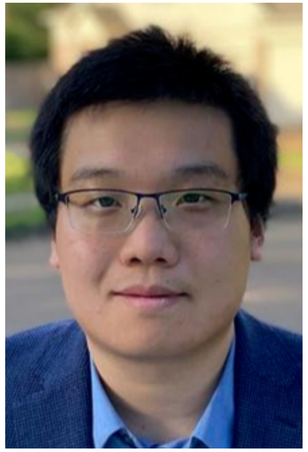
Black-box Adversarial Attacks on Network-wide Multi-step Traffic State Prediction Models

Bibek Poudel and Weizi Li



IEEE ITSC 2021
INDIANAPOLIS, USA

THE UNIVERSITY OF MEMPHIS

Outline

- Introduction
- Network-wide multi-step prediction
- Methodology
- Results
- Conclusion and Future Work

Introduction

Introduction

- Traffic states (density, speed, flow): describe interactions between traffic and infrastructure
- Traffic state prediction: critical for many ITS applications

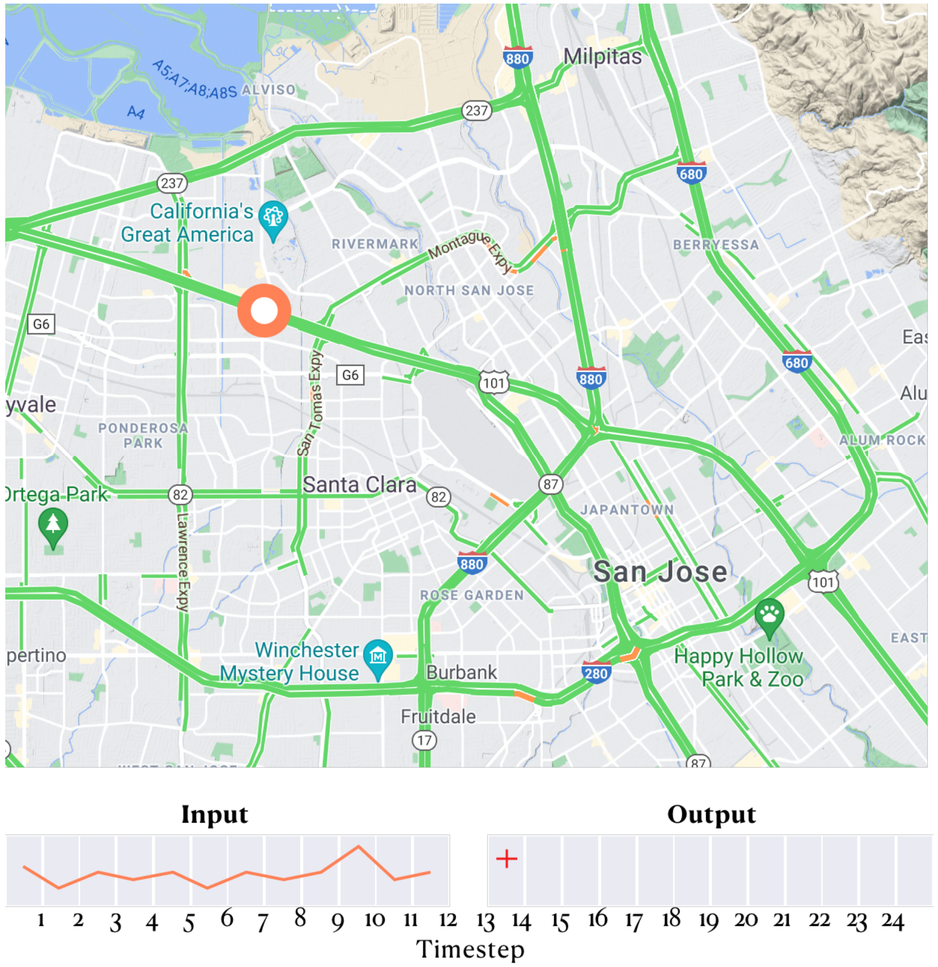
Introduction

- Traffic states (density, speed, flow): describe interactions between traffic and infrastructure
- Traffic state prediction: critical for many ITS applications
- Both traditional and deep learning (state-of-the-art) techniques
- Adversarial attack ➔ Performance degradation
- Deployment: robustness to adversarial attacks

Network-wide multi-step traffic state prediction

Traffic state prediction

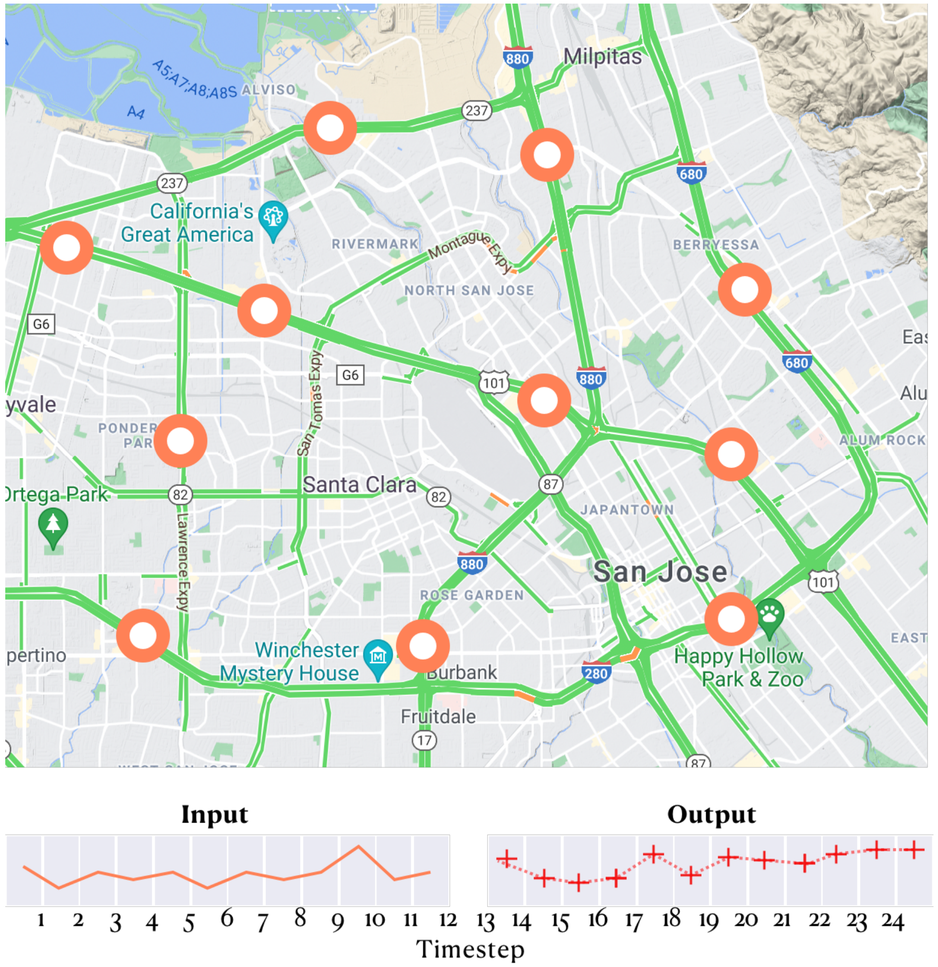
- Traffic state prediction
  - Single node, single-step



Single-node single-step flow prediction
Map source: Google Maps

Traffic state prediction

- Traffic state prediction
  - Single node, single-step
  - Single node, multi-step
  - Network-wide, single-step
  - Network-wide, multi-step



Network-wide multi-step flow prediction
Map source: Google Maps

Network-wide multi-step traffic state prediction

- Traditional techniques
  - Linear Regression
  - Historical Average
  - Vector Autoregression

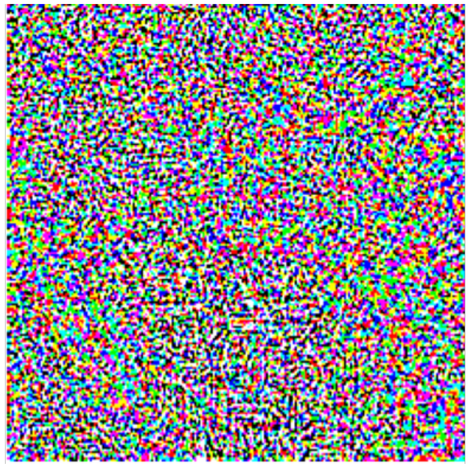
Network-wide multi-step traffic state prediction

