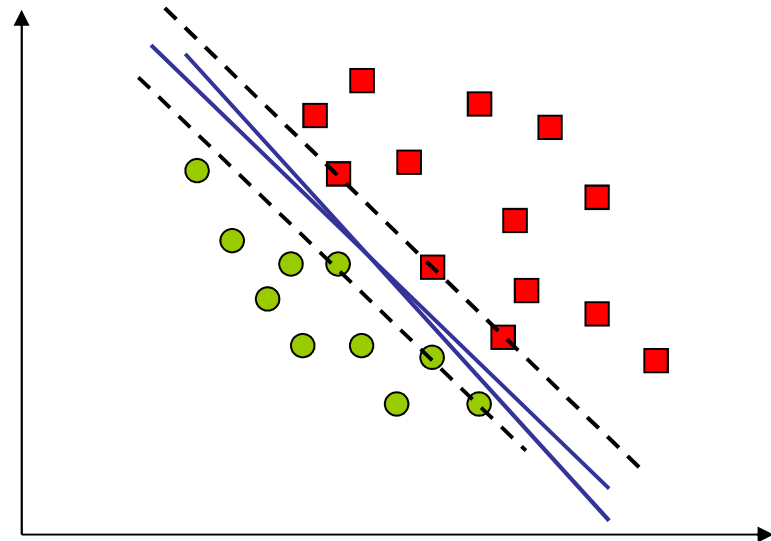
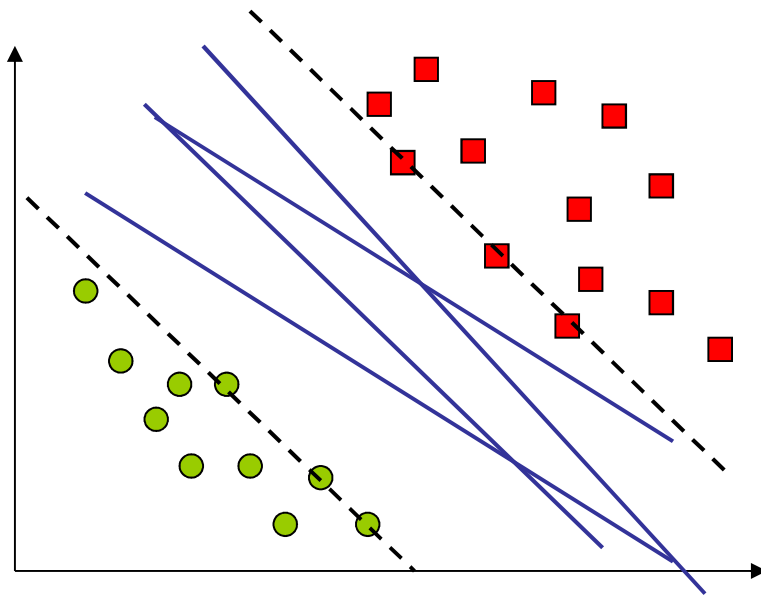
The problem of adversarial learning may be insurmountable.

- Adding layers, changing structures of DNN cannot mitigate attacks
- Potential mitigation: randomization

“A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance”, CoRR abs/1901.10861 (2019), Adi Shamir and Itay Safran and Eyal Ronen and Orr Dunkelman.

Version Space Spread

- If the population of the hypotheses in the version space has a large variance, there is more room for randomization as a defense.



Randomization with Random Noise

- Adding a random noise to:
 - Input x , $f(x, \theta) \rightarrow f(x + \Delta x, \theta)$
 - DNN model parameter θ , $f(x, \theta) \rightarrow f(x, \theta + \Delta \theta)$
 - Input x & model parameter θ , $f(x, \theta) \rightarrow f(x + \Delta x, \theta + \Delta \theta)$
- Goal
 - Detection
 - Classification
 - Make it harder to attack multiple machines at the same time
 - Similar to software randomization

Randomization for Detection

- Based on the assumption that the prediction distribution of adversarial samples is different.
- Establish a valid interval for predictions on clean data.

$$\max(|f_i - l_i|, |f_i - u_i|) > c |u_i - l_i|$$

Where u_i and l_i are the bounds of the confidence interval for prediction probability:

$$[l_i, u_i] = [\bar{f}_i - t_{\alpha/2, n-1} \frac{\hat{S}}{\sqrt{n}}, \bar{f}_i + t_{\alpha/2, n-1} \frac{\hat{S}}{\sqrt{n}}]$$

Randomization for Classification

Randomized Classification (x)

//prediction function $f(\theta, x)$

for each input x ,

 generate a random $\Delta\theta$

 predict with $f(\theta + \Delta\theta, x)$

Experimental Setup

Data set: CIFAR10, MNIST, German Traffic Sign, and Street View House Number

Architecture:

- CIFAR10: VGG16
- MNIST: 2 conv., 64 3x3 filters, 2 dense
- Traffic Sign: 6 conv., 32, 64, and 128 3x3 filters, 3 dense
- House Number: 2 conv., 64 3x3 filters, 3 dense

Experimental Setup (cont.)

Attacks:

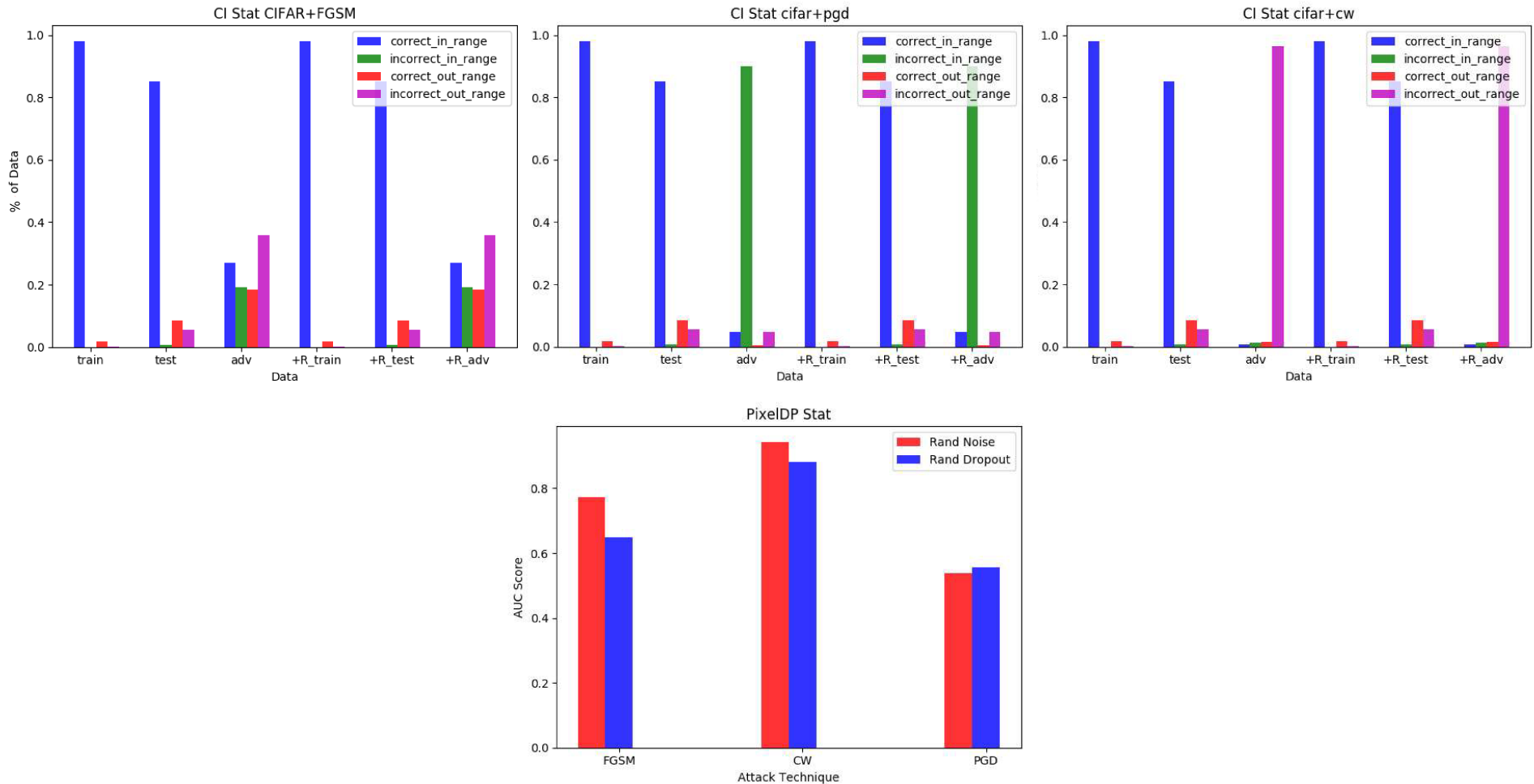
- FGSM attack
 - ε : attack step size
 - Detection: $\varepsilon = 0.2$
- Carlini & Wagner's L_2 attack
 - Greater *confidence* stronger attack
 - Detection: *confidence*=10
 - Classification: *confidence*=1, 5, 10
- Projected Gradient Descent L_∞ attack
 - ε : maximum perturbation allowed
 - Detection: $\varepsilon = 0.3$
 - Classification: $\varepsilon = 0.05, 0.5, 1$

Detection Results

- Confidence Interval vs. Random Dropout
 - Random Dropout: model confidence on adversarial samples is obtained by computing Bayesian uncertainty estimates from randomized DNNs and density estimation in the subspace of deep features learned by the model.

