Symposium On The Science Of Security (HoTSoS 2024)

Presenter: Muneeba Asif

Analytics for Cyber Defense (ACyD) Lab

Florida International University

Miami, FL, USA

Adversarial Data-Augmented Resilient Intrusion Detection System for Unmanned Aerial Vehicles

- **Published:** IEEE Big Data 2023 – Privacy and Security of Big Data (PSBD)
- **Authors:**

  - Muneeba Asif (ACyD Lab, Florida International University (FIU))
  - Mohammad Ashiqur Rahman (ACyD Lab, FIU)
  - Kemal Akkaya (ADWISE Lab, FIU)
  - Hossain Shahriar (Center for Cybersecurity, University of West Florida)
  - Alfredo Cuzzocrea (iDEA Lab, University of Calabria)

Presentation Outline

Research Objectives → Research Motivation → Attack Illustration

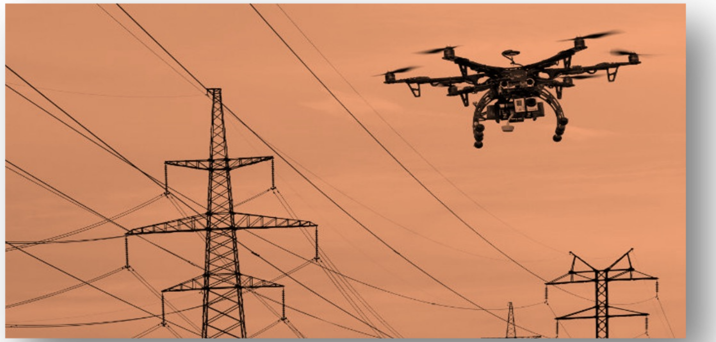
Problem Formulation → Proposed Solution → Case Study

Process Flow of Proposed Solution → Evaluation

UAVs and Their Applications

- Unmanned Aerial Vehicles (UAVs), aka drones, have multidisciplinary applications.
- Big tech companies are including and utilizing the many advantages UAVs bring with them.

Disaster Management



Electric Power Grid Inspections

Precision Agriculture

Defense and Military

Research Objectives and Statistics

RO-1: Comprehensive Security Analysis of UAV Systems in the Face of Adversarial Machine Learning Threats

❖ Conducting a thorough analysis and evaluation of UAV systems, focusing specifically on their robustness in security when confronted with sophisticated adversarial machine learning threats.

RO-2: Enhancing UAV IDSs' Resilience in Response to Adversarial Samples:

❖ Developing and implementing strategies to significantly enhance the resilience of IDSs for UAVs, explicitly identifying and mitigating the impact of meticulously crafted adversarial samples.

UAV industry projected to be $91.23 billion by 2030 ❶

Attacks can have severe consequences: mission thwarting, UAV intercepting/hijacking, etc. ❷

Owing to the nature of its applications, security, and mission precision are vital for UAVs ❸

IDS: intrusion detection system

❶ https://www.fortunebusinessinsights.com/industry-reports/unmanned-aerial-vehicle-uav-market-101603

❷ https://hackaday.com/2015/10/15/hijacking-quadcopters-with-a-mavlink-exploit/

❸ https://fieldlogix.com/news/gps-drone/

Motivation

❶

The New York Times

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk. (Published 2016)

The bot, @TayandYou, was put on hiatus after making offensive statements based on users' feedback, like disputing the existence of the...

→ The input data (tweets and interactions from users) was tainted with harmful content, leading the AI to produce undesired outputs.

❷

Softpedia News

Google reCAPTCHA Cracked in New Automated Attack

A trio of security researchers have devised a new automated attack that can break the CAPTCHA systems employed by Google and Facebook.

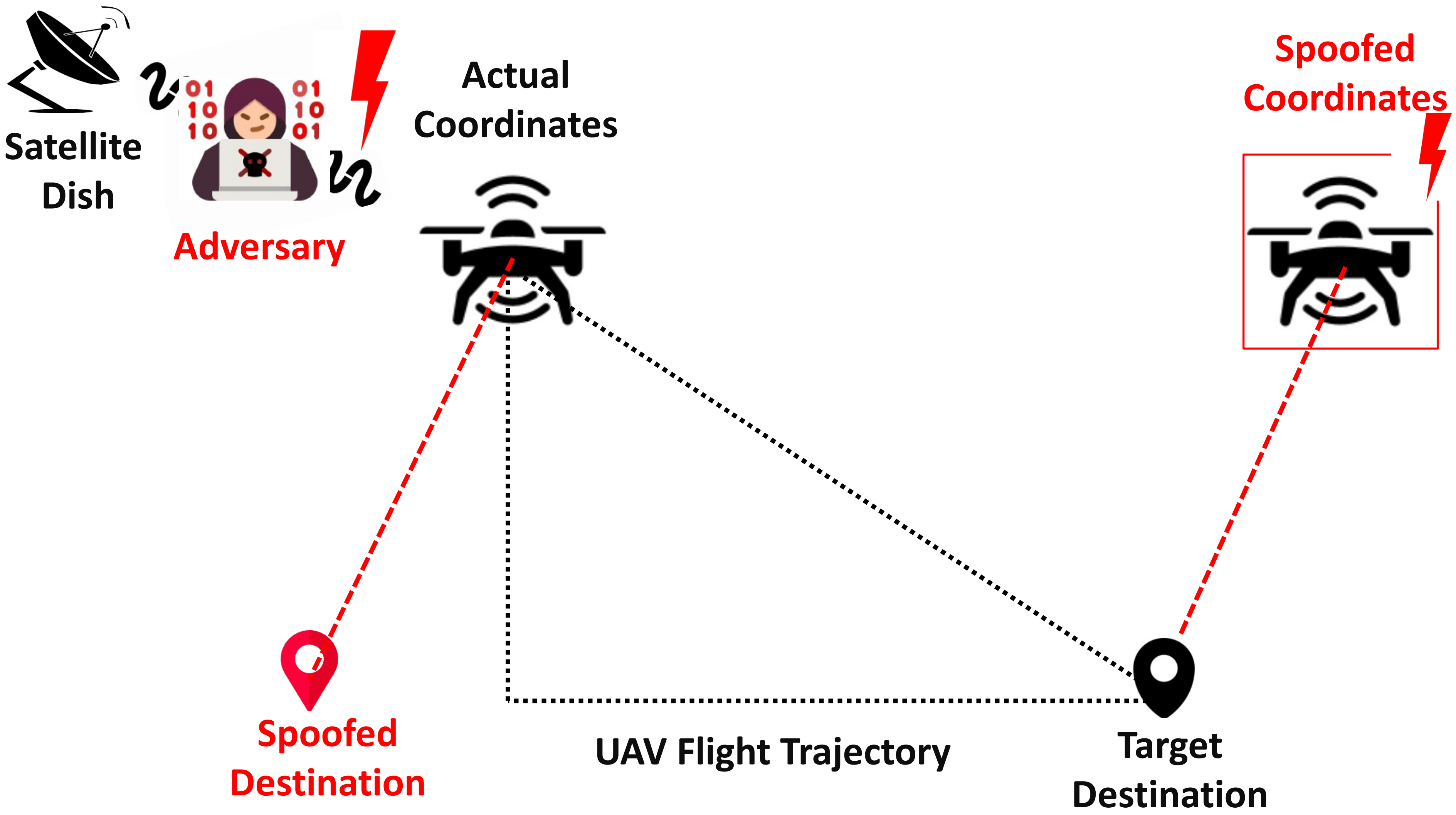
→ Researchers modified the audio CAPTCHAs slightly to mislead the speech-to-text API used for verification, achieving a high success rate in breaking the CAPTCHA.