- Traditional techniques
  - Linear Regression
  - Historical Average
  - Vector Autoregression
- Deep Learning techniques
  - Sequence-to-Sequence model
  - Diffusion Convolution Recurrent Neural Network (DCRNN, Li et al., 2018)
  - Graph Convolution Gated Recurrent Neural Network (GCGRNN, Lin et al., 2021)

Methodology

Adversarial attack

\+ 0.01 x



=

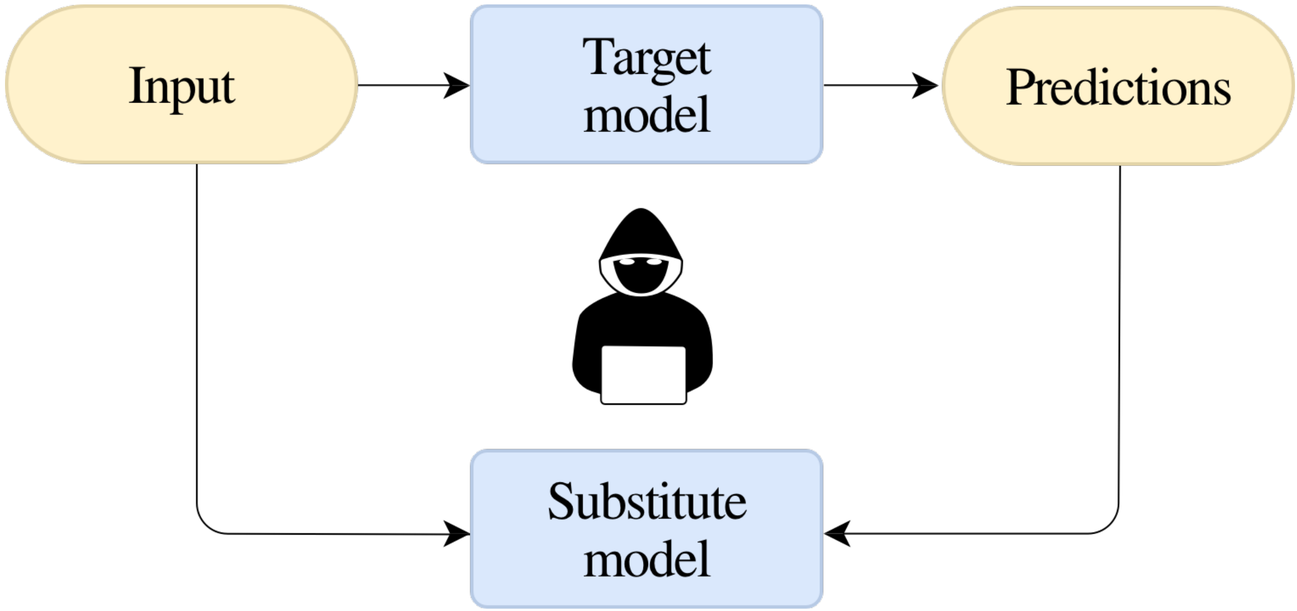
Original Input
Output = Stop

Added
Perturbations

Adversarial Input
Output = Speed limit 45

Adversarial traffic sign
Image source: unsplash.com, @iman_br

Black-box adversarial attack



Attack algorithms

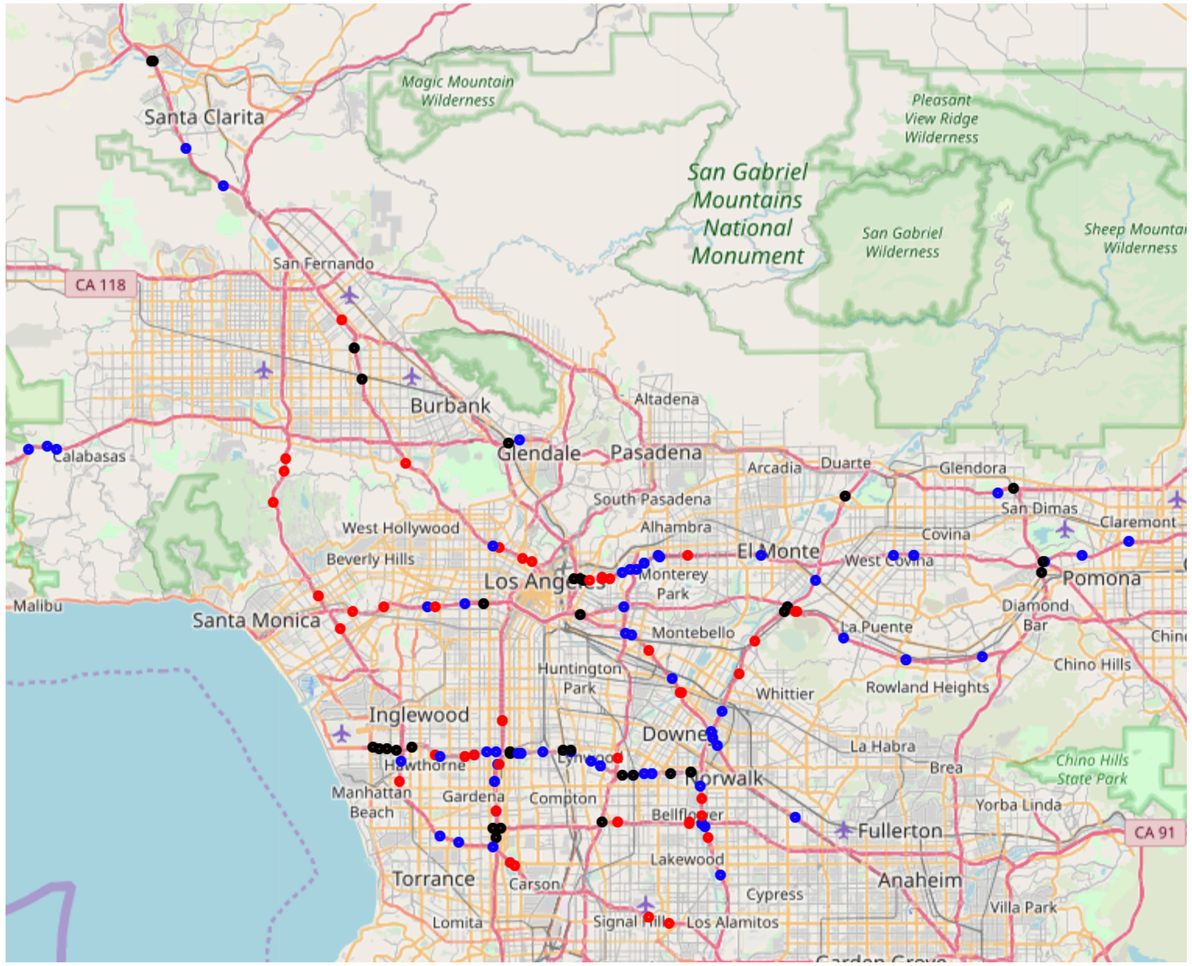
- Gradient-based
- Perturbations maximize the loss

Attack algorithms

- Gradient-based
- Perturbations maximize the loss
- Fast Gradient Sign Method (FGSM):
  - Single step
- Basic Iterative Method (BIM):
  - Iterative
  - More computation

Dataset

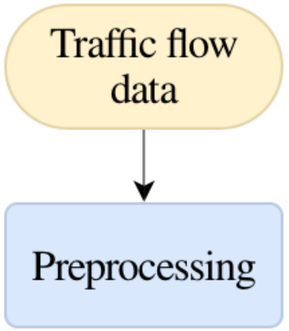
- Caltrans PeMS
- Hourly dataset of Los Angeles
- 150 nodes, 12 timestep
- 6 months
- Total 13,000 data points



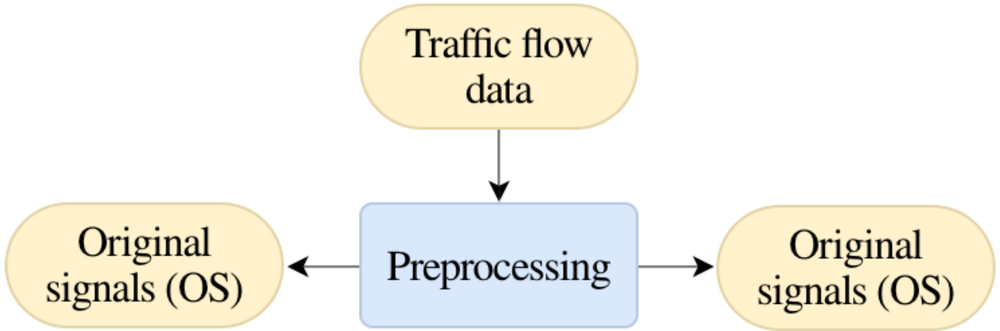
150 traffic nodes in Los Angeles

Image source: GCGRNN, Lin et al.