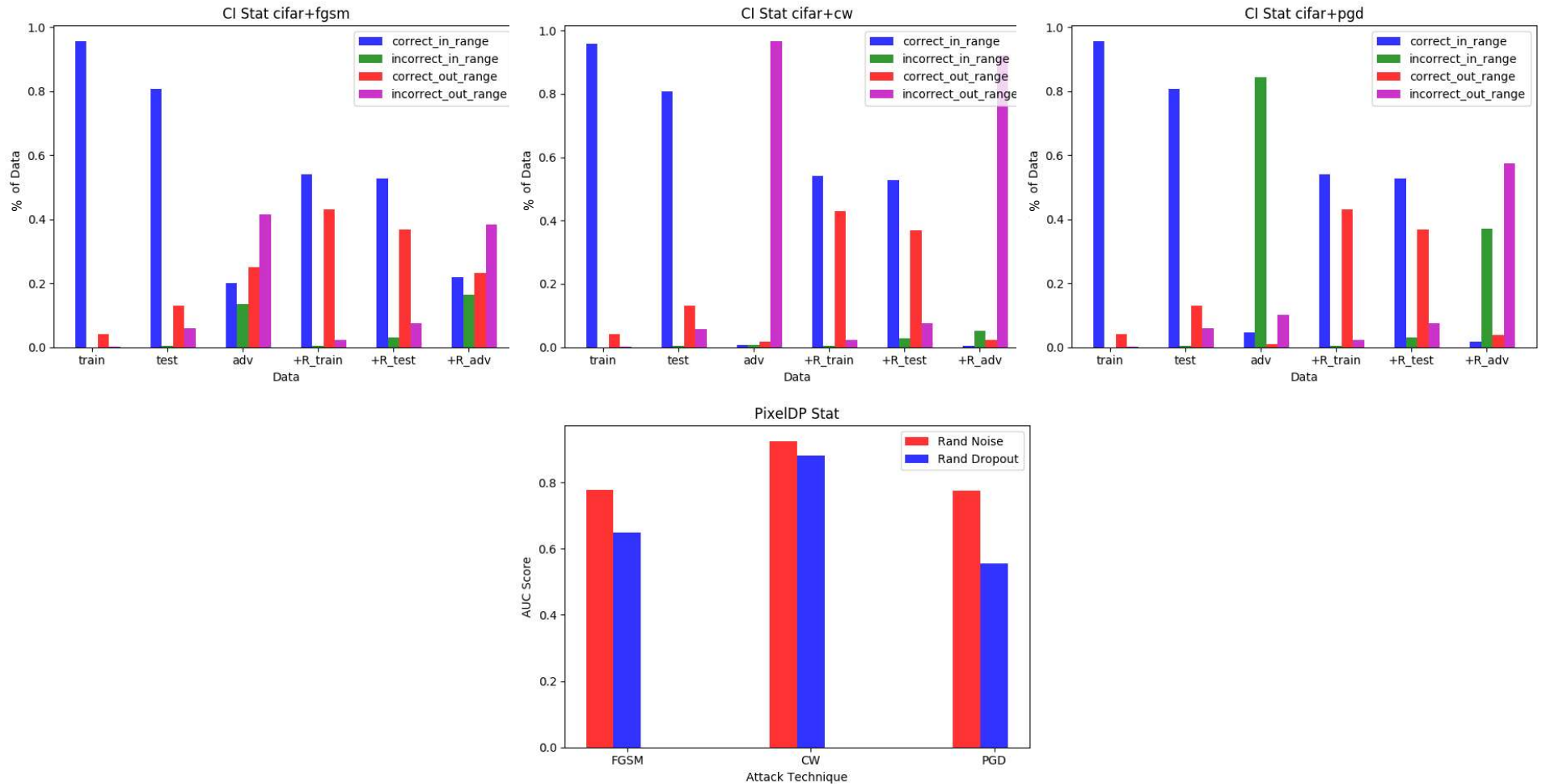
Experimental Results on CIFAR10

$$f(x, \theta) \rightarrow f(x, \theta + \Delta\theta)$$



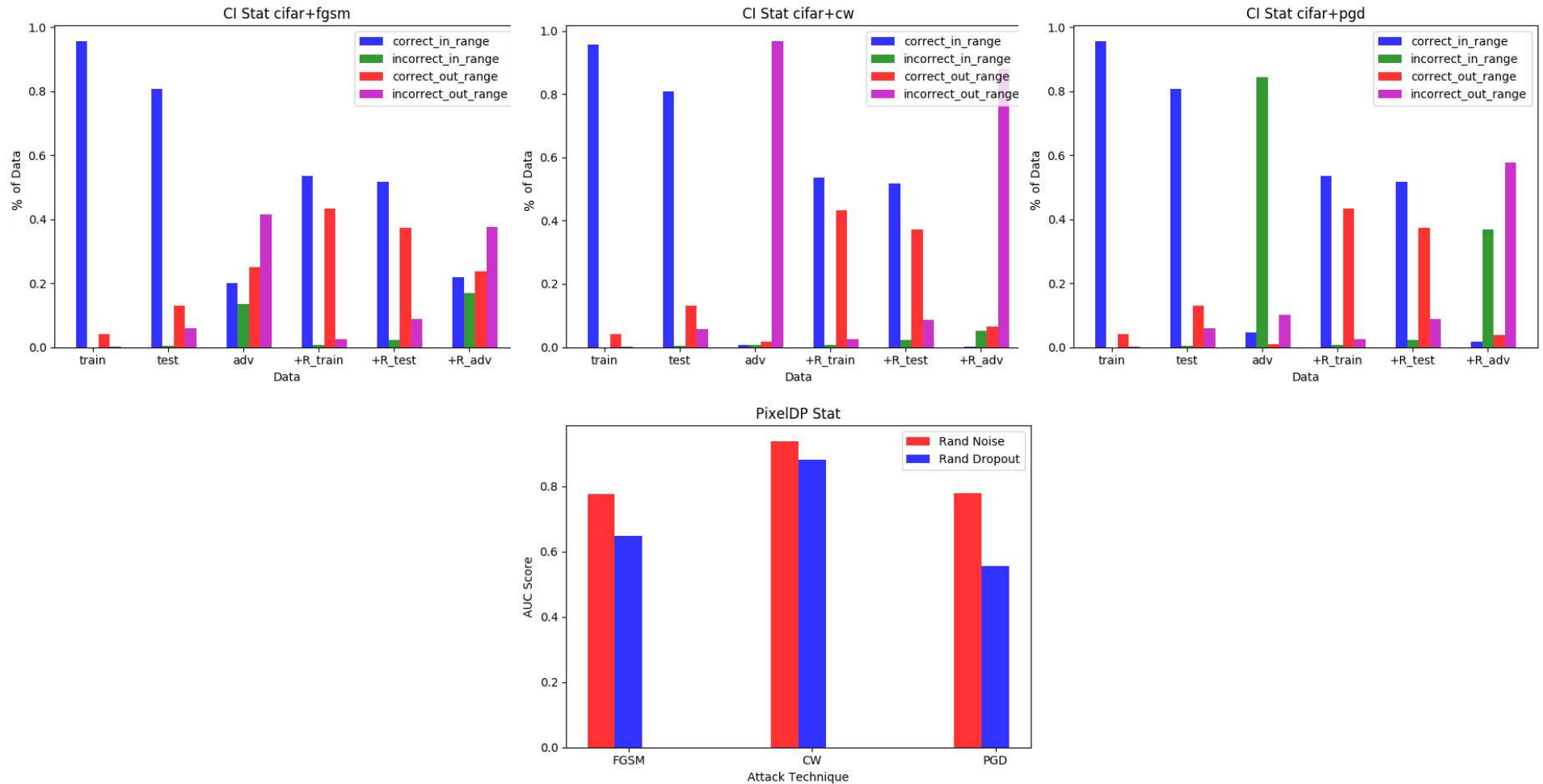
Experimental Results on CIFAR10 (cont.)

$$f(x, \theta) \rightarrow f(x + \Delta x, \theta)$$



Experimental Results on CIFAR10 (cont.)

$$f(x, \theta) \rightarrow f(x + \Delta x, \theta + \Delta \theta)$$



Classification Results

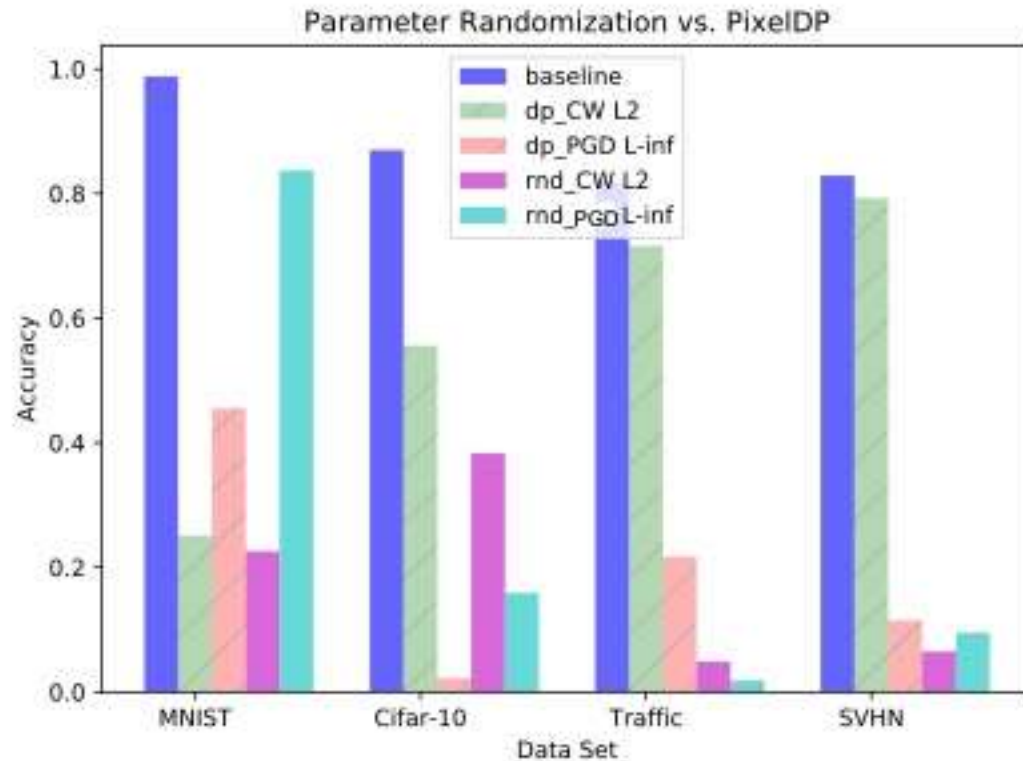
- Randomized Parameter vs. Certified Defense
 - Certified Defense: Differential Privacy based certified defense works by adding a noise layer into a DNN model and turning a deterministic DNN scoring function to a randomized scoring function insensitive to changes in the input.

