Università di Pisa

National Lab RaSS Radar and Surveillance Systems

cnit consorzio nazionale interuniversitario per le telecomunicazioni

Advancing Radar Cybersecurity: Defending Against Adversarial Attacks in SAR Ship Recognition Using Explainable AI and Ensemble Learning

Amir Hosein Oveis (1) , Giulio Meucci (1,2), Francesco Mancuso(1,2), Alessandro Cantelli-Forti (1)
(1) Radar and Surveillance Systems (RaSS) Laboratory (CNIT), Pisa, Italy
(2) Department of Information Engineering, University of Pisa, Pisa, Italy

Overview

01 Introduction

02 FGSM /LIME

03 Proposed Framework

04 Real Data Analysis

05 Conclusions

Introduction

- Vulnerability of ***Synthetic Aperture Radar*** (***SAR***)-based ***ship recognition*** models to ***adversarial attacks***.

- Fast Gradient Sign Method (***FGSM***) to generate adversarial examples
  - Adding ***perturbations*** to SAR ship images to
  - ***mislead*** a ***pre-trained convolutional neural network*** (***CNN***).

- To analyze the impact of these attacks:
  - Local Interpretable Model-agnostic Explanations (***LIME***) algorithm.
  - An ***Explainable Artificial Intelligence*** (***XAI***) method.
  - To ***explain*** the ***contributing area*** in the input image to the CNN's ***decision-making*** process under adversarial conditions.

Introduction

- Finally, we propose an ***ensemble learning*** strategy
  - Combining multiple ***transfer learning-based*** architectures
  - To enhance the ***robustness*** of ship recognition systems
  - Against adversarial examples and ***mitigate*** their ***transferability***.
- Our real data experiment is conducted on ***OpenSARShip*** dataset:
  - Consists of different ship images extracted from 41 images captured by Sentinel-1 SAR satellite.

Introduction

- Synthetic Aperture Radar (***SAR***):
  - one of the most powerful sensors in the ***remote sensing*** field
  - ***high-resolution*** images regardless of ***weather conditions***.
  - Instead of using a ***physically large antenna*** to improve ***resolution***,
    - SAR ***synthesizes*** a much larger, ***virtual antenna***
    - by combining ***radar signals*** collected over time
    - as the platform (e.g., aircraft or satellite) ***moves*** along its flight path.

- While CNNs perform well in SAR ship recognition,
  - their ***decision-making*** processes are not clear.
  - lack of ***transparency*** can make it difficult to ***rely on*** the CNN's decision,
  - especially in ***critical applications*** like ***maritime surveillance***.
  - ***eXplainable AI (XAI)*** techniques to provide ***explanations*** for the ***model's decision***

Introduction

- Questions:
  - ***XAI's*** behavior under ***adversarial attacks***?
  - How to enhance model ***robustness*** and resilience against such attacks?
- What is an ***adversarial attack***?
  - ***Very small changes*** added to the input data
  - To force the model into ***misclassification***
  - Can be ***imperceptible*** even to experts.
  - Can be ***transferable*** to other models (DNNs and even traditional classifiers)

FGSM

- Fast Gradient Sign Method (***FGSM***): A well-known technique to generate ***adversarial samples***
- ***Non-targeted*** specific incorrect class does not matter, just incorrect! (≠ Targeted)
- ***white-box***: full knowledge of the model's architecture and parameters is available for generation of adversarial samples
- ***evasion attack***: deceiving a pre trained model - without poisoning the training data
- ε: ***scaling*** factor for the perturbation
- trade-off:
  - its too small values might ***fail*** to fool the network,
  - its too large values could lead to an ***easily detectable*** image, which raises ***suspicion*** or even the possibility of being ***filtered*** by ***defensive*** algorithms

LIME

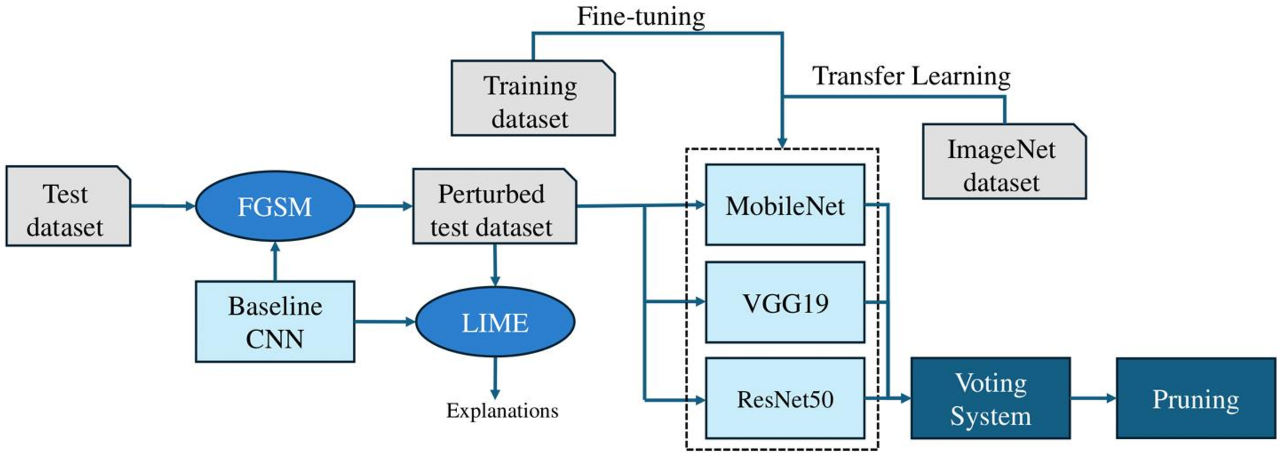
- ***LIME:*** Local Interpretable Model-Agnostic Explanations
  - To ***explain*** the predictions of ***any complex machine learning models*** and
  - Understand their ***decision-making*** process.
- In image classification task:
  - LIME highlights the most influential ***superpixels*** (***features***) of the input image
  - that ***contribute*** to the model's ***decision***.
- LIME generates ***perturbed versions*** of the input image
  - by randomly masking different regions
  - turning superpixels ***on*** and ***off*** and observes how these changes ***affect*** the model's predictions.
- These perturbed instances are ***passed through*** the model
  - their corresponding predictions are collected.

