REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
26-07-2023	Final Report	18-Aug-2017 - 17-Jan-2021

4. TITLE AND SUBTITLE
Final Report: Robustness and Stability for Data Analysis in Security

5a. CONTRACT NUMBER
W911NF-17-1-0405

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 Wisconsin - Madison
Suite 6401
21 N Park Street
Madison, WI 53715 -1218

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)
69224-NC.1

12. DISTRIBUTION AVAILIBILITY 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 contrued 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	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Somesh Jha
UU	UU	UU	UU		19b. TELEPHONE NUMBER 608-262-9519

RPPR Final Report
as of 27-Jul-2023

Agency Code: 21XD

Proposal Number: 69224NC **Agreement Number: W911NF-17-1-0405**
INVESTIGATOR(S):

Name: Somesh Jha Ph.D.
Email: jha@cs.wisc.edu
Phone Number: 6082629519
Principal: Y

Organization: **University of Wisconsin - Madison**
Address: Suite 6401, Madison, WI 537151218
Country: USA
DUNS Number: 161202122 EIN: 396006492

Report Date: 17-Apr-2021 Date Received: 26-Jul-2023
Final Report for Period Beginning 18-Aug-2017 and Ending 17-Jan-2021
Title: Robustness and Stability for Data Analysis in Security
Begin Performance Period: 18-Aug-2017 **End Performance Period:** 17-Jan-2021
Report Term: 0-Other

Submitted By: Somesh Jha Email: jha@cs.wisc.edu
Phone: (608) 262-9519

STEM Degrees: **STEM Participants:**

Major Goals: Motivated by safety-critical applications, test-time attacks on classifiers via adversarial examples has recently received a great deal of attention. In this project we will undertake projects related to understanding this phenomenon deeply. The significant research thrusts are summarized in the next section.

Accomplishments: (1) Semantic View: There is a general lack of understanding on why adversarial examples arise; whether they originate due to inherent properties of data or due to lack of training samples remains ill-understood. For example, in an autonomous vehicle using deep learning for perception, not every adversarial example for the neural network might lead to a harmful consequence. Moreover, one may want to prioritize the search for adversarial examples towards those that significantly modify the desired semantics of the overall system. Along the same lines, existing algorithms for constructing robust ML algorithms ignore the specification of the overall system. In one line of ARO-funded research we explore that the semantics and specification of the overall system has a crucial role to play in this line of research.
(2) Using confidence measures: Another line of research investigates leveraging confidence information induced by adversarial training to reinforce adversarial robustness of a given adversarially trained model. Based on this insight, we devised Highly Confident Near Neighbor (HCNN), a framework that combines confidence information and nearest neighbor search, to reinforce adversarial robustness of a base model. Experimental results demonstrate that HCNN can enhance robustness by 40-50%.
(3) Theoretical analysis of k-nearest neighbors: This research thrust proposes a framework to analyze the robustness of a canonical non-parametric classifier - the k-nearest neighbors. Our analysis shows that its robustness properties depend critically on the value of k - the classifier may be inherently non-robust for small k, but its robustness approaches that of the Bayes Optimal classifier for

fast-growing k. Based on these insights a novel modified 1-nearest neighbor classifier is constructed, which guarantees robustness in the large sample limit. Experimental results suggest that this classifier may have good robustness properties even for reasonable data set sizes. Our experiments demonstrate that our algorithm performs better than or about as well as both standard 1-nearest neighbors and nearest neighbors with adversarial training – a popular and effective defense mechanism.

Training Opportunities: The main Ph.D student funded on this project is Uyeong Jang, who is in his 4th year. Occasionally we have put various other personnel to help with the projects.

RPPR Final Report
as of 27-Jul-2023

Results Dissemination: Note: Some of the papers appeared after the grant finished, but the work was started during the period of the grant. Moreover, the material is very relevant to the topic of the grant. This is not the complete list.

Introduction to the Special Issue on Automotive CPS Safety & Security: Part 1
S Chakraborty, S Jha, S Samii, P Mundhenk - ACM Transactions on Cyber-Physical Systems, 2023

Robust learning against relational adversaries
Y Wang, M Alhanahnah, X Meng, K Wang, S.Jha, NeurIPS, 2022

Overparameterization from computational constraints
S Garg, S Jha, S Mahloujifar, M Mahmoody, M Wang, NeurIPS, 2022

A quantitative geometric approach to neural-network smoothness
Z Wang, G Prakriya, S Jha, NeurIPS, 2022

Federated boosted decision trees with differential privacy
S Maddock, G Cormode, T Wang, C Maple, S Jha, ACM CCS, 2022

Graphite: Generating automatic physical examples for machine-learning attacks on computer vision systems
R Feng, N Mangaokar, J Chen, E Fernandes, S Jha, EuroSP, 2022

A separation result between data-oblivious and data-aware poisoning attacks
S Deng, S Garg, S Jha, S Mahloujifar, M Mahmoody, NeurIPS, 2021

A general framework for detecting anomalous inputs to dnn classifiers
J Raghuram, V Chandrasekaran, S Jha, S Banerjee, ICML, 2021

Face-off: Adversarial face obfuscation
V Chandrasekaran, C Gao, B Tang, K Fawaz, S Jha, ProPETS, 202

Concise explanations of neural networks using adversarial training
P Chalasani, J Chen, AR Chowdhury, X Wu, S Jha, ICML 2021

Informative outlier matters: Robustifying out-of-distribution detection using outlier mining
J Chen, Y Li, X Wu, Y Liang, S Jha, ECML, 202

Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning
S Yeom, I Giacomelli, A Menaged, M Fredrikson, S Jha, Journal of Computer Security, 2020

Detecting adversarial examples using data manifolds
S Jha, U Jang, S Jha, B Jalaian, MILCOM 2018.

Tommaso Dreossi, Somesh Jha, Sanjit A. Seshia:
Semantic Adversarial Deep Learning. CAV, 2018.

Yizhen Wang, Somesh Jha, Kamalika Chaudhuri:
Analyzing the Robustness of Nearest Neighbors to Adversarial Examples.
ICML 2018.

Xi Wu, Uyeong Jang, Jiefeng Chen, Lingjiao Chen, Somesh Jha:
Reinforcing Adversarial Robustness using Model Confidence Induced by Adversarial Training.
ICML 2018.

RPPR Final Report
as of 27-Jul-2023

Honors and Awards: During the project, PI Jha became an IEEE fellow, ACM fellow, and AAAS fellow.

Protocol Activity Status:

Technology Transfer: Nothing to report.

PARTICIPANTS:

Participant Type: PD/PI
Participant: Somesh Jha
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Uyeong Jang
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Somesh Jha
Signature Date: 7/26/23 12:08PM

There are several publications related to this grant.
They can be found on. There were too many publications to upload.

https://scholar.google.com/citations?user=BaI7l8QAAAAJ&hl=en&oi=ao