Some Real-Life Attack Instances

❶ https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html

❷ https://news.softpedia.com/news/google-recaptcha-cracked-in-new-automated-attack-502677.shtml

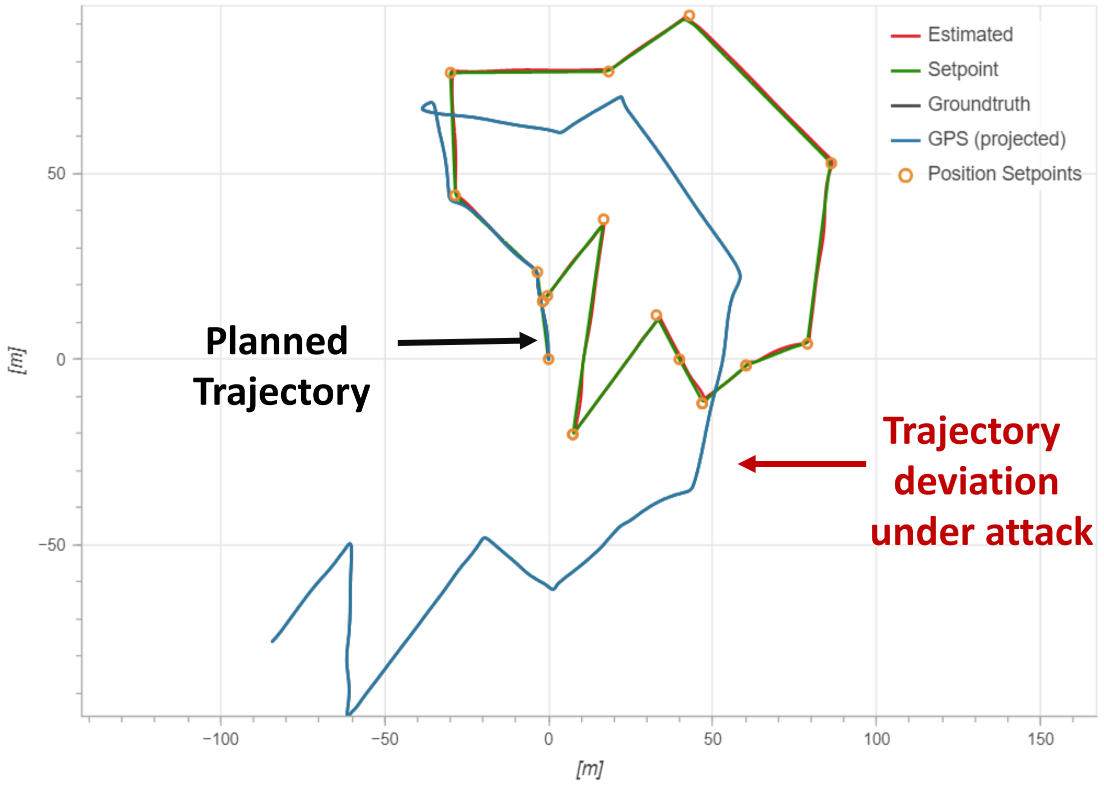
GPS Spoofing Attack Illustration



GPS Spoofing Attack Simulation

- Introduction
- Attack Illustration
  -- Concept
  --Simulation
- Problem Formulation
- Proposed Solution
- Case Study
- Methodology
- Evaluation
- Conclusion