LIME

- These perturbed instances, along with their corresponding predictions:
  - are then used to ***fit*** a ***simple***, ***interpretable*** model like a ***linear regression***.
- This ***surrogate model*** g:
  - Locally ***approximates*** the ***behavior*** of the ***original model*** f
  - in the ***local neighborhood*** of the ***input image***.

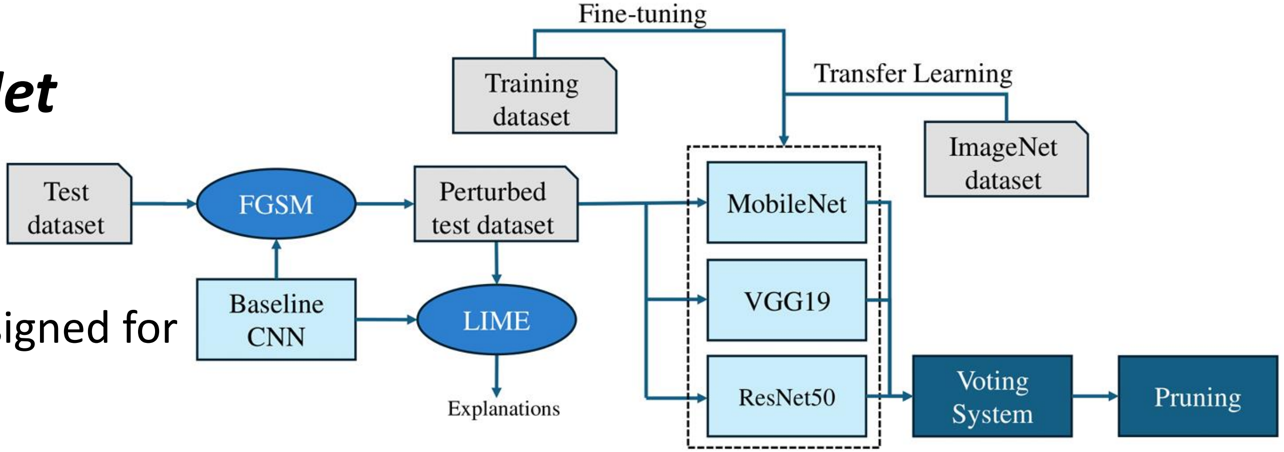
SYSTEM OVERVIEW

1. Training a baseline ***CNN*** model (using the training dataset.)
2. Applying ***FGSM*** method
   a. by an adversary in practical scenarios,
   b. to generate ***adversarial perturbations***
   c. aimed at ***confusing*** the ***pre-trained*** baseline model
   d. when applied to the ***test dataset***.
3. Inputs to the ***LIME*** method for generating explanations:
   a) All ***perturbed test samples*** at different levels of perturbation,
   b) along with ***the pre-trained model***



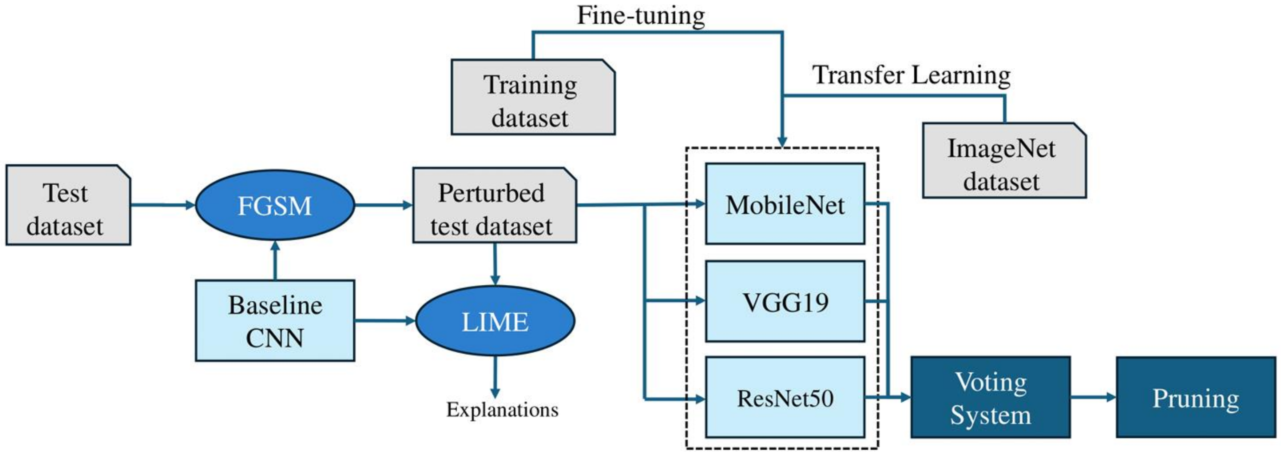
SYSTEM OVERVIEW

- ***LIME***: treats the pre-trained model as a ***black-box***
  - Enables: understand the ***most important part*** of the ***input*** images.
- Transfer learning: ***VGG19***, ***ResNet50***, and ***MobileNet***
  - Pre-trained on the ***ImageNet*** dataset
  - ***fine-tuned*** using the ***training*** dataset
  - Evaluate using the ***perturbed test dataset*** (designed for the ***baseline*** model.)
  - Assess the ***transferability*** of perturbed samples across different models.
  - A ***voting*** mechanism:
    - ***ensemble*** learning strategy
    - Selecting the ***most frequently predicted class*** among the models.