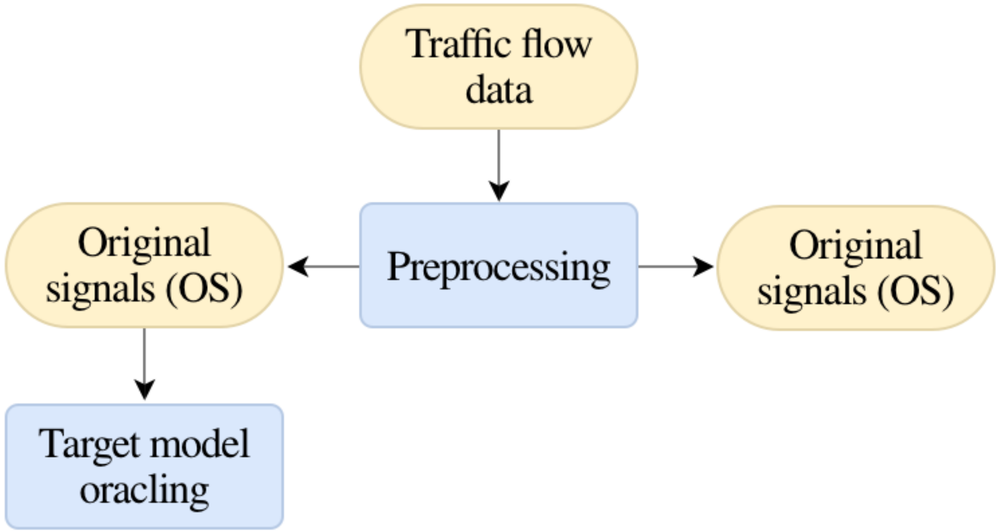
Systematic Diagram



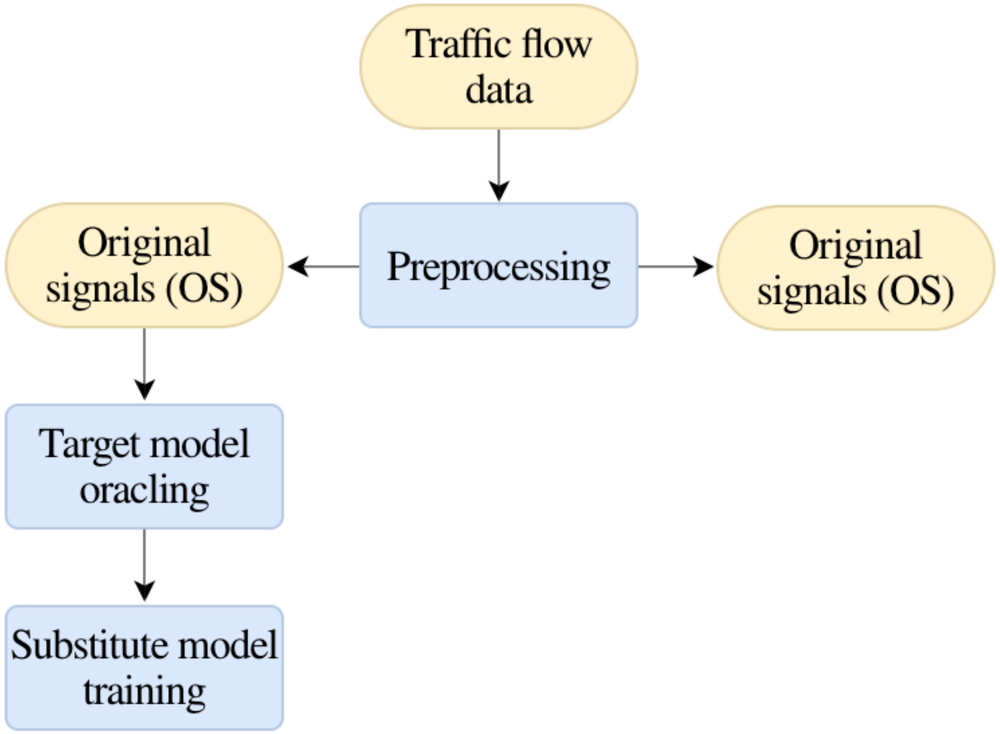
Systematic Diagram



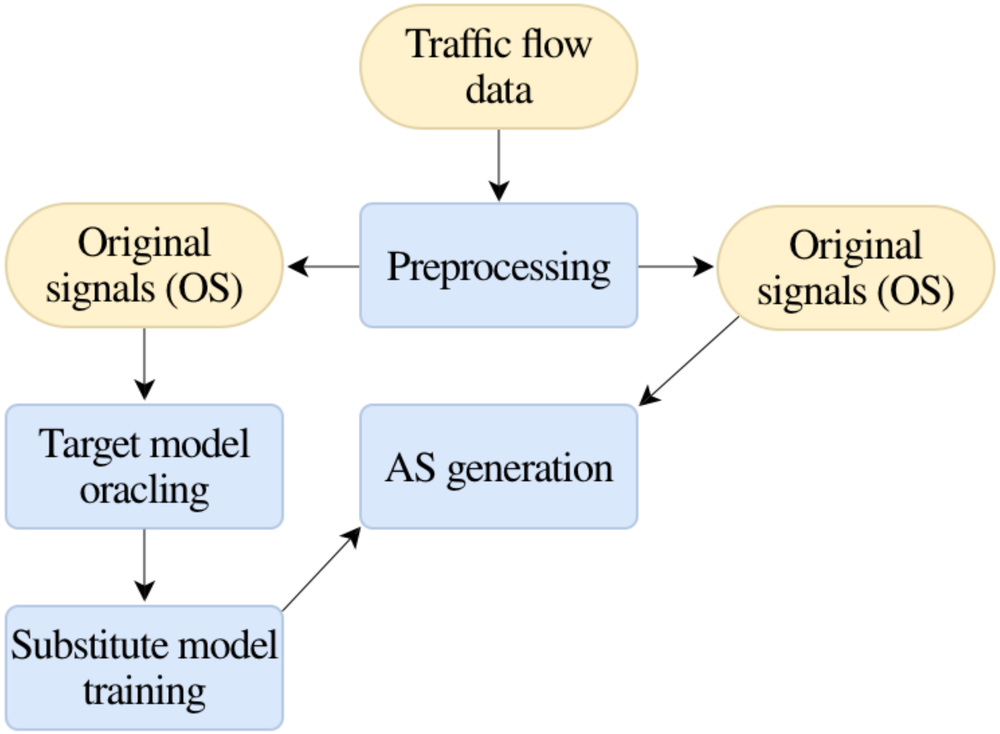
Systematic Diagram



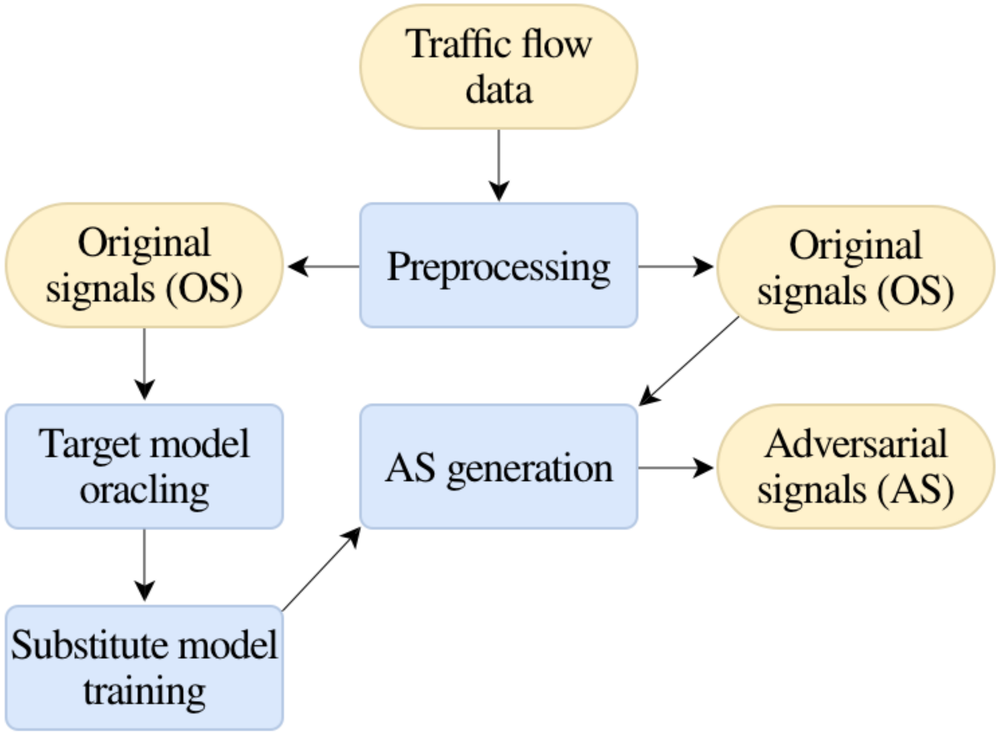
Systematic Diagram