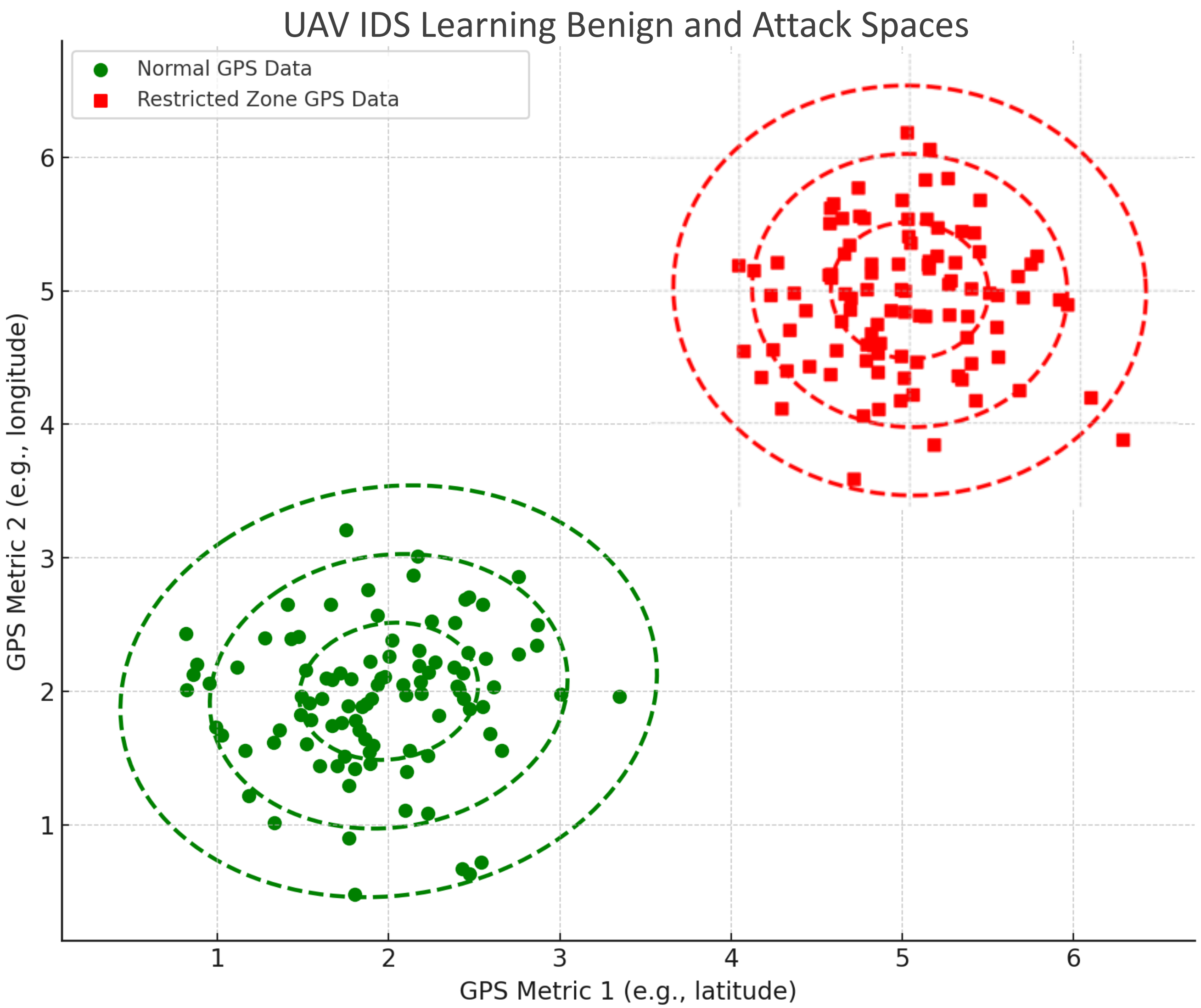
So, what is the existing detection mechanism for this?



GPS spoofing attack simulated using jMAVsim

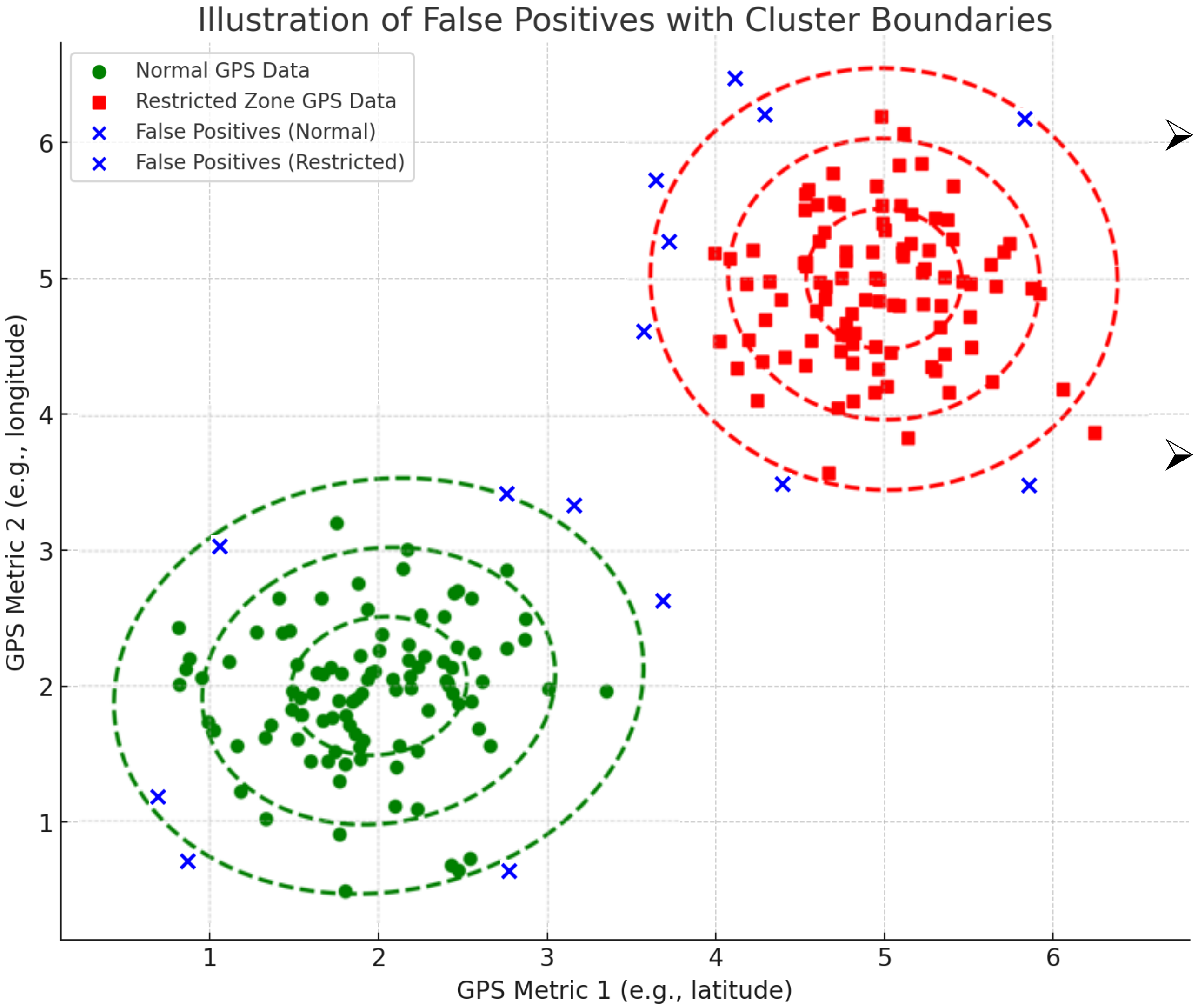
And what is the research gap, this paper seeks to address?