SYSTEM OVERVIEW

- A ***rejection*** mechanism:
  - labeling a prediction as ***unreliable***
  - if ***significant variation*** exists among predictions from different models.
- This ***rejection mechanism***
  - enhance the ***robustness*** and ***reliability*** of the ***ensemble*** predictions,
  - particularly in scenarios where the ***cost*** of ***misclassification*** can be high.



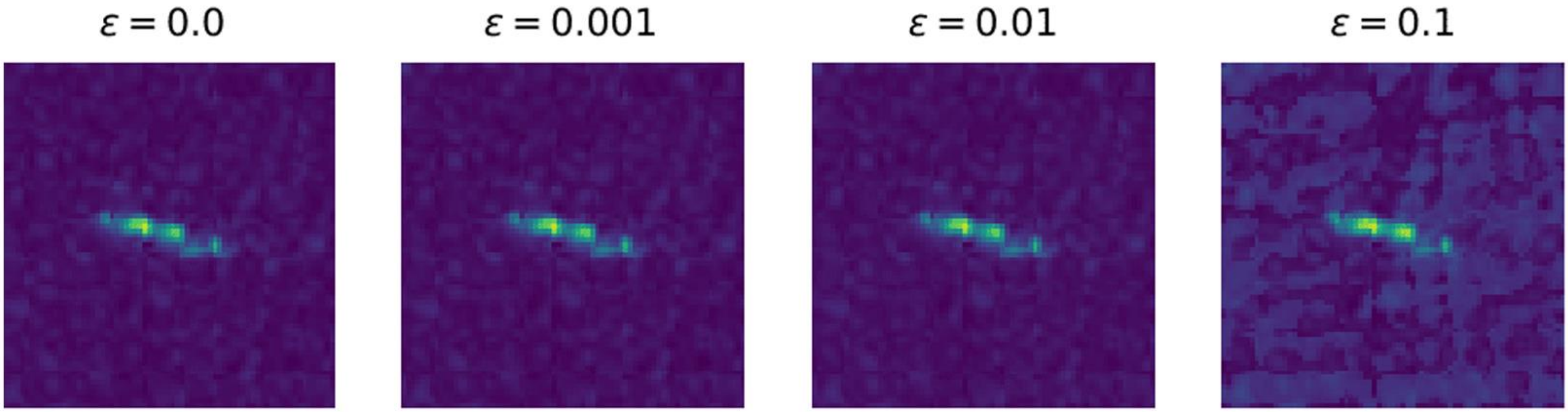
Real Data Analysis

- ***OpenSARShip-v1***:
  - Ship patches extracted from 41 Sentinel-1 C band ***SAR*** satellite images captured under different conditions.
  - 11346 SAR ship chips
  - We constructed:

    a balanced ***three-category*** scenario: bulk carriers, container ships, and tankers
  - We used ***169 training*** images per class (in total 507 images).
  - Curated the test dataset by selecting ***120 correctly classified*** images per class (overall 360 images).
  - Establish a starting point of ***100% overall accuracy*** before introducing adversarial