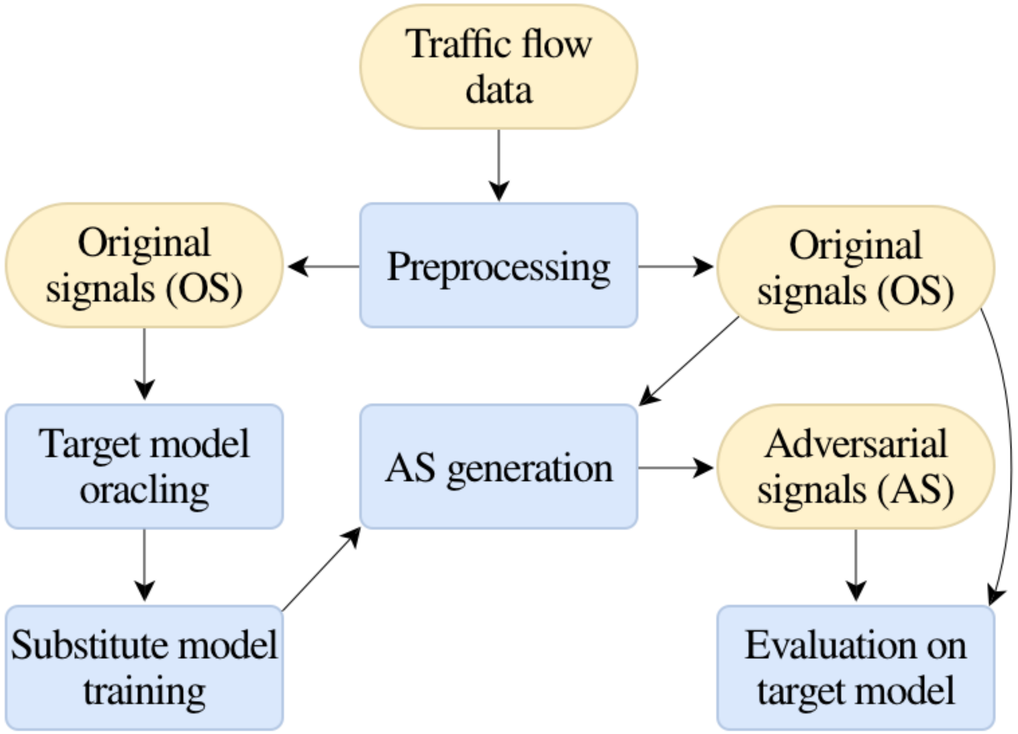
Systematic Diagram



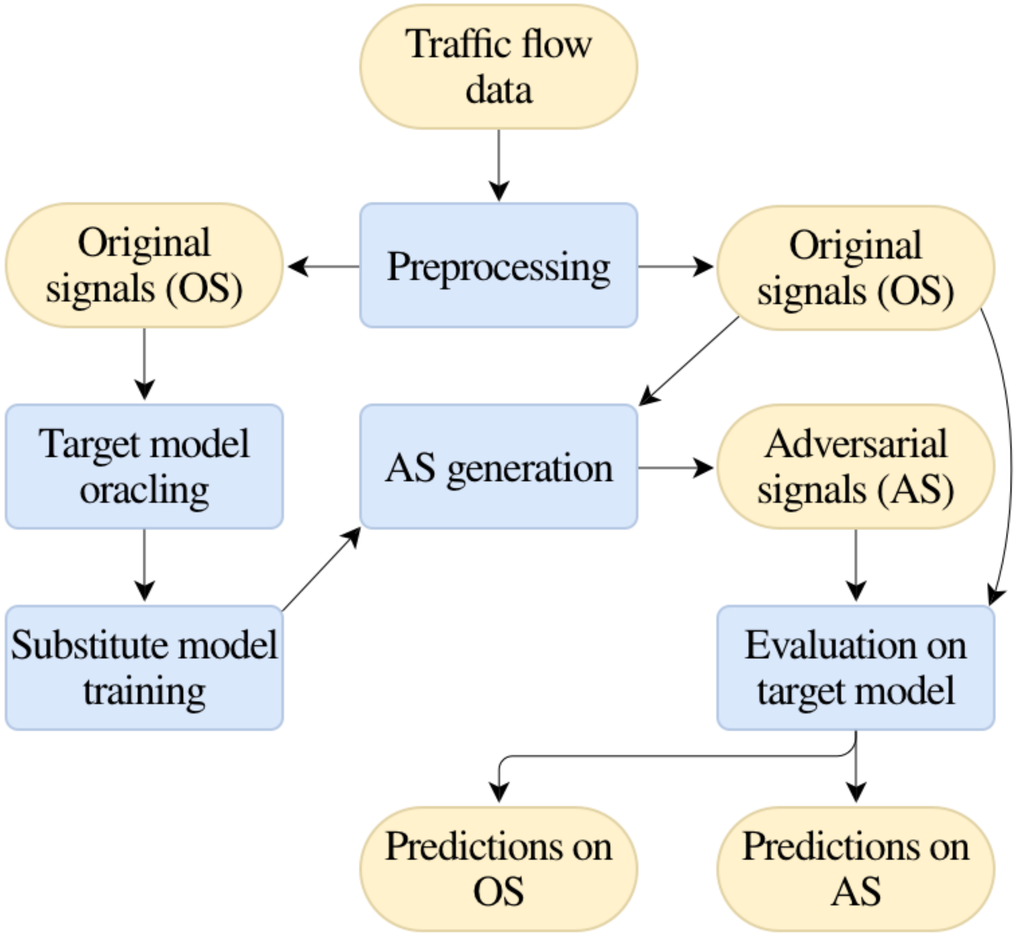
Systematic Diagram



Systematic Diagram



Systematic Diagram



Results

Results

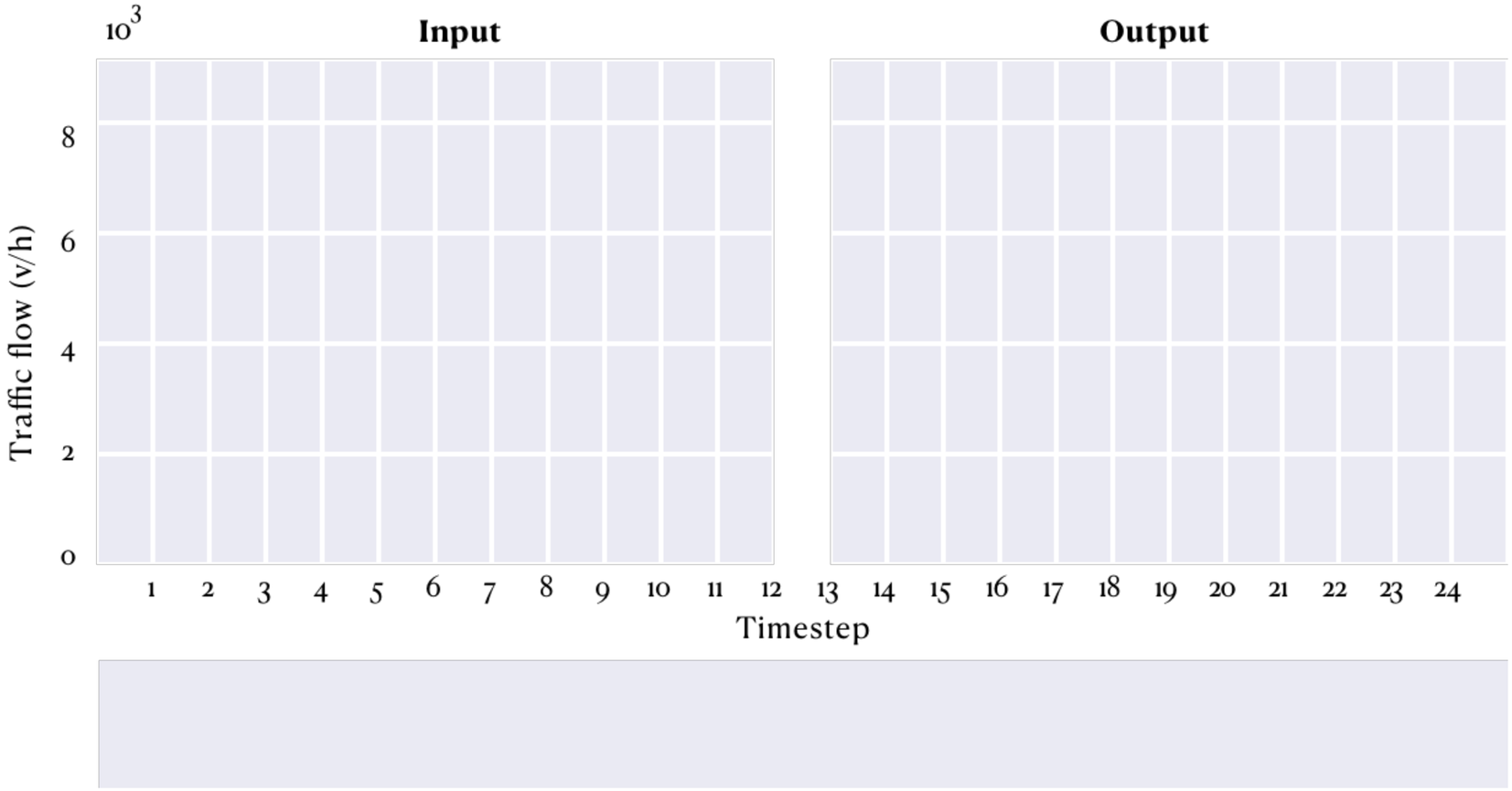
4 Target Models x 2 Attack Algorithms