Experimental Results on Classification

Mild attack:

confidence=1

$\varepsilon = 0.05$

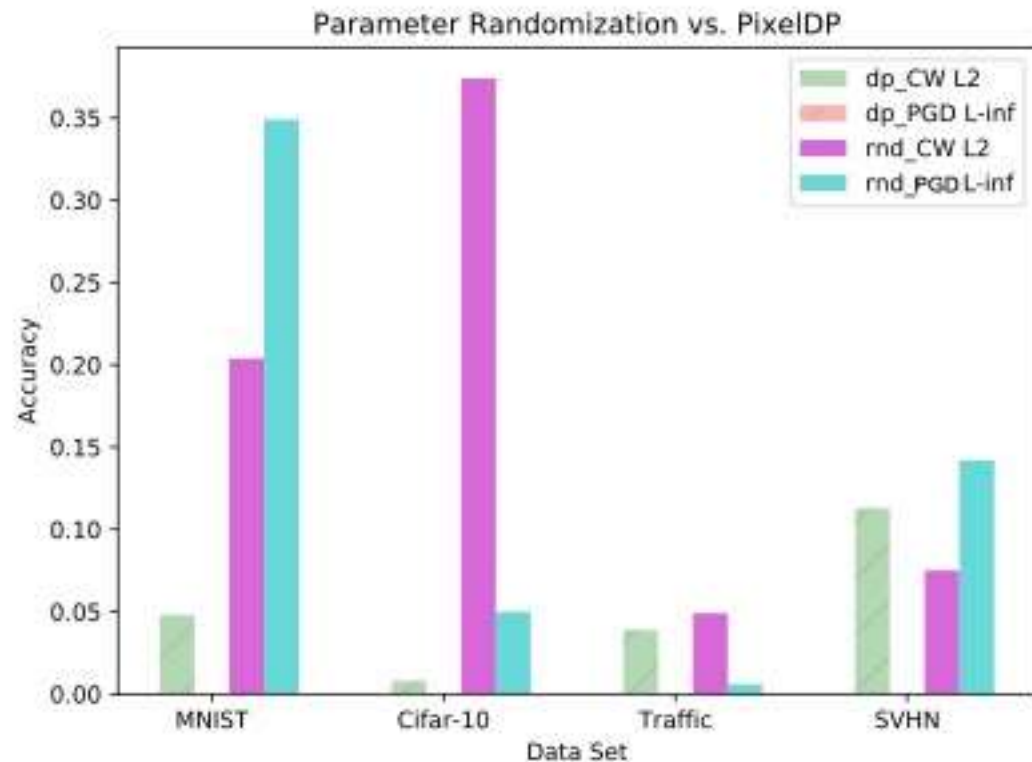


Experimental Results on Classification (cont.)

Moderate attack:

Confidence=5

$\varepsilon = 0.5$

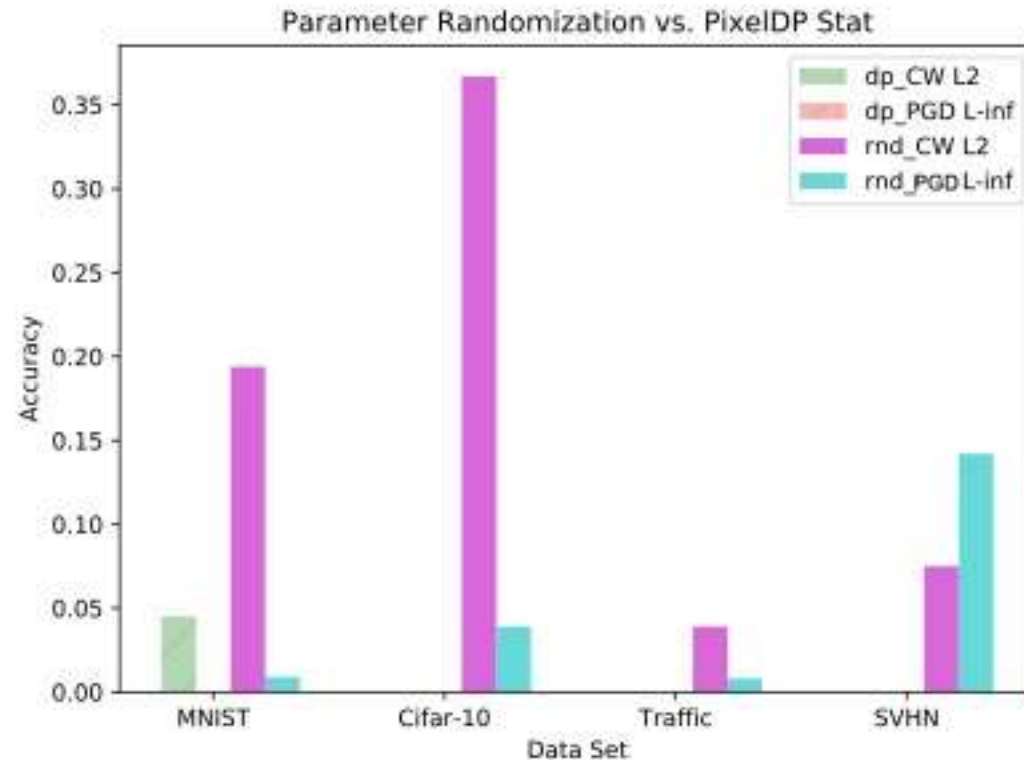


Experimental Results on Classification (cont.)

Severe attack:

Confidence=10

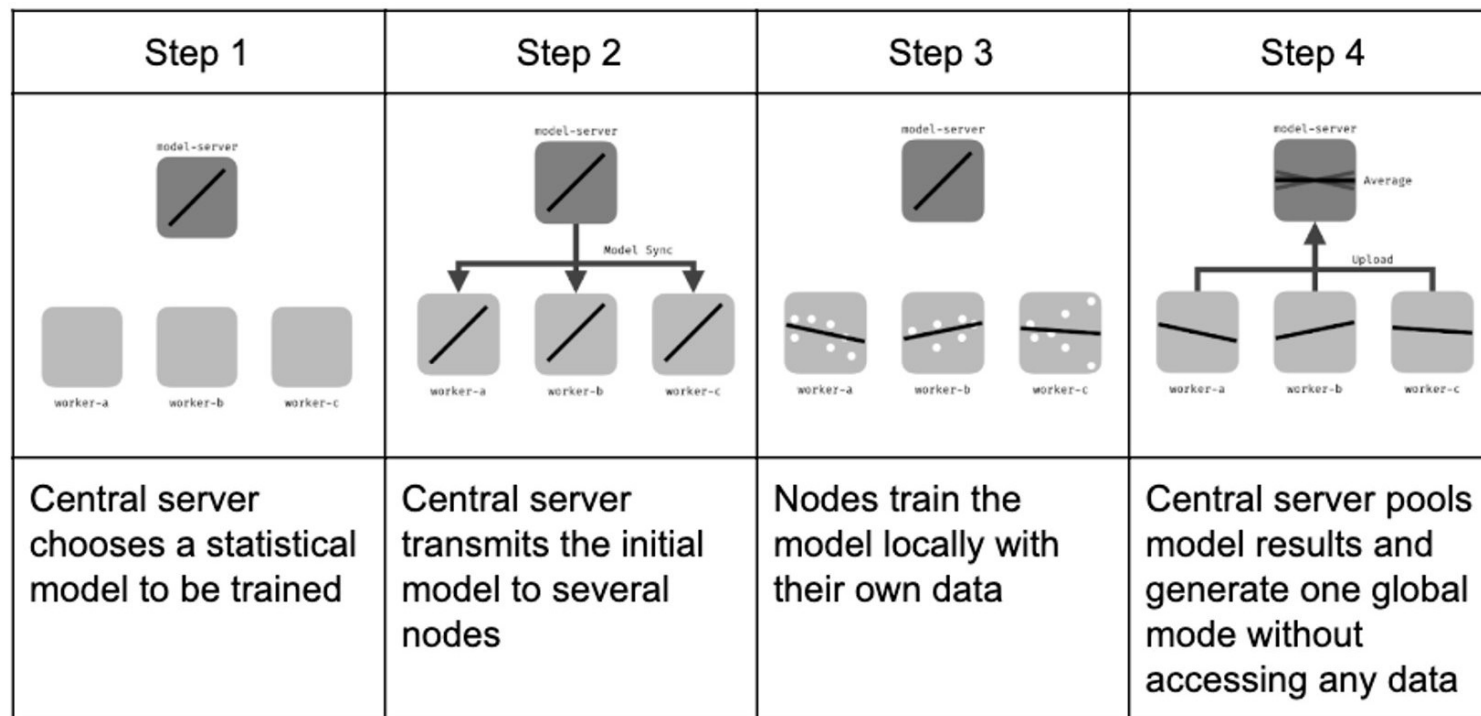
$\varepsilon = 1$



Summary

- Randomizing input and learning models can help detect abnormal prediction distribution of adversarial samples.
- Randomized learning models can improve robustness against adversarial attacks