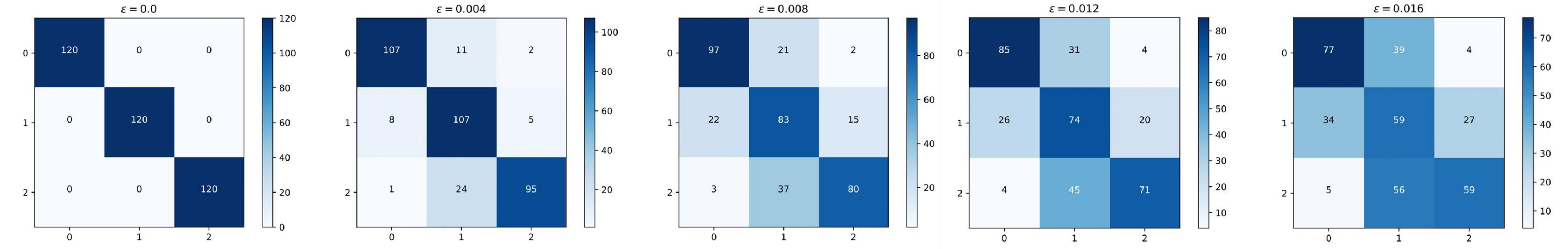
Real Data Analysis

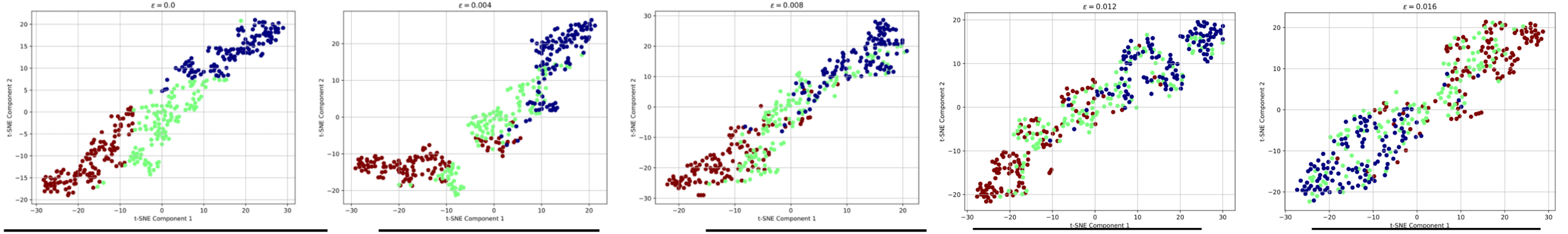
- **Visual** effects of **FGSM-generated** adversarial perturbations on **a test image** from the OpenSARShip-v1 database



Real Data Analysis

- The **effects** of increasing adversarial **perturbation levels**





Class	Precision	Recall
0	1	1
1	1	1
2	1	1
OA = 1		

Precision	Recall
0.92	0.89
0.75	0.89
0.93	0.79
OA = 0.86	

Precision	Recall
0.80	0.81
0.59	0.69
0.82	0.67
OA = 0.72	

Precision	Recall
0.74	0.71
0.49	0.62
0.75	0.59
OA = 0.64	

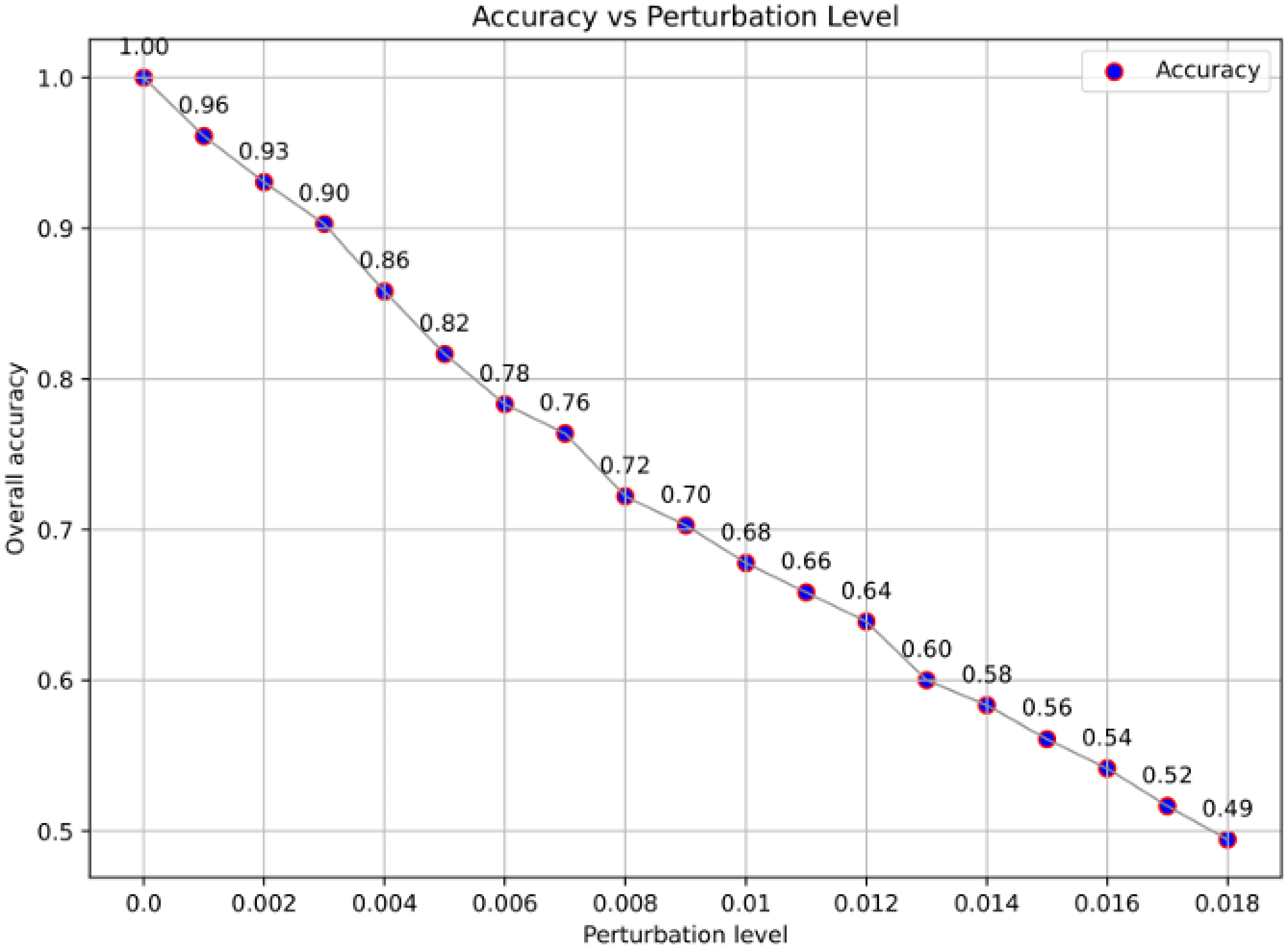
Precision	Recall
0.66	0.64
0.38	0.49
0.66	0.49
OA = 0.54	