Traditional: LR , HA

Deep Learning: DCRNN, GCGRNN

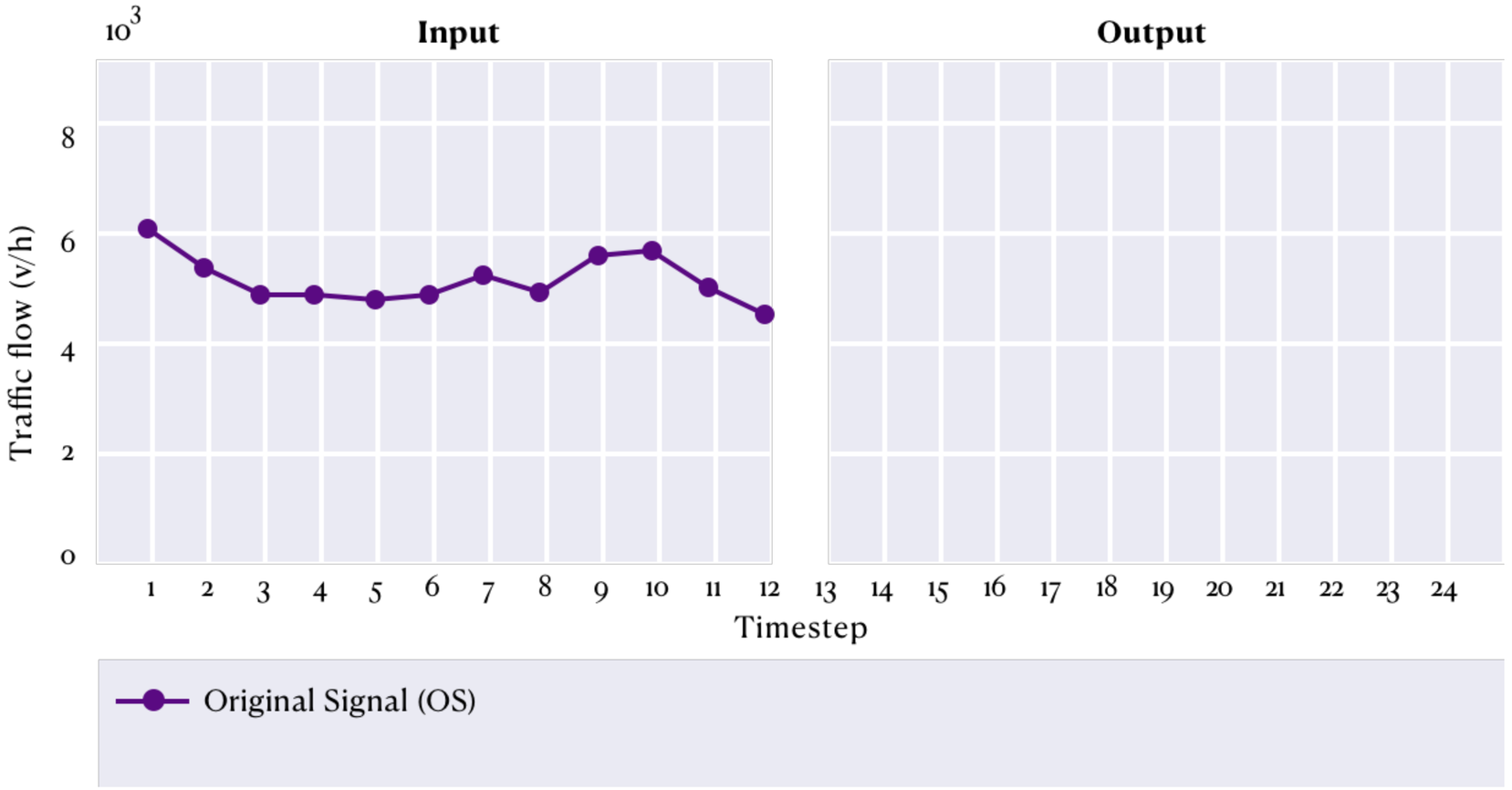
FGSM
BIM

Results



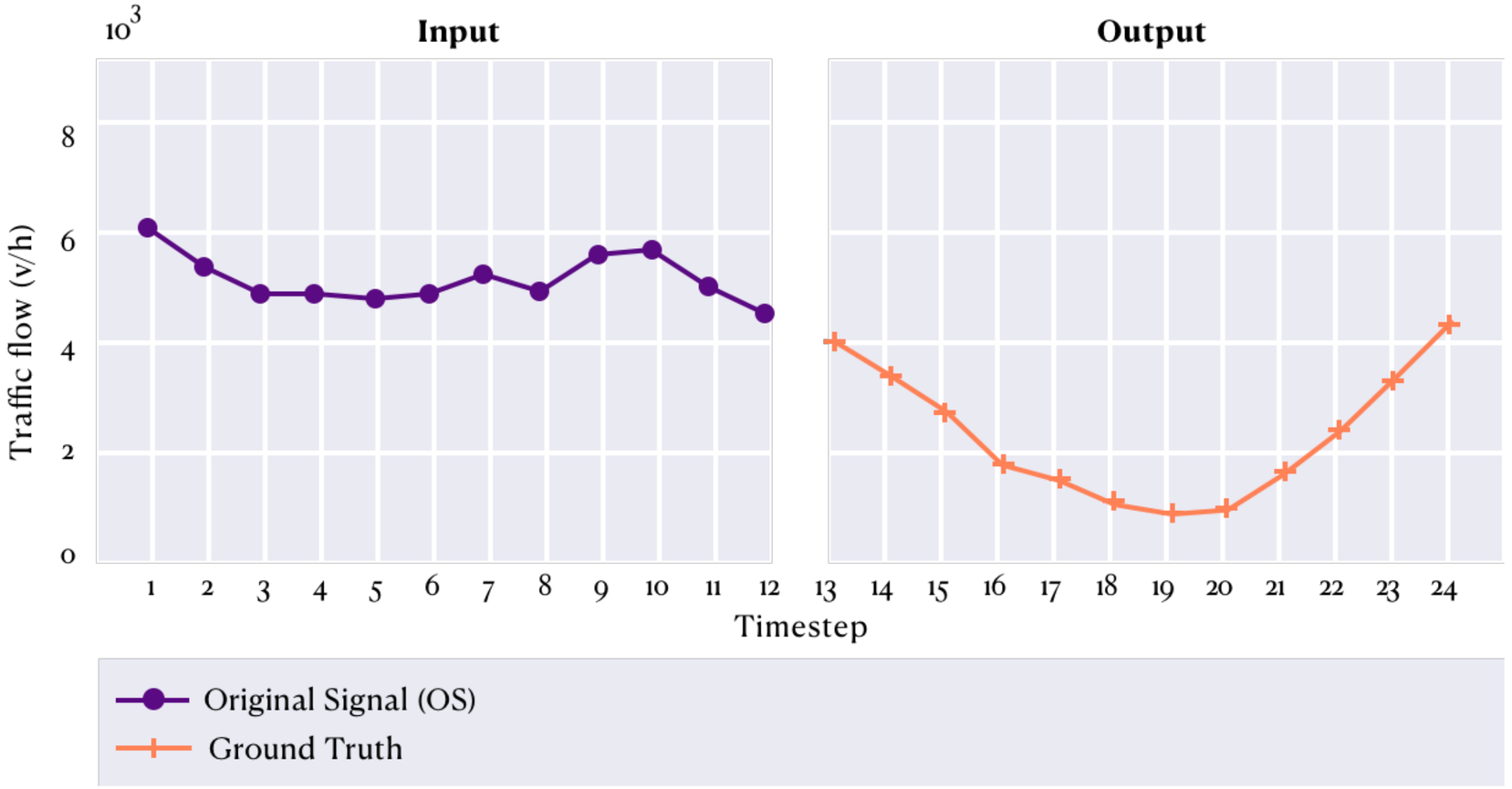
Performance degradation of the GCGRNN model by an adversarial signal from FGSM