Federated Learning: Security vs Robustness*

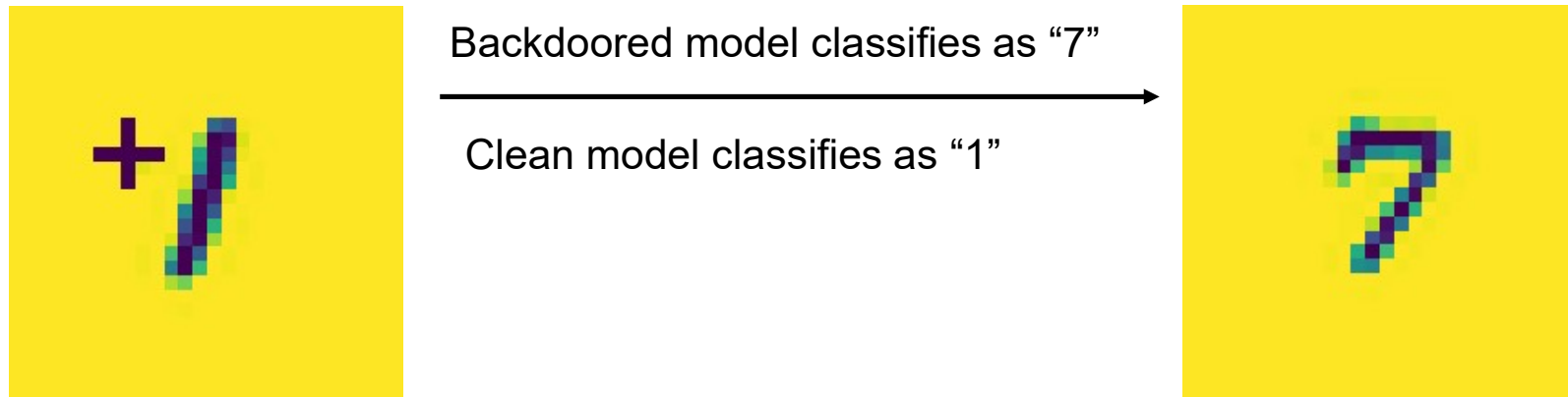


https://en.wikipedia.org/wiki/Federated_learning

* AAAI 2021

Backdoor Attacks

- Backdoor: a targeted misclassification functionality
- Can be introduced via **data poisoning** in centralized setting
 - Gain access to training data
 - Add pixel-pattern to some samples of a class, re-label them to a target class



Backdoor Attacks in FL context

- Backdoor attacks can be carried via **model poisoning** [1-2]
 - Corrupt agents train their models on poisoned data
 - Send the malicious update to server for aggregation
- Aggregation function should be robust
 - A single adversary can arbitrarily skew FedAvg

$$w_{t+1} = w_t + \eta \frac{\sum_{k \in S_t} n_k \cdot \Delta_t^k}{\sum_{k \in S_t} n_k}$$

w_t : weights at round t

S_t : selected agents at round t

Δ_t^k : update of k'th agent at round t

n_k : dataset size of k'th agent

η : server's learning rate

Our Defense: Robust Learning Rate (RLR) -1

- A hyperparameter called learning threshold, θ , at server-side
- For a dimension i , if sum of signs is less than θ , negate learning rate for dimension i
 - To maximize loss on that dimension

$$\eta_{\theta,i} = \begin{cases} \eta & |\sum_{k \in S_t} \text{sgn}(\Delta_{t,i}^k)| \geq \theta \\ -\eta & \text{otherwise.} \end{cases}$$

$$w_{t+1} = w_t + \eta_{\theta} \odot \frac{\sum_{k \in S_t} n_k \cdot \Delta_t^k}{\sum_{k \in S_t} n_k}$$

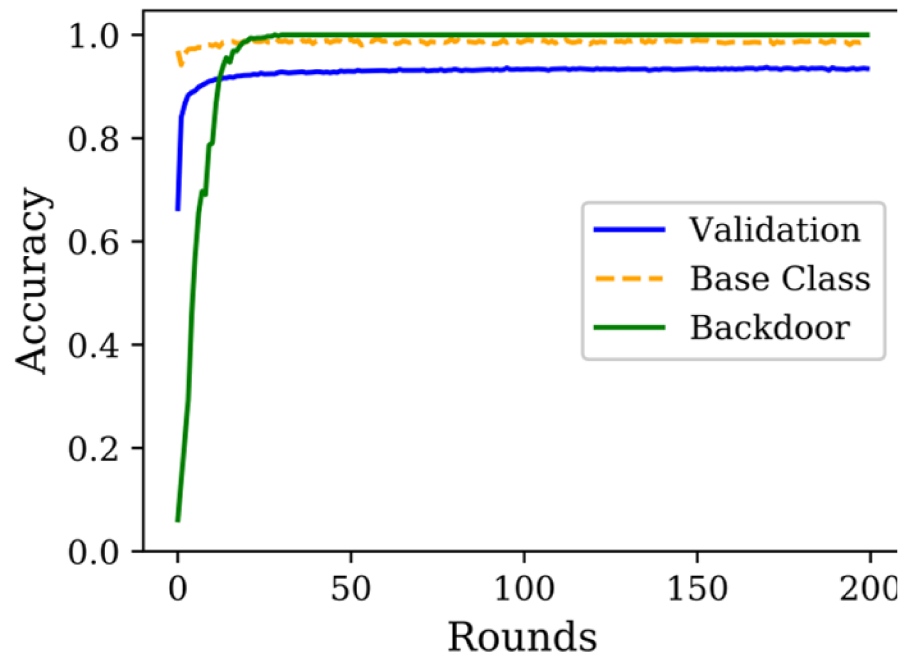
Our Defense: Robust Learning Rate (RLR)-2

- Let w_{adv} , w_{hon} be two distinct points on parameter space
 - w_{adv} : minimizes loss on backdoor, and main tasks
 - w_{hon} : minimizes loss on main task
- For some dimensions, honest and corrupt agents will try to move the model to different directions
- Sign information of updates can be treated as votes for directions.

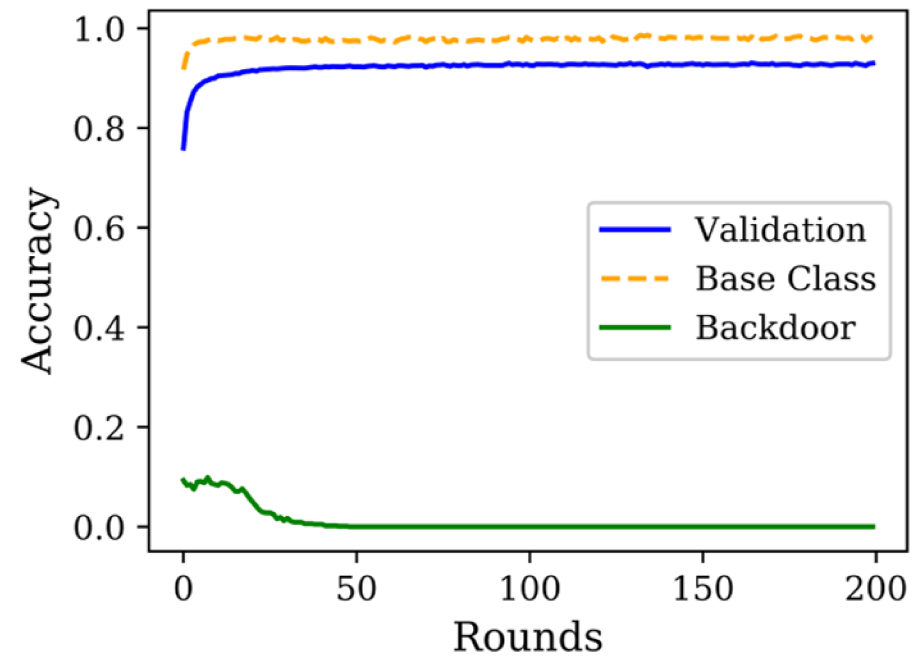
Experiments

- Simulate FL where 10% of agents are corrupt
 - Corrupt agents poison their local data via trojans
 - IID setting with 10 agents on Fashion-MNIST [4]
 - NIID setting with ~3k agents on FEMNIST [5]
 - ~ 30 agents per round
- Three metrics measured at each round
 - **Validation accuracy**
 - Backdoor accuracy
 - e.g., whether trojaned “1”s are classified as “7”
 - **Base class accuracy**
 - e.g., whether clean “1”s are classified as “1”