Real Data Analysis

- ***t-SNE***: a statistical nonlinear ***dimensionality reduction*** technique for embedding ***high-dimensional*** data for ***visualization*** in a low-dimensional space of ***two*** or ***three*** dimensions.
- In our analysis:
  - to visualize how ***similarities*** between test samples are affected by ***perturbations***
- The first column:
  - ***perturbation free*** scenario:
  - ***well-separated features*** and an ***ideal accuracy of 1*** (data-selection to ***isolate*** the specific impact of perturbations)
- As perturbation levels increase:
  - features in the ***t-SNE*** plot becoming ***less distinct***,
  - ***confusion matrices*** become ***less diagonal***,
  - ***precision*** and recall ***metrics*** in the classification reports ***deteriorate***.
  - model's ability to differentiate between classes is ***compromised***.

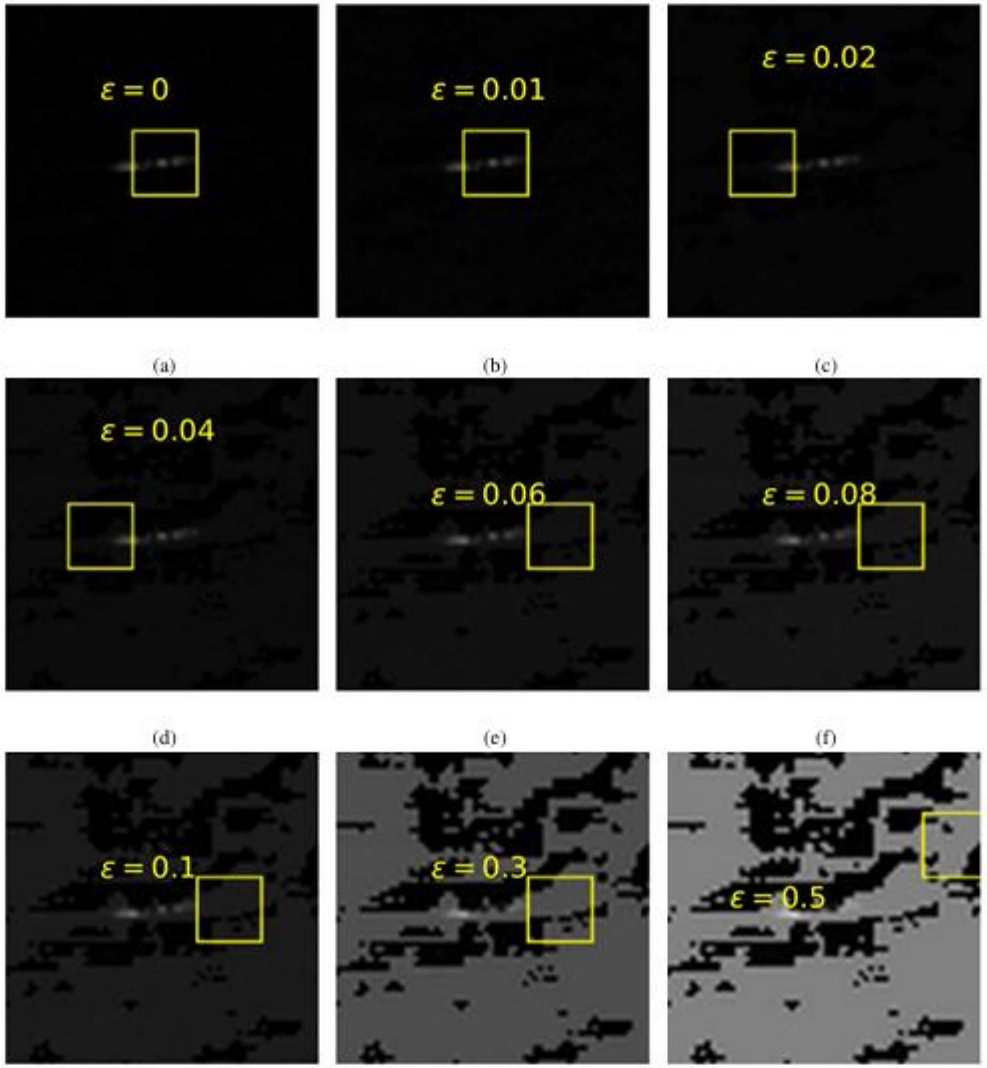
Real Data Analysis

➢ How ***overall accuracy decreases*** as the adversarial perturbation level ***ϵ increases***.