Problem Formulation (1/3)



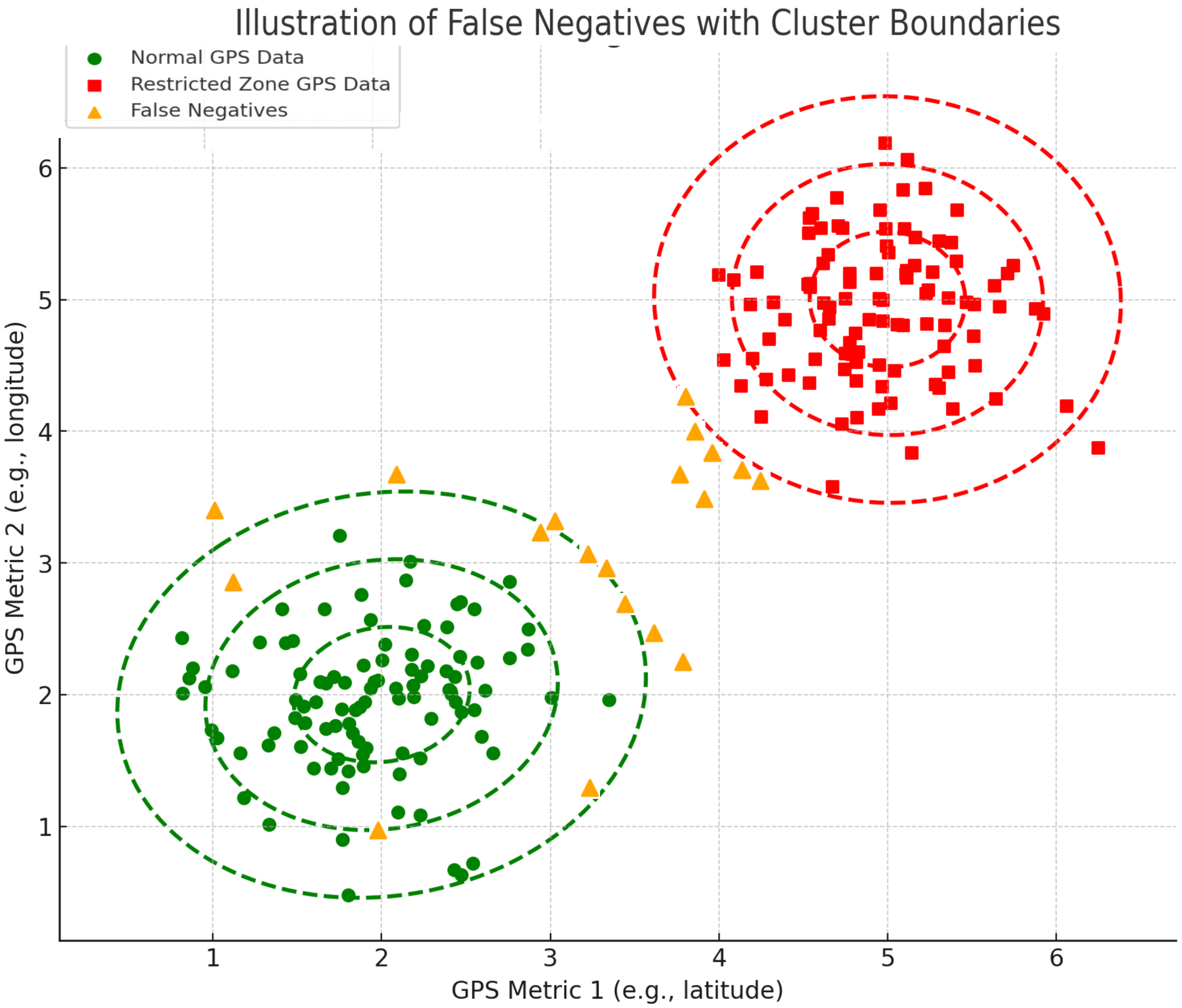
- ➢ Comprehensive UAV data collection is challenging
  - ❑ extensive flight time
  - ❑ operational coverage
  - ❑ variable environmental conditions
    - ❖ expensive
    - ❖ time consuming
- ➢ Collected data
  - ❑ sparse and limited
    - ❖ benign
    - ❖ malicious