Results



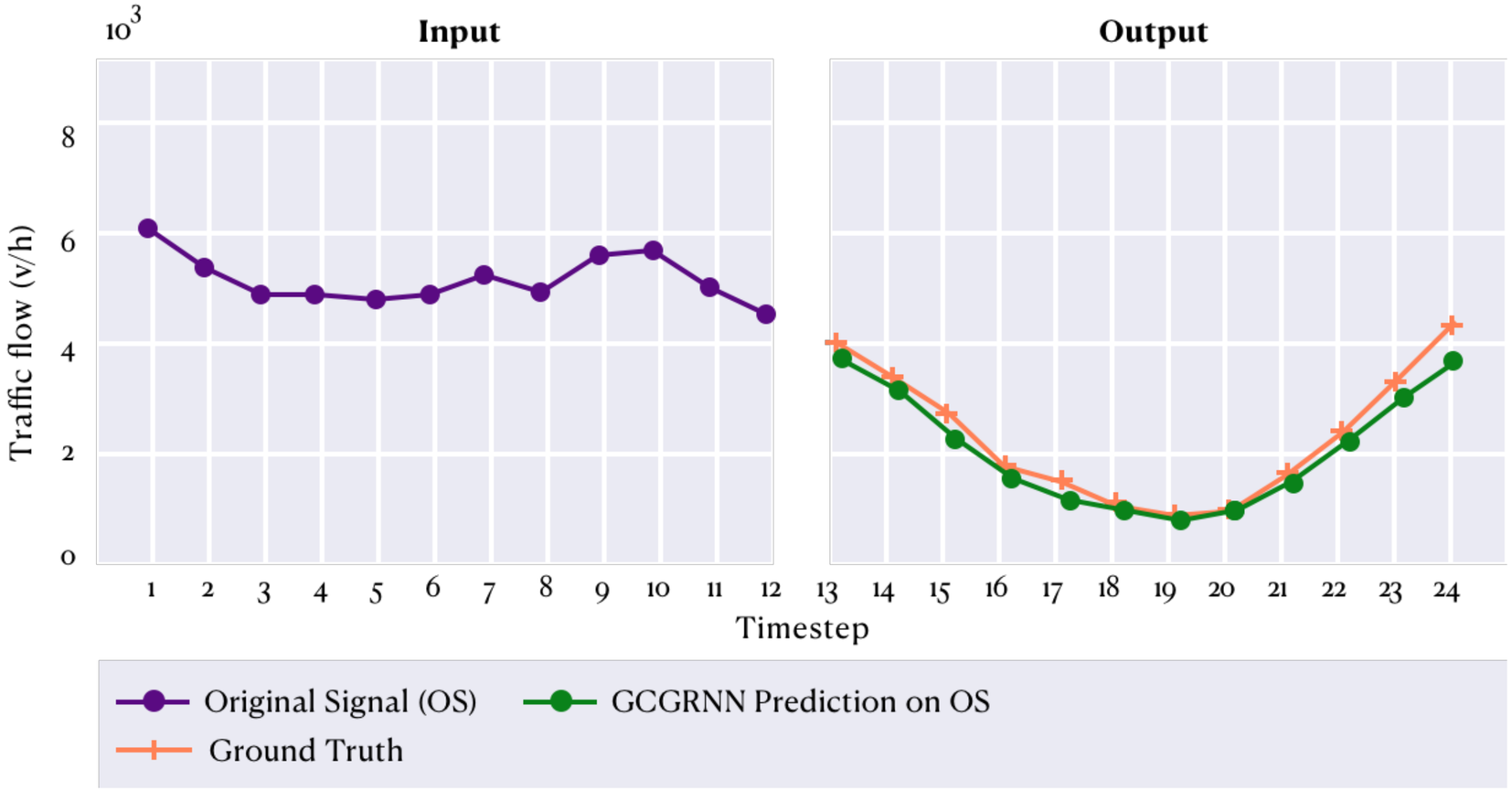
Performance degradation of the GCGRNN model by an adversarial signal from FGSM

Results



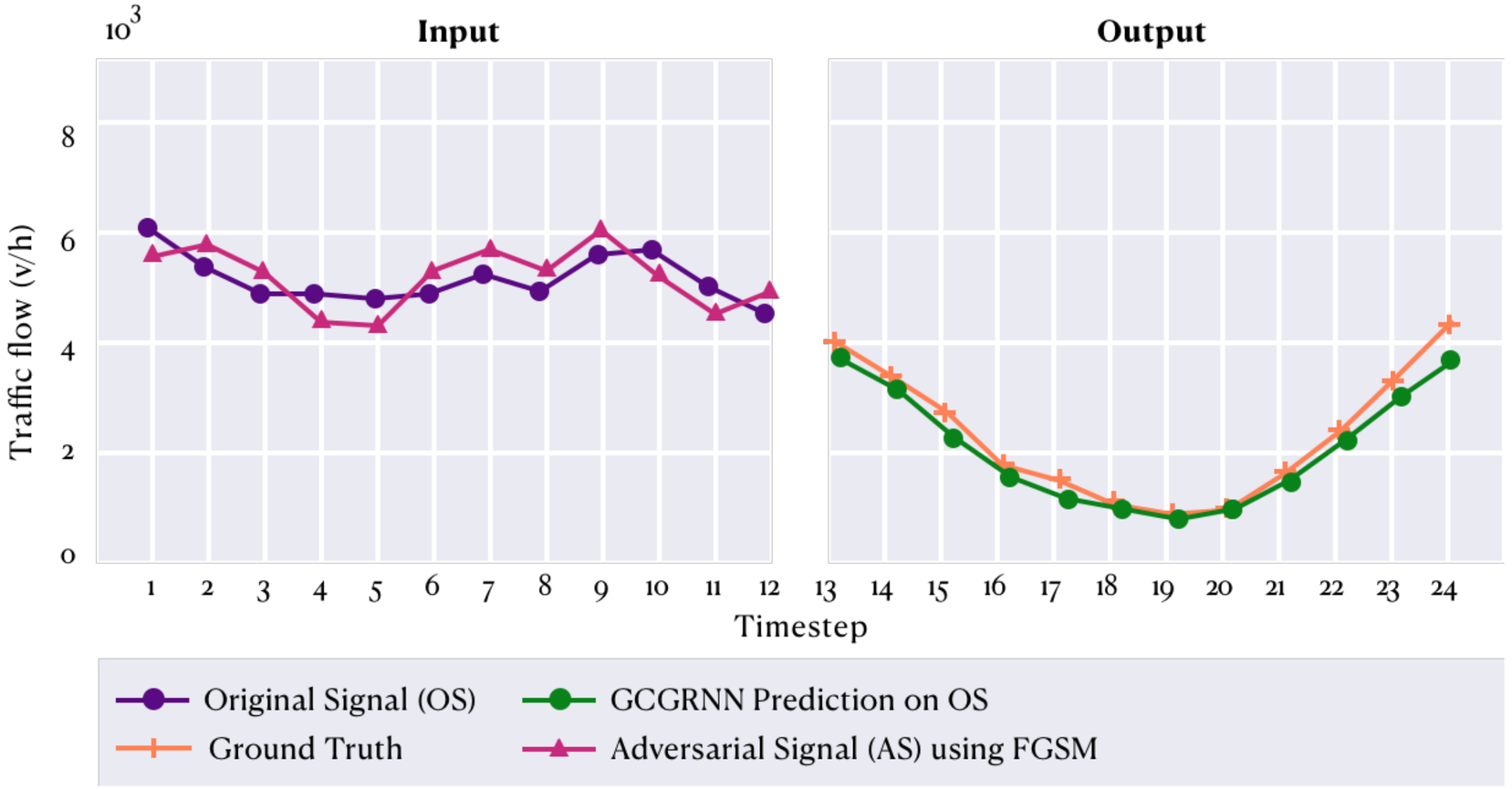
Performance degradation of the GCGRNN model by an adversarial signal from FGSM