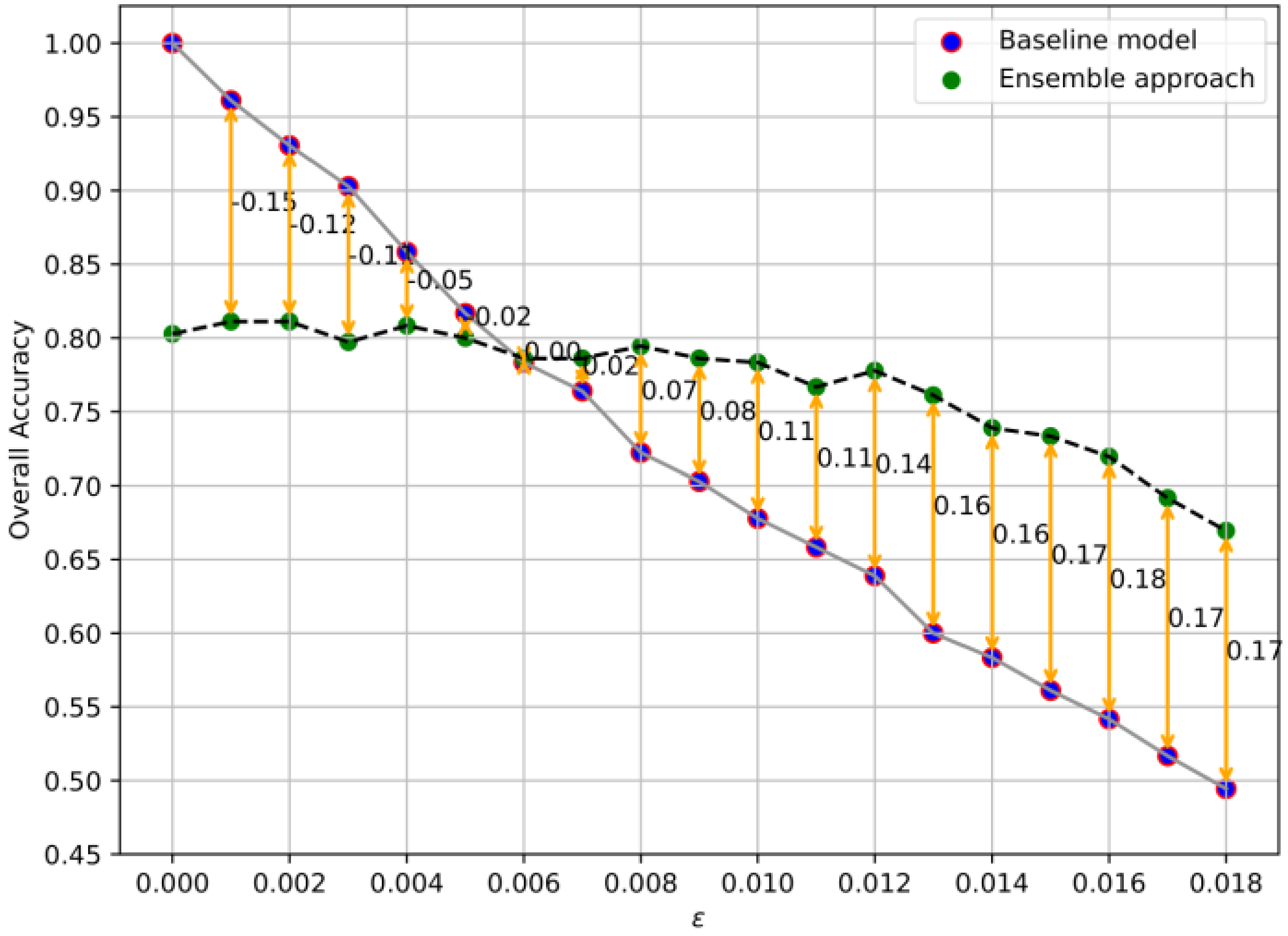
Real Data Analysis

- ***LIME's*** explanation for the most probable class, under ***different perturbation levels***
- The image belongs to ***class 2***
  - The image is ***correctly*** classified at ε = 0 and 0.01
  - Is misclassified as ***class 1*** at ε = 0.02 and 0.04
  - Is misclassified as ***class 0*** when ε = 0.06, 0.08, 0.1, 0.3, and 0.5.