Problem Formulation (2/3)



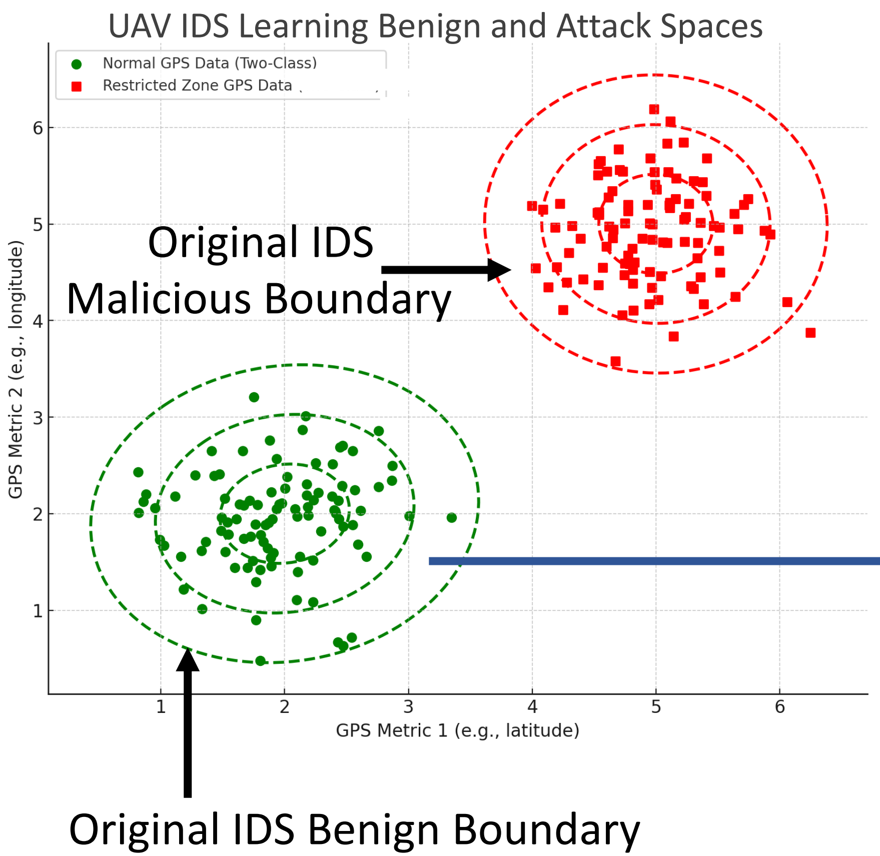
- IDS misclassifies data points
  - false positive
    - a benign GPS coordinate identified as an attacked sample
- Reason
  - less accurate boundaries due to data sparsity
    - false positive data points lie closer to or outside the boundaries

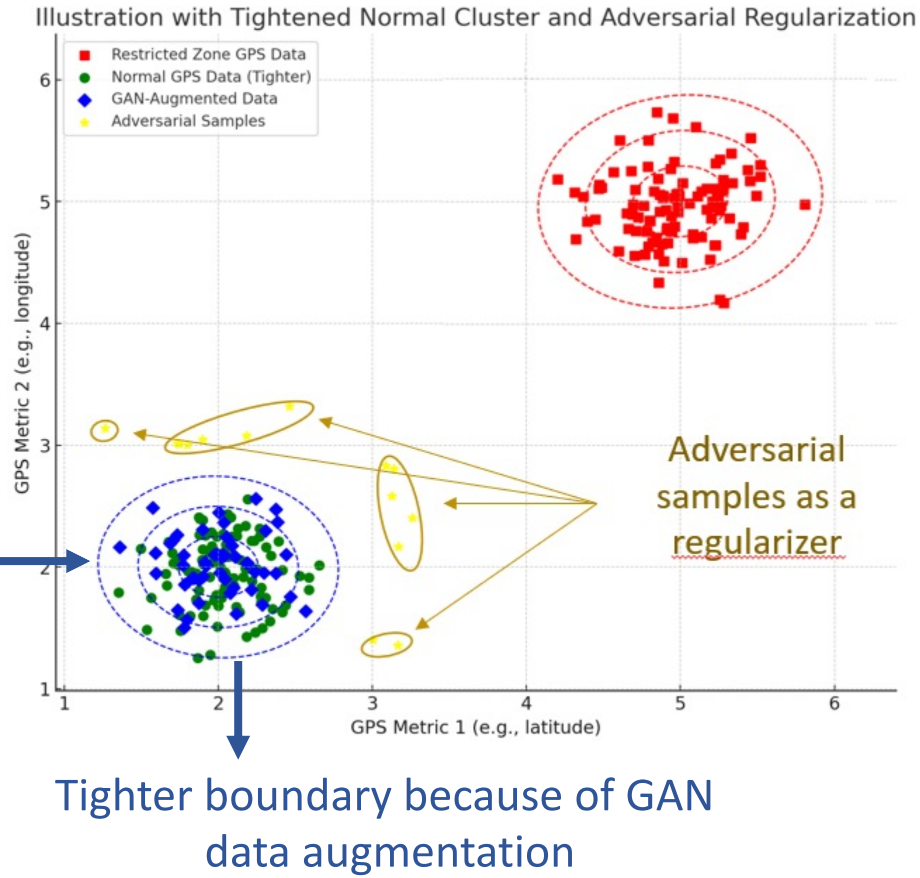
Problem Formulation (3/3)