Learning Curves - IID

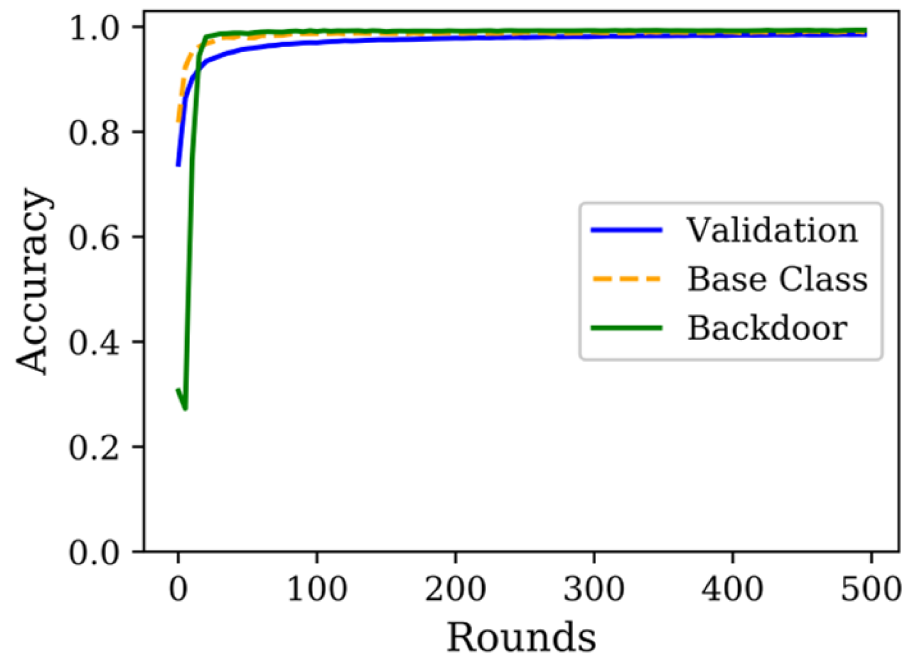


FedAvg

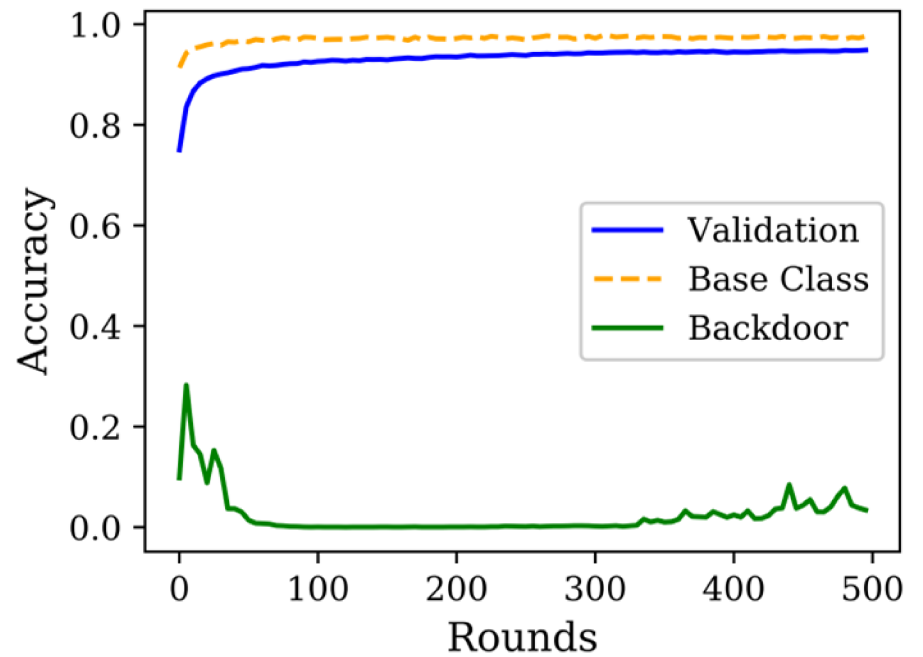


FedAvg with RLR

Learning Curves - NIID



FedAvg



FedAvg with RLR

Comparison with Other Defenses - IID

Aggregation	M	σ	Backdoor (%)	Validation (%)	Base (%)
FedAvg- <i>No Attack</i>	0	0	1	93.5	98.5
FedAvg	0	0	100	93.4	98.5
FedAvg	4	1e-3	100	93.2	99.1
FoolsGold	0	0	100	93.1	98.9
FoolsGold	4	1e-3	100	93.3	98.5
Comed	0	0	100	92.8	99.0
Comed	4	1e-3	99.5	92.8	98.4
Sign	0	0	100	92.9	98.7
Sign	4	1e-3	99.7	93.1	98.6
FedAvg with RLR	0	0	0	92.9	98.3
FedAvg with RLR	4	1e-3	0.5	92.2	97.4

DP can be applied to limit contribution of each agent

- For fairness/privacy purposes [6]
- Can deter label-flipping backdoors [7]

M : L_2 -norm threshold on updates

σ : std.dev of Gaussian noise

FoolsGold [8]

Comed [9]

Sign [3]

Comparison with Other Defenses - NIID

Aggregation	M	σ	Backdoor (%)	Validation (%)	Base (%)
FedAvg*- <i>No Attack</i>	0	0	21.1	98.6	99.1
FedAvg	0	0	99.3	98.5	99.0
FedAvg	0.5	1e-3	99.2	98.0	98.7
FoolsGold	0	0	98.5	98.9	99.5
FoolsGold	0.5	1e-3	99.1	97.9	98.6
Comed	0	0	82.3	96.3	98.4
Comed	0.5	1e-3	95.2	95.5	98.1
Sign	0	0	99.8	97.6	98.7
Sign	0.5	1e-3	99.7	97.8	98.5
FedAvg with RLR	0	0	3.4	94.8	97.6
FedAvg with RLR	0.5	1e-3	0.4	93.2	97.7

Conclusion: FL Poisoning Attacks?

- A simple defense that requires no changes to FL
- Agnostic to the aggregation function
- Outperforms some of the recent defenses
- Full version contains the following.
 - Distributed backdoor attacks [10]
 - Combining RLR with other aggregation functions
 - Extended set of experiments
 - More trojan patterns, higher corruption rates
 - M , σ values etc.

What if attacks are still successful ? *

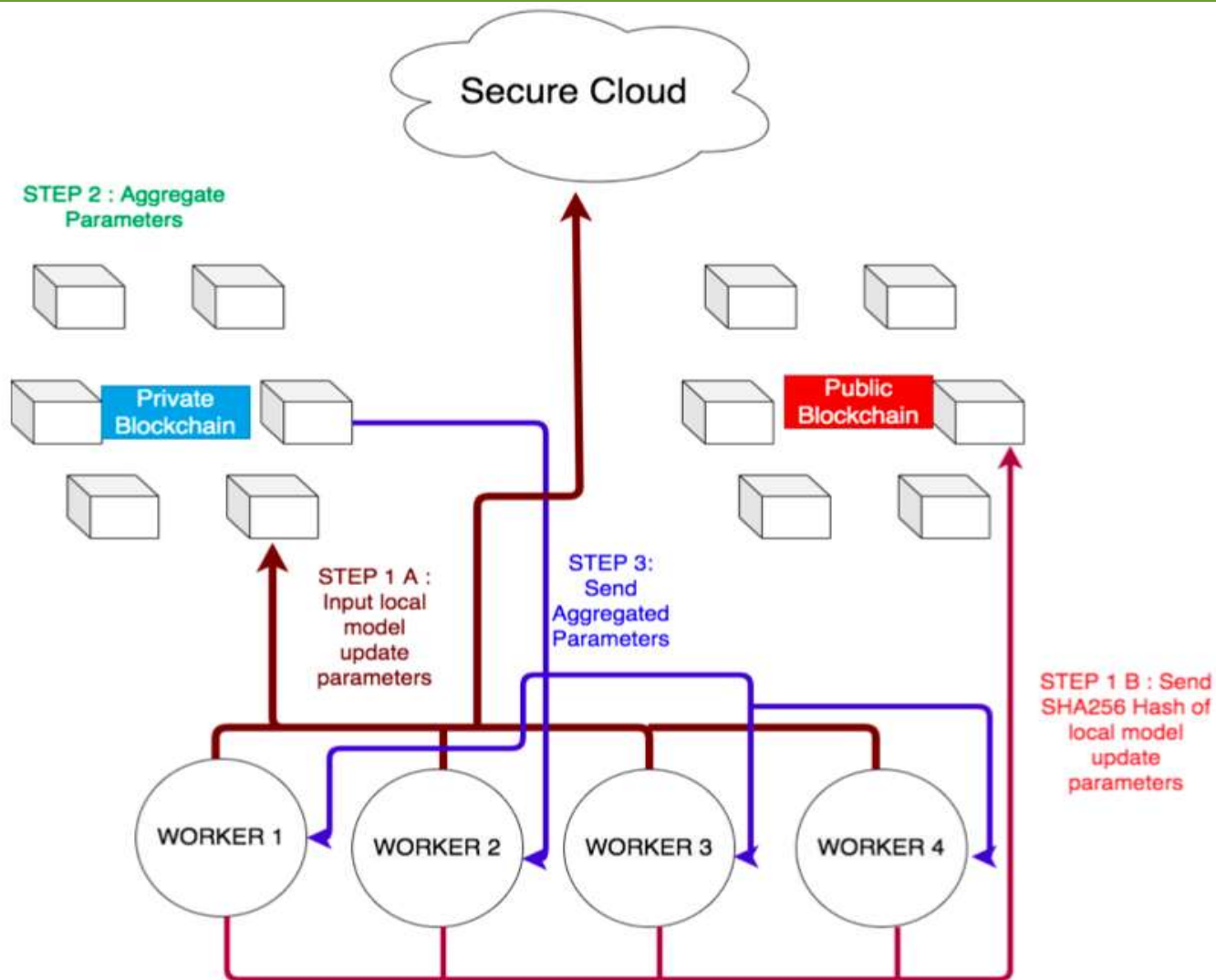


- **What if still defenses against poisoning attack?**
 - Leverage Private Blockchain and Public Blockchain to attain accountability, privacy and robustness altogether.
 - Store immutable training logs on blockchain
 - Trojan detection algorithm running off chain
 - Cloud services to store logs for tracing purposes