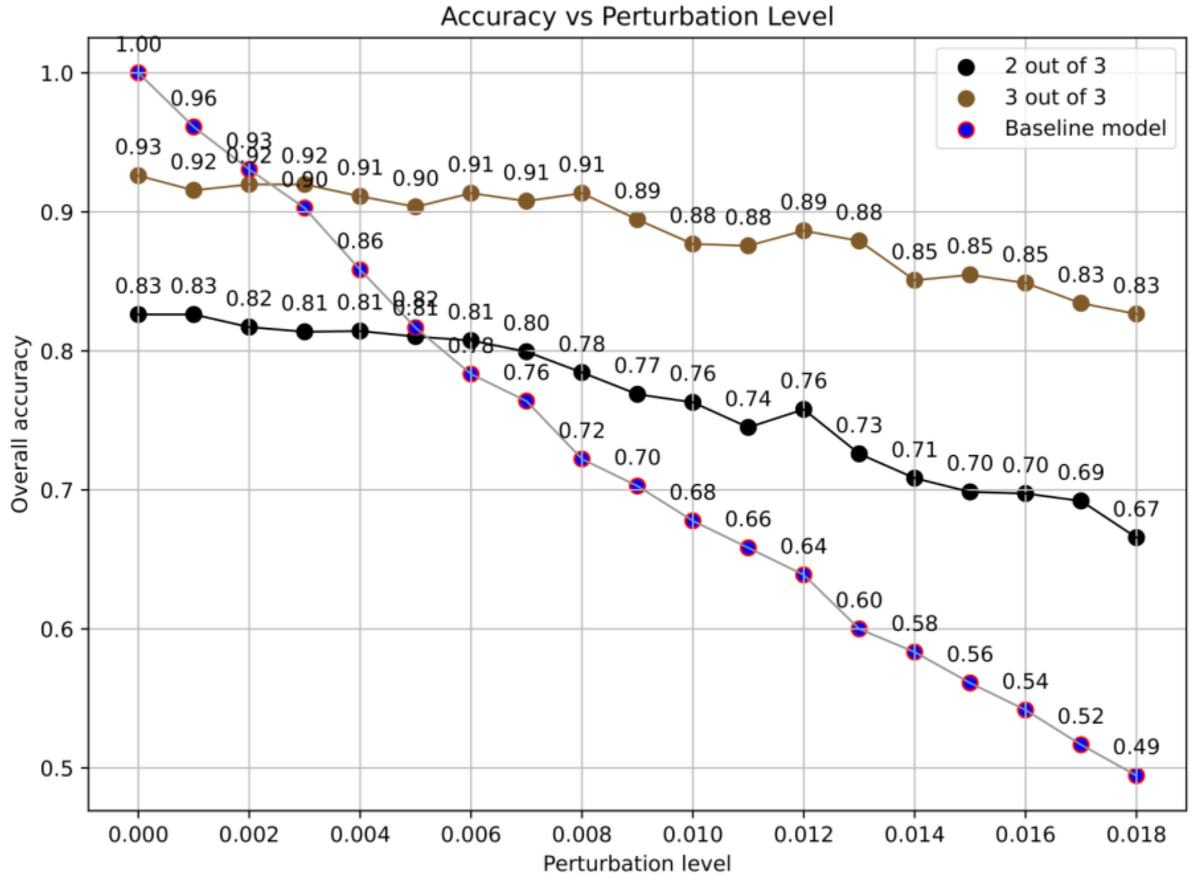
Real Data Analysis

- ***Transfer learning-***based ensemble approach:
  - VGG19, ResNet50,and MobileNet models
- ***Majority voting***: “***Ensemble approach***”
  - significantly ***outperforms*** the “***Baseline model***” as perturbation levels ***rise***



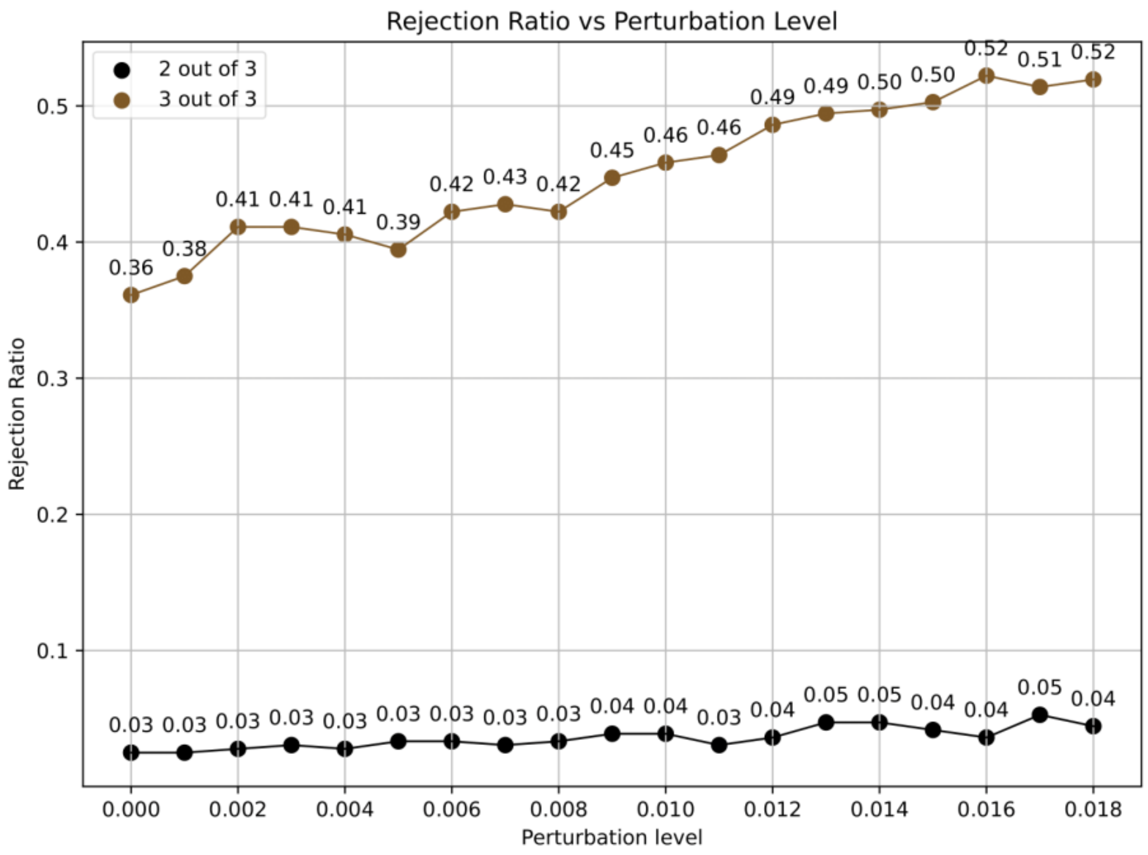
Real Data Analysis

- Another ***voting mechanism***: to ***assess*** the ***reliability*** of ***each prediction***
- This ***reliability measure*** can be interpreted
  - as a form of classifier with ***rejection***,
  - and it falls under the scope of ***open-set recognition*** (***OSR***) algorithms
- ***Accuracy*** after ***excluding unreliable predictions*** based on ***two voting criteria***:
- "***2 out of 3***": considers a prediction ***reliable*** if ***at least two predictions*** are the ***same***,
- "***3 out of 3***" requires ***all*** three predictions to ***match*** for reliability.