- IDS misclassifies data points
  - false negative
    - a malicious GPS coordinate identified as a benign sample
- Reason
  - less accurate boundaries due to data sparsity
    - false negative data points lie closer to or outside the boundaries
- Adversarial exploitation
  - craft benign-looking adversarial samples
    - evade existing IDSs

Proposed Solution for Enhanced Resilience





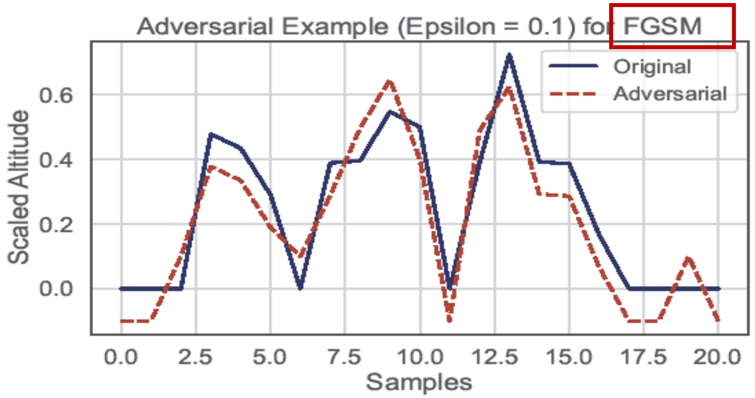
Case Study (Crafting Samples)

- Adversarial attacks subtly alter GPS signals, causing IDS to misidentify benign signals as threats or miss actual threats, risking UAV security.
- Perturbations (δ_{GPS}) are optimized to craft signals without exceeding the defined threshold (ϵ_{GPS}), maintaining stealth and causing misclassification.

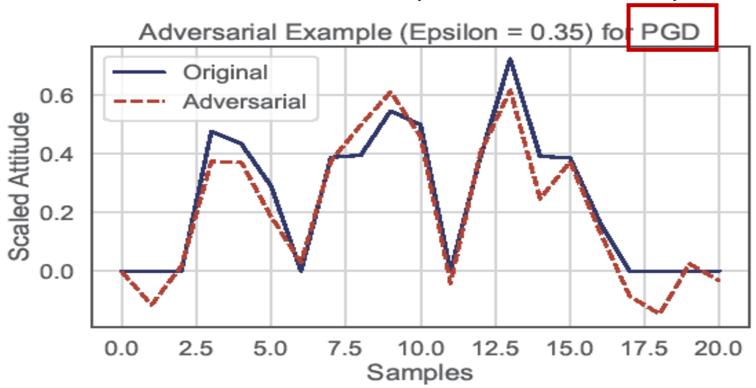
$$\min_{\delta_{GPS}} ||\delta_{GPS}|| \text{ subject to: } IDS(GPS_{orig} + \delta_{GPS}) \neq IDS(GPS_{orig})$$

$$\textbf{Constraint: } ||\delta_{GPS}|| \leq \epsilon_{GPS}$$

- The constraint $||\delta_{GPS}|| \leq \epsilon_{GPS}$ makes the adversarial perturbation go undetected by the IDS.



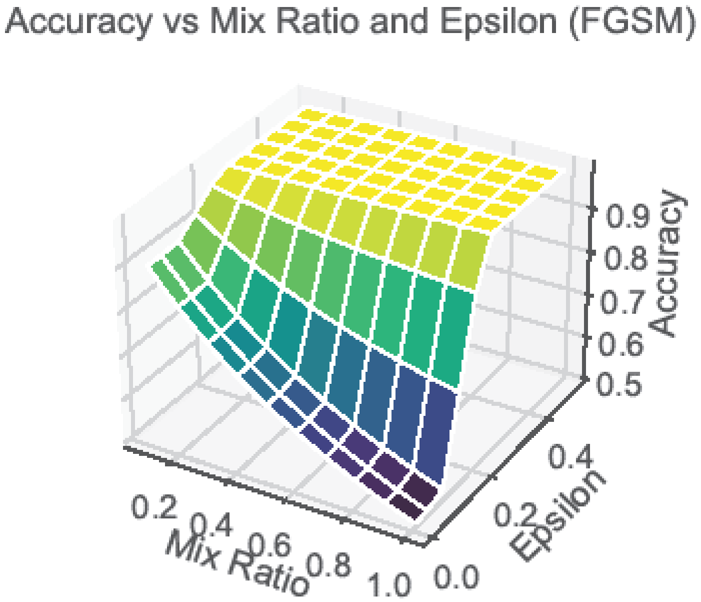
$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y_{\text{true}}))$$

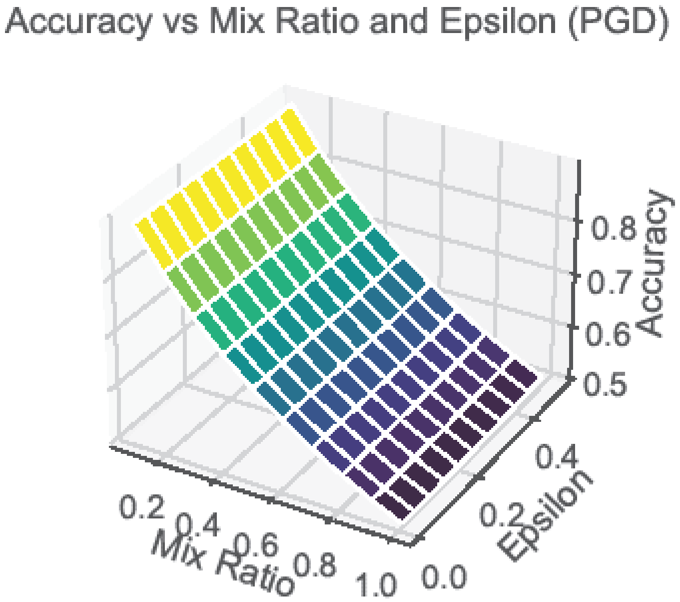


Projected Gradient Descent (PGD) refines perturbations iteratively within the allowed range for a stronger adversarial example.