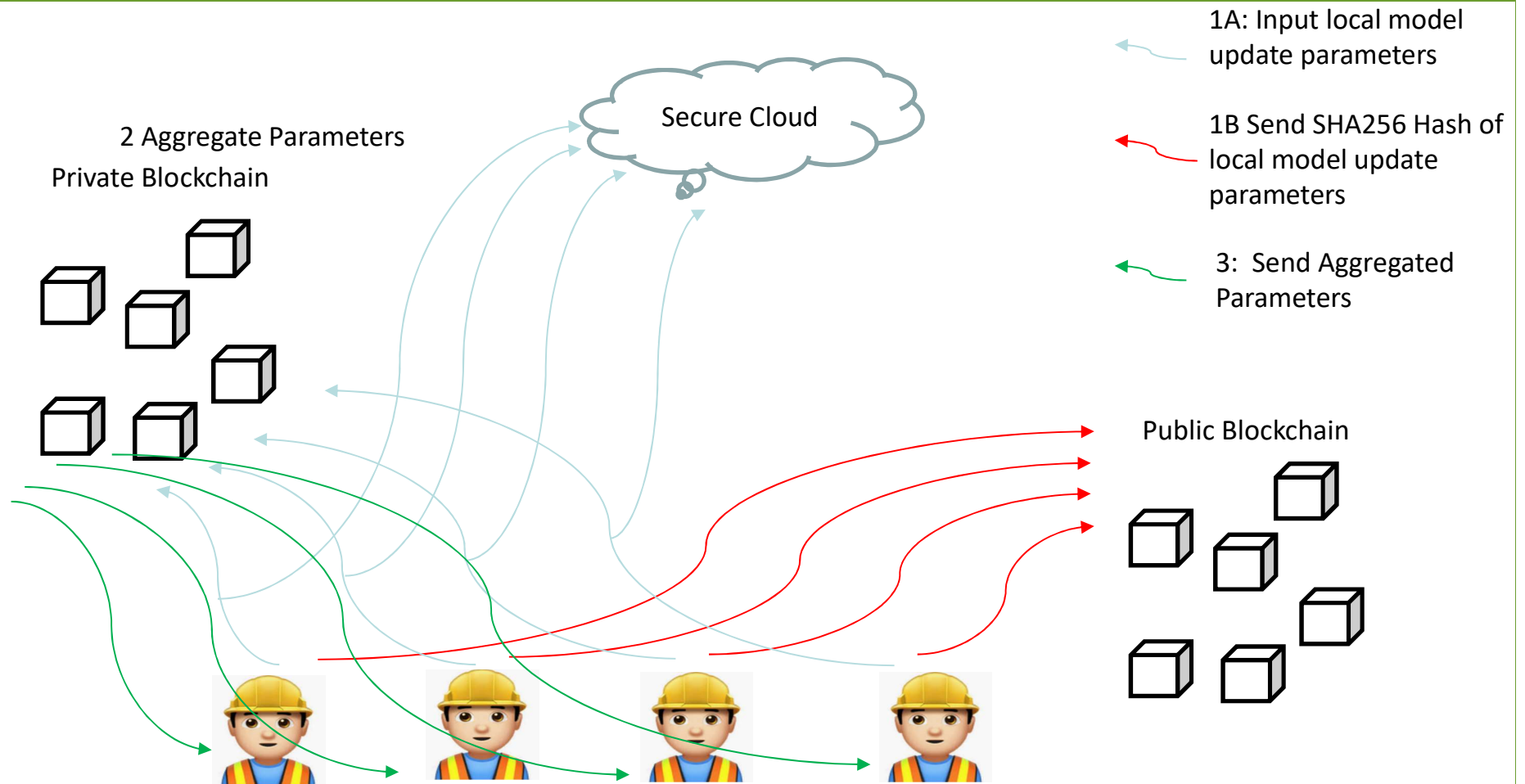


*ACM Codaspy 2021

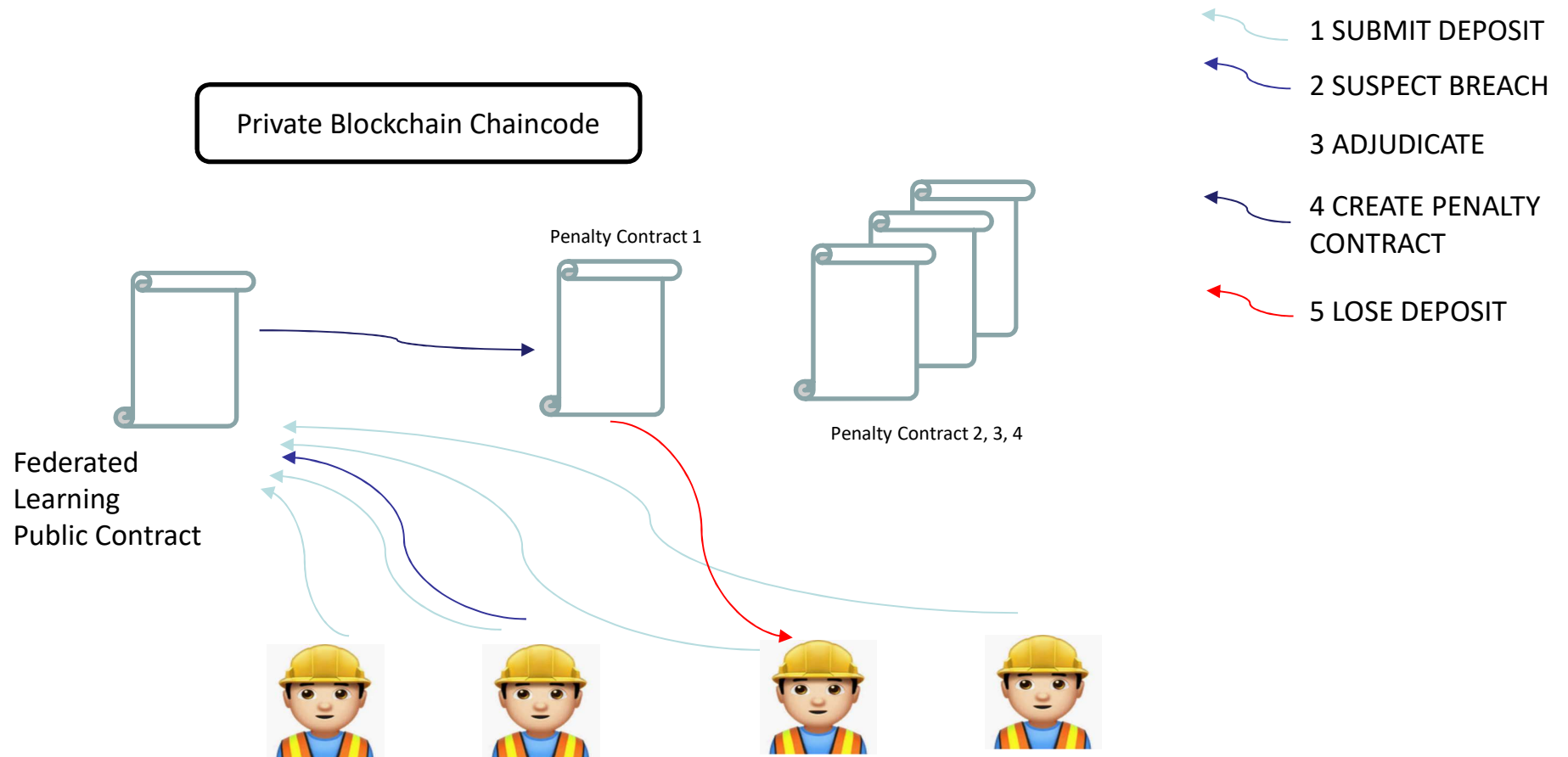
Blockchain Architecture



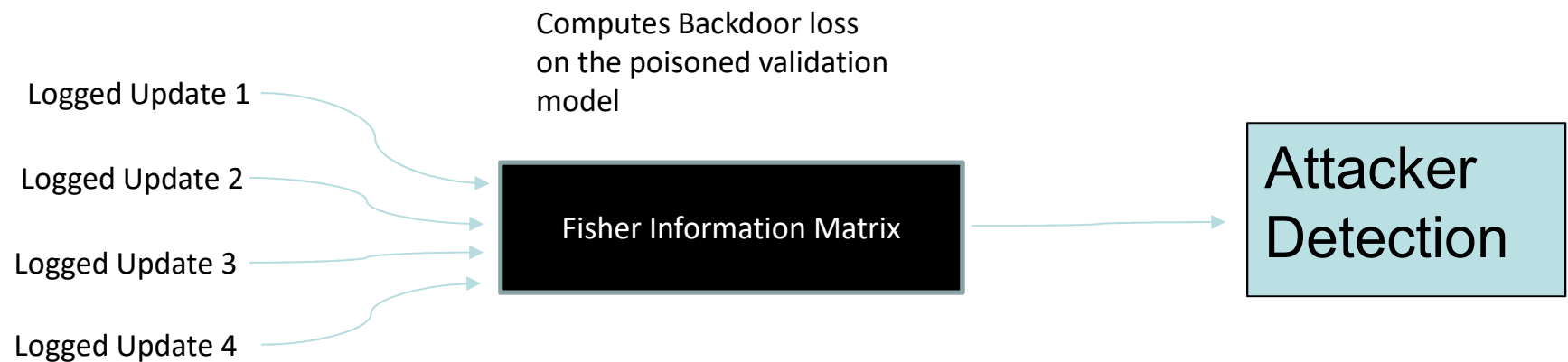
Architecture



Architecture: Cont



Attacker Detection Algorithm for Trojaning Attacks



Performance of our Attacker Detection Algorithm

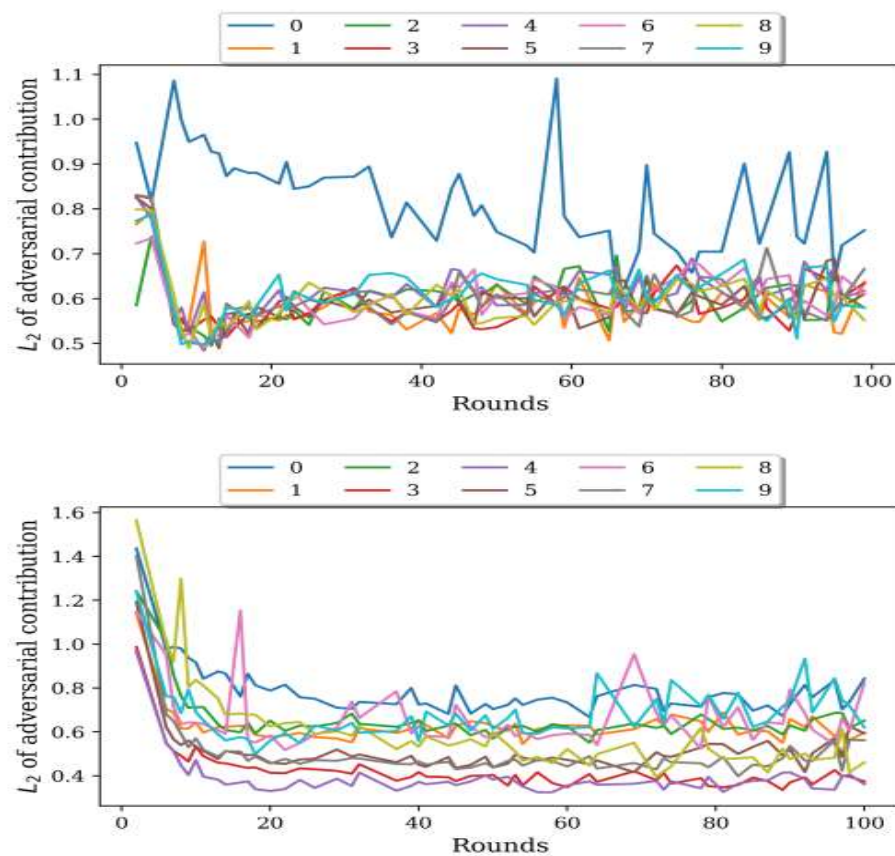


Figure 6: Samples from trojaned base class, and target class. The trojan pattern is a 5-by-5 plus pattern that is put to the top-left of base class instances. The goal of adversary is to make the model classify instances of dog class with the trojan pattern (left) as a horse (right).

Results of trojan detection algorithm in iid (top), and non-iid (bottom) settings.