Results



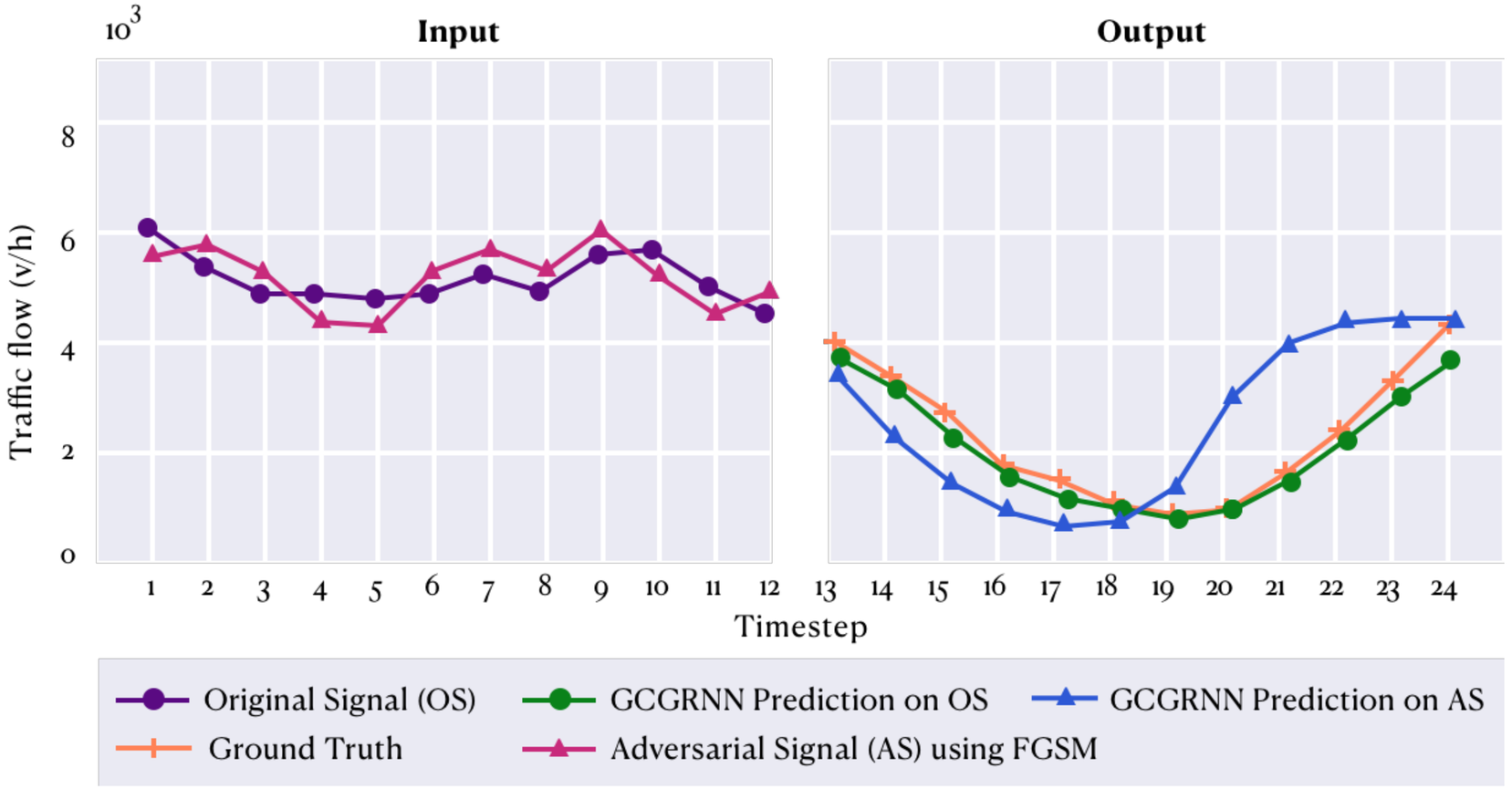
Performance degradation of the GCGRNN model by an adversarial signal from FGSM

Results



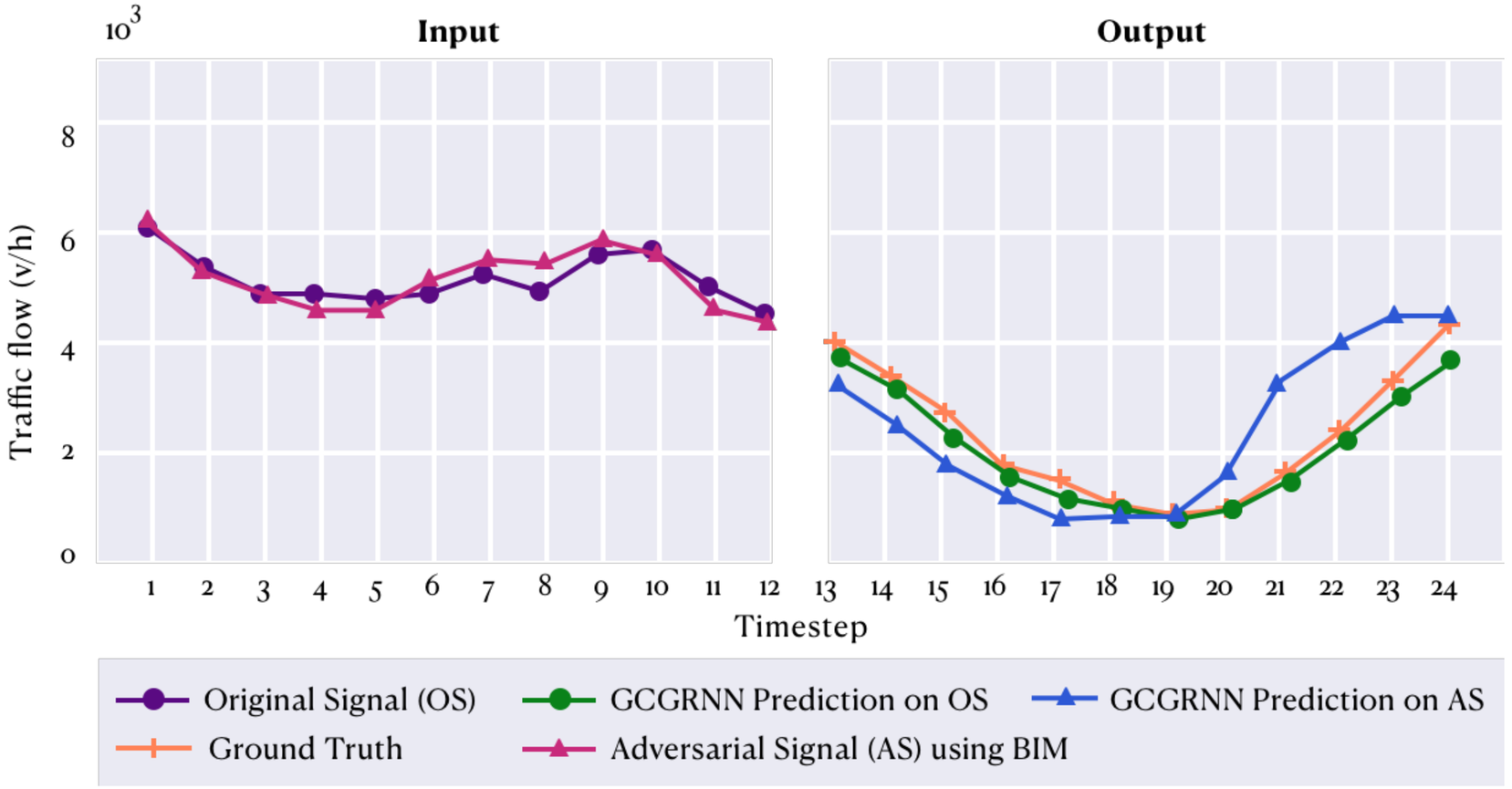
Performance degradation of the GCGRNN model by an adversarial signal from FGSM