Impact of Adversarial Samples

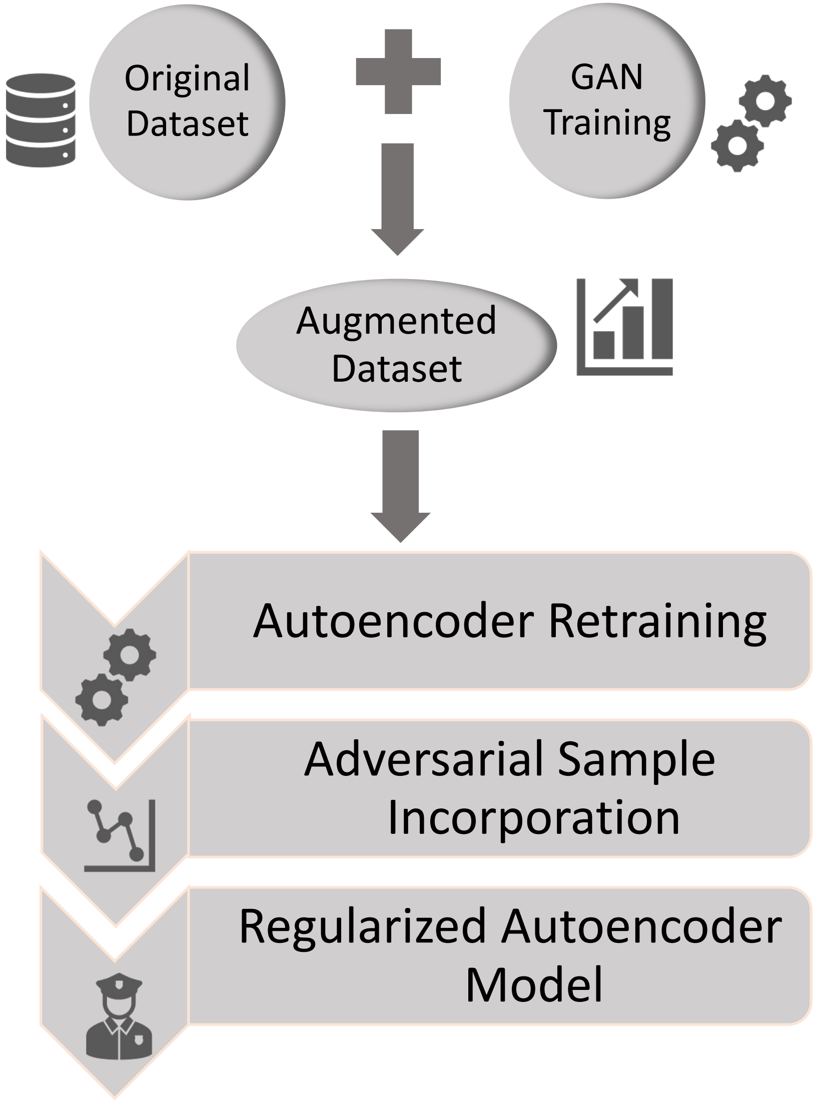
- The IDS is tested against both FGSM and PGD attacks
  - variable perturbation limit, ϵ.
  - variable mix ratios → ratios of benign and adversarial samples in the dataset
- PGD attacks trigger a sharper decline (i.e., over 50%) compared to FGSM (i.e., 30%)
- FGSM causes notable decrease after $\epsilon = 0.5$
- PGD degrades performance immediately and significantly with $\epsilon = 0.1$.





Proposed Framework

- **Data Augmentation with GANs:**
  - InfoGAN and WGAN leverage original data to produce synthetic samples, augmenting the training dataset for model resilience.
- **Autoencoder Retraining:**
  - The autoencoder is retrained with a mix of original and GAN-generated data to represent weak points in data distribution better.
- **Adversarial Regularization:**
  - Adversarial samples are used as regularizers in training, bolstering the autoencoder's robustness and anomaly detection.
- **Optimization and Performance Evaluation:**
  - The autoencoder's loss now includes a regularization term penalizing adversarial reconstruction. Hence, the model is less sensitive to input manipulations.