Real Data Analysis

- Higher accuracy ***doesn't*** always translate to ***better*** performance!
    - as it may result from ***rejecting*** a significant portion of the test set
- "***3 out of 3***" criterion:
    - gives the ***highest*** overall accuracy.
    - However, with the ***expense*** of ***rejecting*** a substantial portion of test images
    - which is ***not*** be ***acceptable*** in real-world scenarios.
- The rejection ratio of "***2 out of 3***"
    - falls between ***3*** to ***5*** percent
    - shows a better ***balance*** between ***accuracy*** and the ***proportion of rejected*** samples.



Conclusions

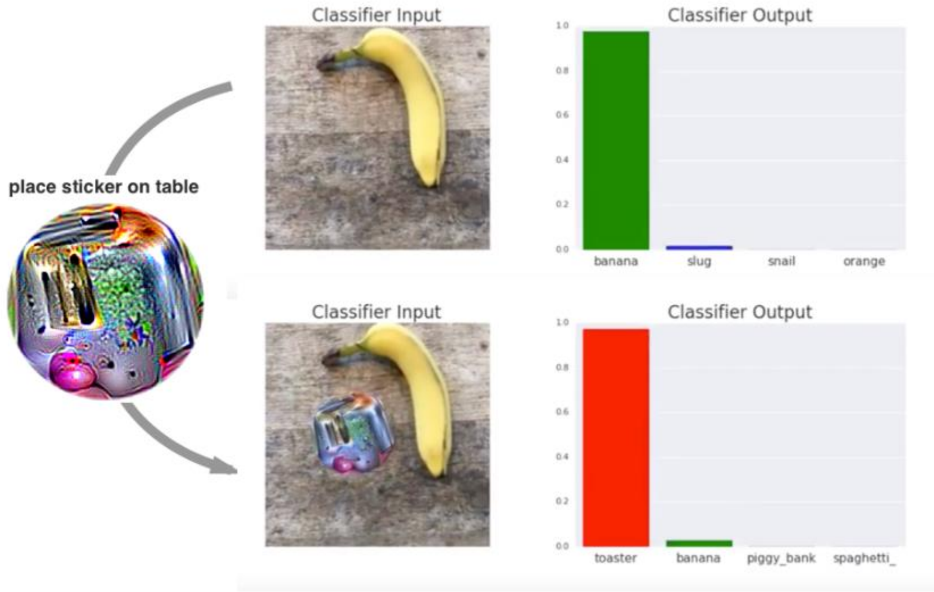
- Investigation of the ***vulnerability*** of ***SAR-based ship recognition models*** to ***adversarial attacks***
- Our analysis:
  - How ***adversarial*** perturbations ***degrade*** the CNN's classification performance
  - How ***LIME*** method can also be ***misleading***.
- The ***mitigating*** the impacts of ***adversarial attacks*** on such systems, especially in ***critical maritime surveillance*** applications, is necessary.
- The ***reliability*** of the explanations provided by LIME:
  - Depends on how much the input data is ***perturbed***.
- Explanations: can ***vary*** significantly under adversarial perturbations:
  - Inconsistent and possibly unreliable interpretations.
- An adversary, by strategically perturbing the input:
  - can ***manipulate LIME*** to emphasize features that are ***irrelevant*** or even ***incorrect***, in order to ***deceive*** the ***end user***.

Conclusions

- ***Over-reliance*** on ***LIME***, without a comprehensive understanding of its constraints:
  - can lead to a ***false sense of trust*** in the model's decisions.
- Since LIME ***approximates*** the complex decision boundary of a CNN with a ***simpler*** model, it is prone to producing ***inaccurate*** and ***oversimplified*** explanations.
- We proposed:
  - a ***transfer learning***-based ***ensemble learning*** strategy
  - to enhance the ***robustness*** of ship recognition models ***against adversarial*** examples.
- We analyzed:
  - the ***reliability*** of ***each prediction*** through a ***voting*** mechanism, along with an ***option to reject*** the ***unreliable*** predictions

Conclusions

- ***Question***:
1. One could ***realistically alter*** the images ***before*** model ***inference*** in a real attack,
2. What kind of ***access*** would be ***needed*** to the model.
- The neural network's input: likely ***well-protected*** and ***not exposed*** to adversaries.
- Makes ***direct manipulation*** of ***input data*** challenging.
- Nevertheless, there must be an ***interface*** for ***capturing*** and ***feeding*** data
- One ***potential*** approach:
  - to ***attach*** a ***small***, ***carefully designed patch*** or ***sticker*** to the ***target***.
- Exploiting the ***vulnerabilities*** of the ***imaging radar***
  - this ***patch*** could ***sufficiently alter*** the ***radar signature*** to ***deceive***



T. B. Brown, D. Mané, A. Roy, M. Abadi and J. Gilmer, "**Adversarial patch**" in arXiv:1712.09665, 2017.

Conclusions

- ***Future research directions*** include different ***defensive*** strategies:
  - ***adversarial training: re-training*** with perturbed images
  - ***ensemble learning*** with non-CNN models
- Adversarial perturbation can be applied to ***object detection*** task in maritime applications
- with SAR images.
- ***Incorporating LIME results*** in a feedback loop to help build a better classifier is crucial.

Thank you for your attention

Any questions?

cnit Italian national interuniversity consortium for telecommunications

National Lab RaSS Radar and Surveillance Systems

IEEE COMPUTER SOCIETY

IN SUPREMÆ DIGNITATIS 1343 UNIVERSITÀ DI PISA