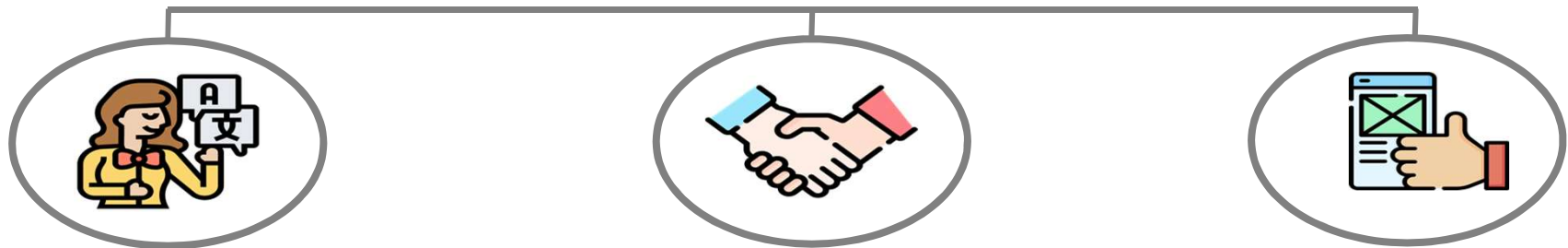
Can we ask for human help?

- **Can humans help ML/AI models to detect and mitigate issues?**
 - Humans help with detecting Trojans
 - Humans help with detecting potential bias

Explainable AI to Rescue ?*



- Prior research



Interpretability

Trust

Usability

*AAAI 2021

Goal

- Limited empirical research for its *impact on human decision-making*.
- To fully achieve fairness and accountability, explainable AI should lead to **better human decisions and detect errors**.

Contribution

- Empirically investigate whether explainable AI improves human decision-making using:
 - Human subject experiments.
 - Two choice classification tasks (objective human decision accuracy).
 - Real-world data.
 - Anchor LIME as explainable AI model.
 - Three different conditions where a user needs to make decision.

Datasets

- Heavily used in prior research that tries to understand algorithmic fairness and accountability issues.
- Involve real-world contexts.

COMPAS

- Scoring system used to assign risk scores to criminal defendants to determine their likelihood of reoffending.
- Has been used to detect potential biases in criminal justice system.

Census income

- Used to test many ML techniques.
- Involves predictions of individuals' income status.
- Has been associated with potential biases in making decisions.

AI Conditions

Control

Participants were provided with no prediction or information from the AI.

AI

Participants were provided with only an AI prediction.

AI with Explanation

Participants received an AI prediction, as well as an explanation of the prediction using anchor LIME.

Hypotheses

H.1

Explainable AI enhances decision-making process compared to the other conditions

H.2

A participant's decision accuracy is positively associated with AI accuracy

H.3

Average participant's decision accuracy does not outperform AI accuracy

H.4

Participants outperform AI accuracy more often with explainable AI over AI prediction alone

H.5

Participants follow explainable AI recommendation more often than AI only recommendation

H.6

Explainable AI increases participants' decision confidence

H.7

A participant's decision confidence is positively correlated with the accuracy of his/her decision

Trials

- Two-alternative forced choice for the classification
- Each choice was correct on 50% of the trials.
- AI prediction and/or explanation may appear, depending on the AI condition.
- 10 practice trials with feedback.

Control Case

Age ⓘ	Charge Degree ⓘ	Race ⓘ	Risk Score ⓘ	Sex ⓘ	Number of Priors ⓘ
Greater_than_45	Felony	Caucasian	Low	Female	1

Choose whether or not you think this person will commit another crime within 2 years.

Will Not Re-offend

Will Re-offend



AI Case

Age ⓘ	Charge Degree ⓘ	Race ⓘ	Risk Score ⓘ	Sex ⓘ	Number of Priors ⓘ
Greater_than_45	Felony	Caucasian	Low	Female	1

AI Predicts: Will NOT Re-offend

Choose whether or not you think this person will commit another crime within 2 years.

Will Not Re-offend

Will Re-offend



AI with Explanation

Age ⓘ	Charge Degree ⓘ	Race ⓘ	Risk Score ⓘ	Sex ⓘ	Number of Priors ⓘ
Greater_than_45	Felony	Caucasian	Low	Female	1

AI Predicts: Will NOT Re-offend

AI Explanation:

- race: Caucasian, AND
- risk_score: Low, AND
- priors_count: priors_count is between 0.00 and 1.00

Choose whether or not you think this person will commit another crime within 2 years.

Will Not Re-offend

Will Re-offend



Study Example

— Control
— AI
— AI Explanation

Age ⓘ	Charge Degree ⓘ	Race ⓘ	Risk Score ⓘ	Sex ⓘ	Number of Priors ⓘ
Greater_than_45	Felony	Caucasian	Low	Female	1

AI Predicts: Will NOT Re-offend

AI Explanation:

- race: Caucasian, AND
- risk_score: Low, AND
- priors_count: priors_count is between 0.00 and 1.00

Choose whether or not you think this person will commit another crime within 2 years.

- After a decision is made, participants are asked to enter their confidence in that choice.
 - 1 (No Confidence) to 5 (Full Confidence).
- Feedback is then displayed.

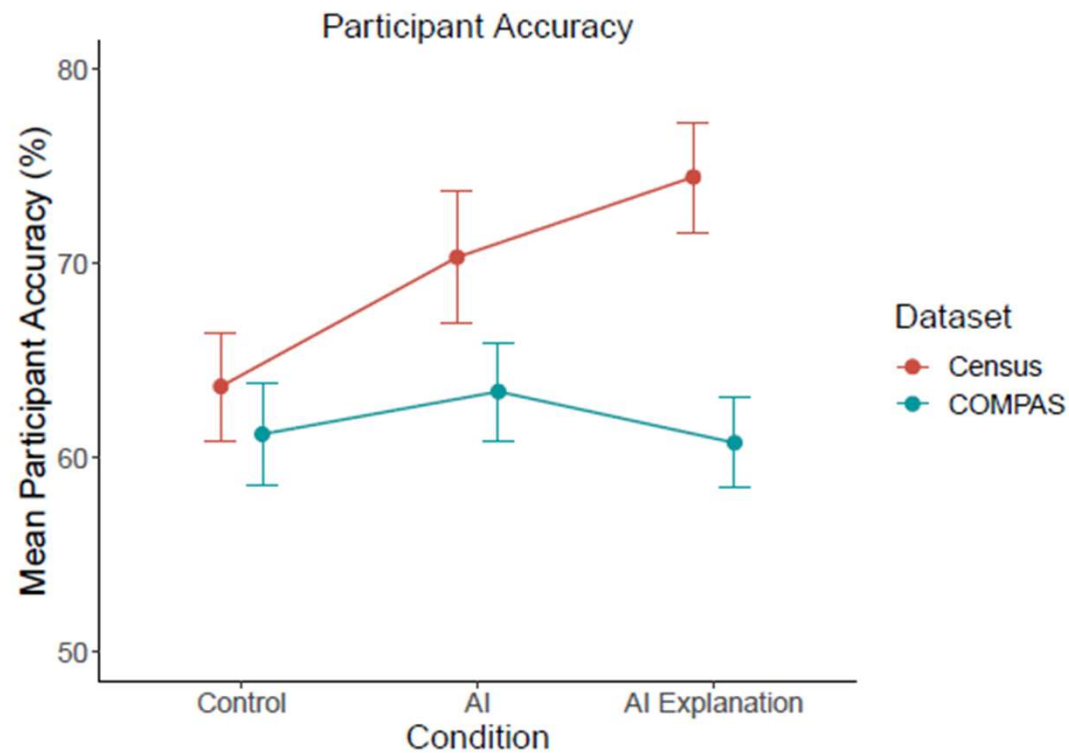
Machine Learning Models

- Compared the prediction accuracy of:
 - Logistic Regression.
 - Multi-layer Perceptron Neural Network.
 - Random Forest.
 - Support Vector Machine (SVM) with rbf kernel.
- Selected the best classifier for each dataset.
 - Multi-layer Perceptron Neural Network for Census income data $\Rightarrow$ accuracy of 82%.
 - SVM with rbf kernel for COMPAS data $\Rightarrow$ accuracy of 68%.

Participant Recruitment and Procedure

- Amazon Mechanical Turk (AMT).
- 50 participants in each of the six experimental conditions.
- Total of 300 participants (57.67% male).
- Most participants were 18 to 44 years old (80.67%).
- Participants completed 10 practice trials, followed by 50 test trials and a brief questionnaire.
- Median time to complete the practice and test trials was 18 minutes.