Results



Performance degradation of the GCGRNN model by an adversarial signal from FGSM

Results



Performance degradation of the GCGRNN model by an adversarial signal from BIM

Results

	Change on original signal (L2)		Change on prediction (L2)	
Model	FGSM	BIM	FGSM	BIM
GCGRNN	3.35	1.67	3.27	1.80
DCRNN	3.35	1.45	17.5	10.1
LR	3.35	1.45	3.16	1.38
HA	132	129	0	0

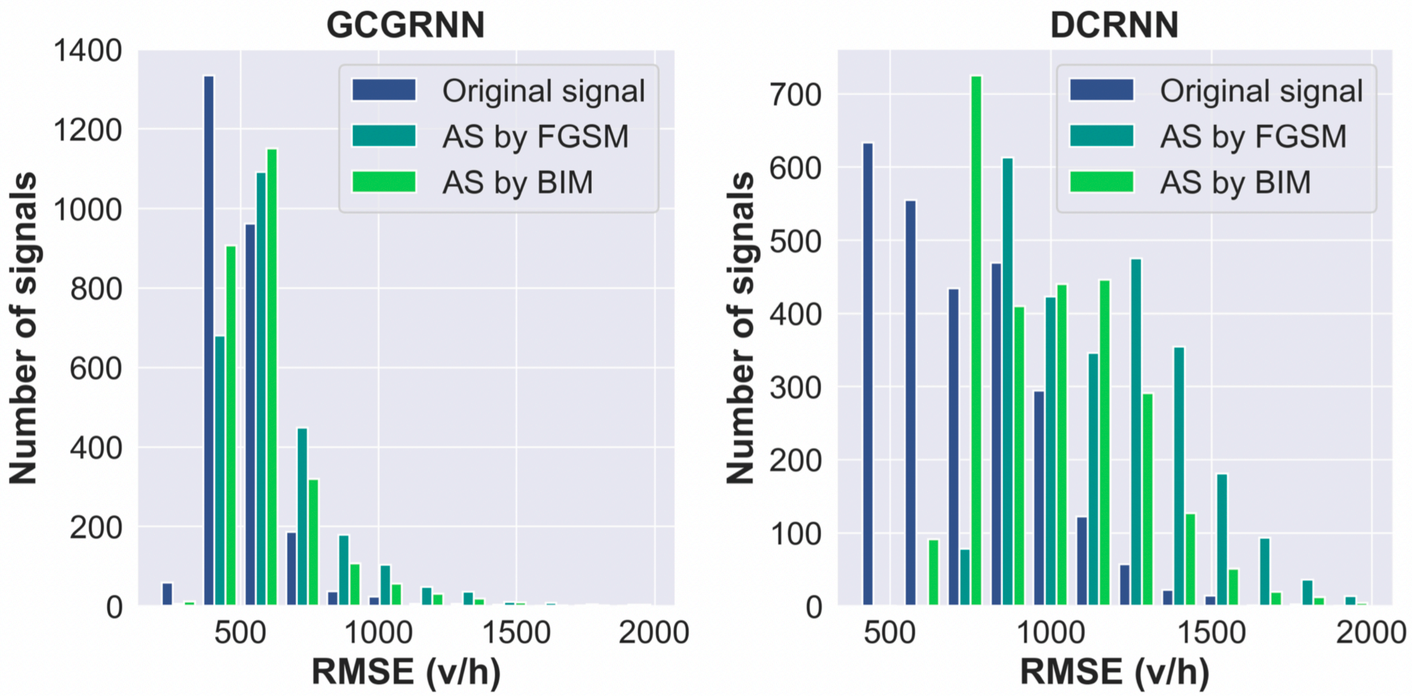
Average changes in original signals and corresponding average changes in predictions

Results

	Change on original signal (L2)		Change on prediction (L2)	
Model	FGSM	BIM	FGSM	BIM
GCGRNN	3.35	1.67	3.27	1.80
DCRNN	3.35	1.45	17.5	10.1
LR	3.35	1.45	3.16	1.38
HA	132	129	0	0

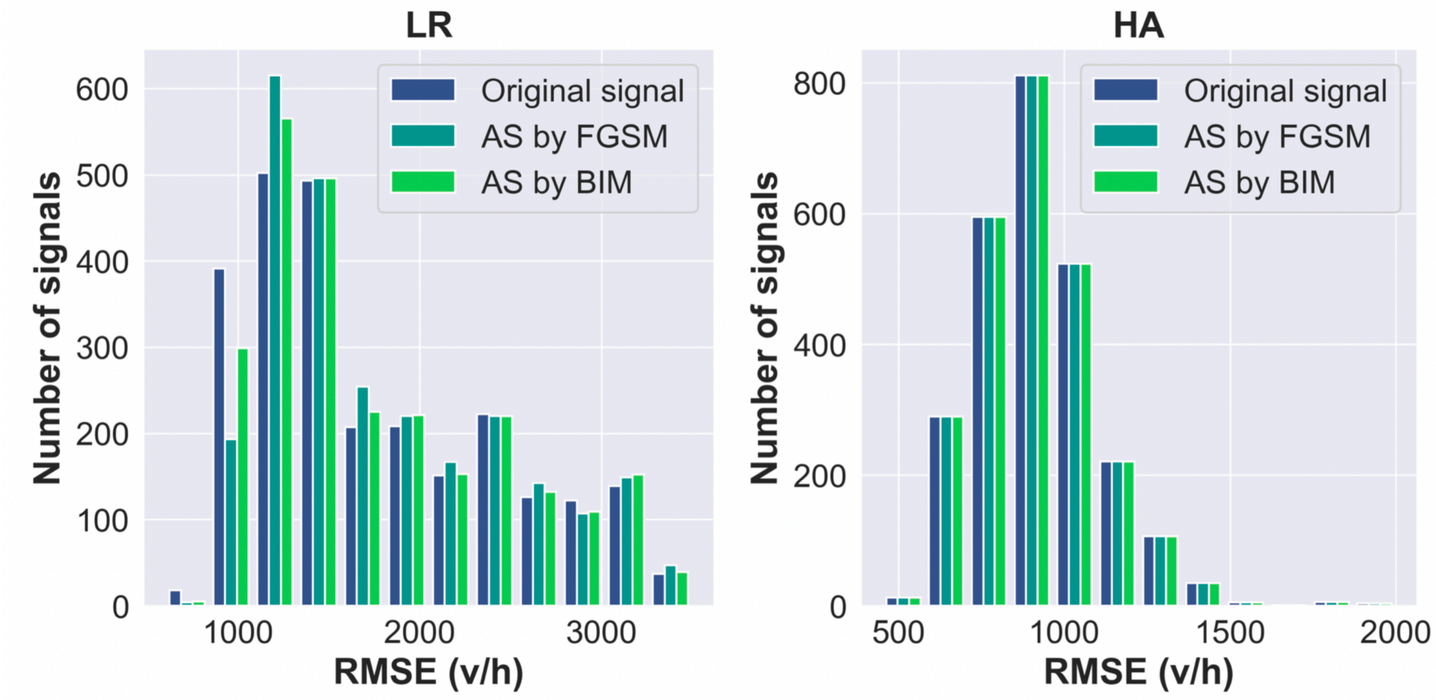
Average changes in original signals and corresponding average changes in predictions

Results



Target models' RMSE Distribution on Original and Adversarial signals as inputs

Results



Target models' RMSE Distribution on Original and Adversarial signals as inputs

Conclusion and Future Work

Conclusion & Future Work

- Conclusion
  - Higher accuracy ≠ Higher adversarial robustness
  - DCRNN ➔ 54%, GCGRNN ➔ 26%, LR ➔ 2.5%, HA ➔ No degradation

Conclusion & Future Work

- Conclusion
  - Higher accuracy ≠ Higher adversarial robustness
  - DCRNN ➔ 54%, GCGRNN ➔ 26%, LR ➔ 2.5%, HA ➔ No degradation
- Future Work
  - Develop defense techniques
  - Attack connected autonomous vehicles

Thank You!

Bibek Poudel
bpoudel@memphis.edu