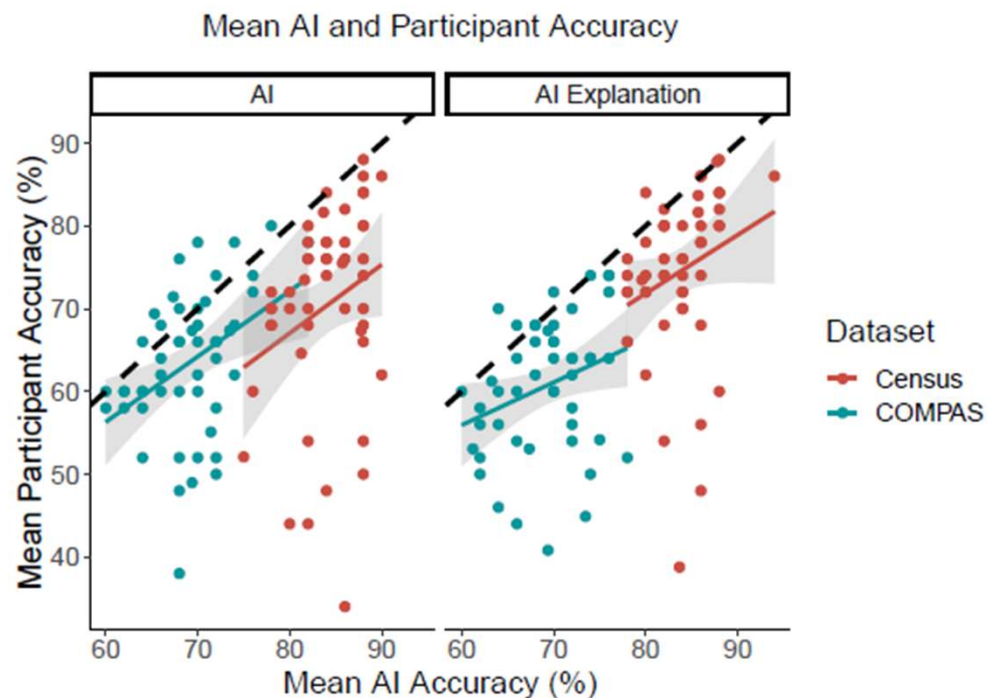
Participant Decision-Making Accuracy



H.1

Explainable AI enhances decision-making process compared to the other conditions

AI Accuracy and Participant Decision-Making Accuracy



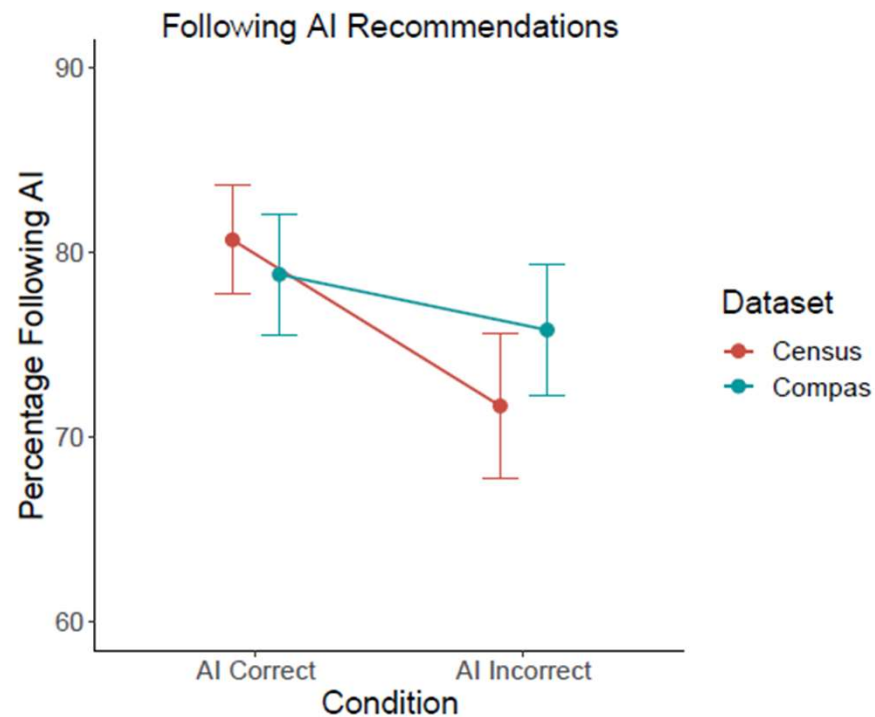
H.2

A participant's decision accuracy is positively associated with AI accuracy

H.3

Average participant's decision accuracy does not outperform AI accuracy

Adherence to AI Model Predictions



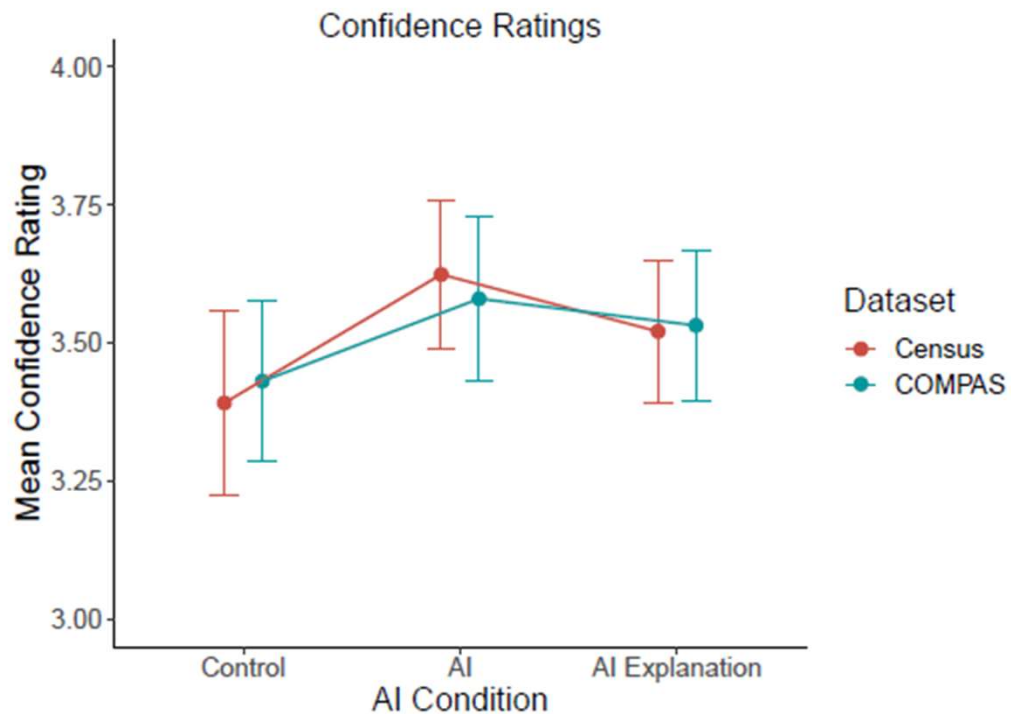
H.2

A participant's decision accuracy is positively associated with AI accuracy

H.5

Participants follow explainable AI recommendation more often than AI only recommendation

Confidence Ratings



H.6

Explainable AI increases participants' decision confidence

H.7

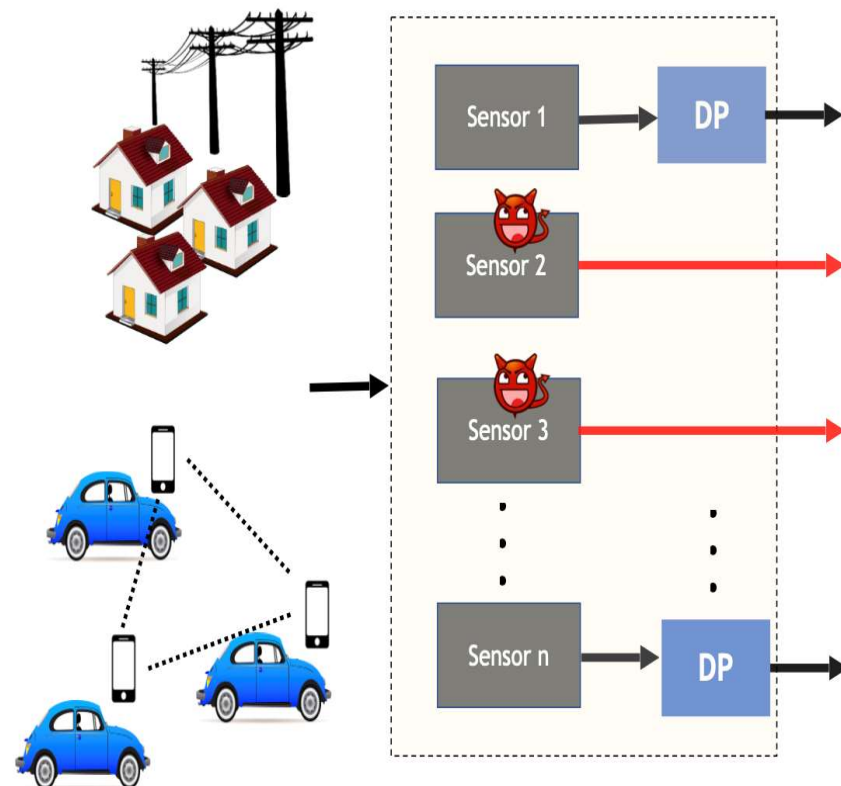
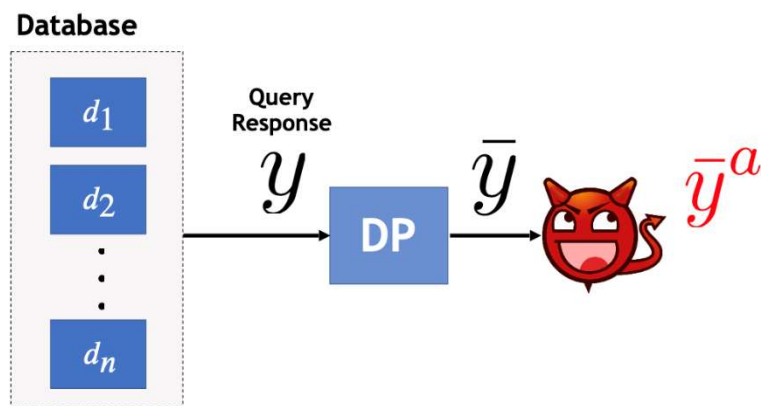
A participant's decision confidence is positively correlated with the accuracy of his/her decision

Conclusion: Is explainable AI enough?

- Providing an AI prediction enhances human decision accuracy.
- No conclusive evidence that *explainable* AI has a meaningful impact.
- The strongest predictor for human decision accuracy was AI accuracy.
- AI with or without explanation can increase the confidence of human users.

Security and Privacy for Data Science: Adversarial ML and Privacy for CPS*

- Consumer data protected by Differential Privacy (DP)
- Our adversary: data poisoning by hiding their attacks in DP noise
- How to detect attacks while protecting privacy in CPS?



*NDSS 2020

Future directions

- **Robust FL for IIoT devices**
 - Non-iid makes it hard to detect attacks
 - Devices may be limited so models could be shallow
 - Learn robust models via FL under Non-iid settings
- **Diversify the models more**
 - Develop classifiers that have some overlaps in certain classes but not exactly the same.
 - Build $f_1()$, ... $f_n()$ where f_i has different features and combine them using exp-AI ideas to build more robust models .
 - **Inspiration, Nature:** Two different genes can have an overlapping but not exact function.
 - If one gene fails, the other can help etc.
 - **Hypothesis:** class boundaries would not be the same across models, hence attacks will not directly transition over.

Future Directions

- **Digital twins security**
 - Digital twins have many use cases for modeling cyber-physical systems
 - Combination of control theoretical models and AI systems
 - Limited work on understanding attacking digital twins
 - Poisoning attacks
 - Evasion attacks
 - Robust deployment of digital twins for attack detection

Future Directions

- **When to ask humans?**
 - Robust estimation of uncertainty
 - Have more robust models to decide when to get the human expert involved
 - Best ways to convey uncertainty to humans?
 - Cross cutting work with ARL colleagues
 - Applicable in FL, robust ML or Digital twins setting