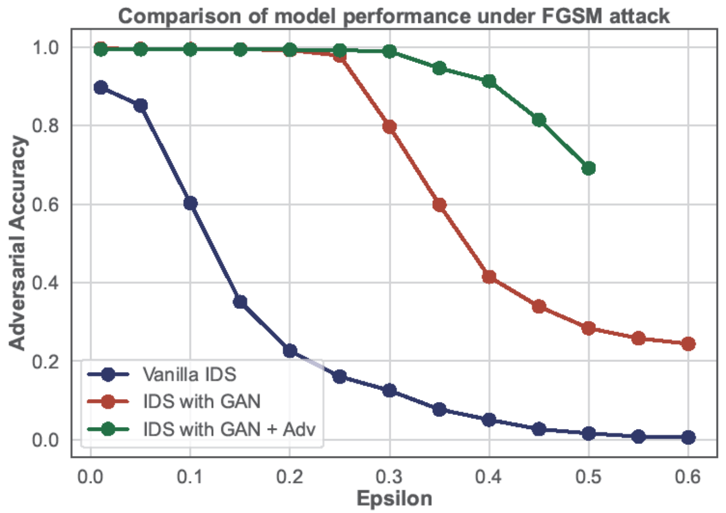
Evaluating Adversarial Attack Defense

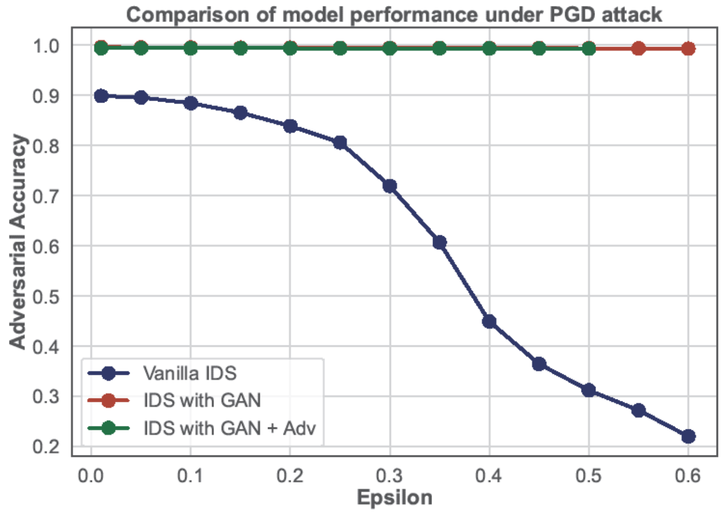
- Introduction
- Attack Illustration
- Problem Formulation
- Proposed Solution
- Case Study
- Methodology
- **Evaluation**
  --RQ1
  --RQ2
  --RQ3
  --RQ4
- Conclusion

➢ The initial IDS model's adversarial accuracy decreases with higher epsilon values
  ❖ drops to as low as 0.016042 for FGSM and 0.220658 for PGD attacks

➢ GAN-augmented IDS shows enhanced resiliency
  ❖ maintains over 0.99 accuracy for FGSM and PGD attacks at ϵ values up to 0.25
  ❖ a noted decrease at higher ϵ, particularly for FGSM

➢ The adversarially regularized IDS remains stable across varying epsilon values
  ❖ indicates superior resilience provided by adversarial learning





Evaluating Impact on Real GPS Attacks

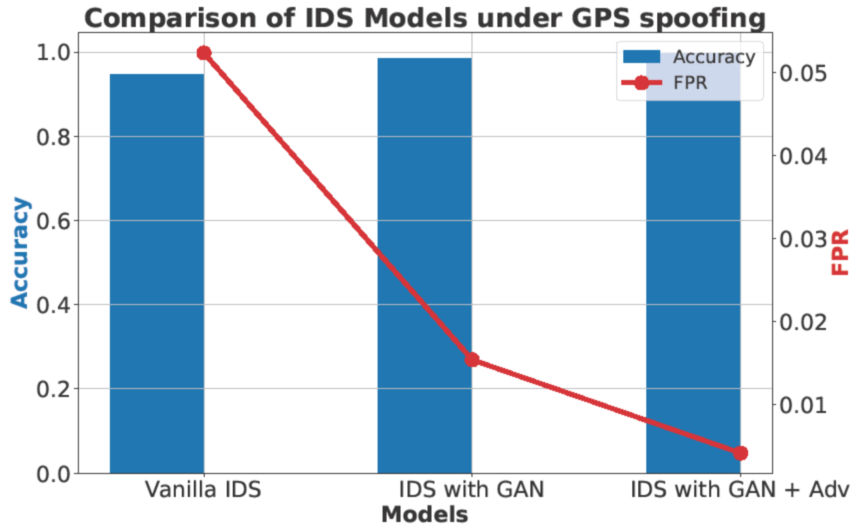
- Introduction
- Attack Illustration
- Problem Formulation
- Proposed Solution
- Case Study
- Methodology
- **Evaluation**
  --RQ1
  --RQ2
  --RQ3
  --RQ4
- Conclusion

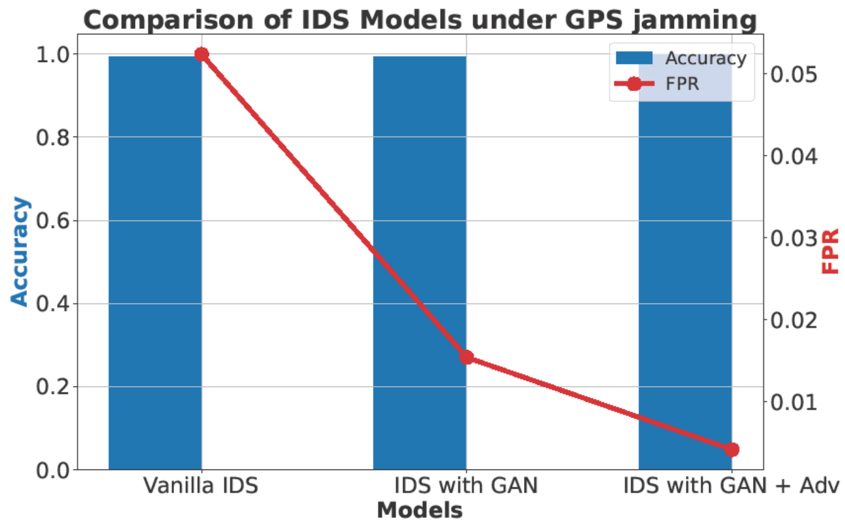
➢ For GPS spoofing attacks, the original IDS had an accuracy of 0.9476 and an FPR of 0.0523.

➢ GAN augmentation improved accuracy to 0.9845 and reduced FPR to 0.0154.

➢ The integration of GAN and adversarial learning in IDS yielded the highest accuracy (0.9957) and the lowest FPR (0.0042).

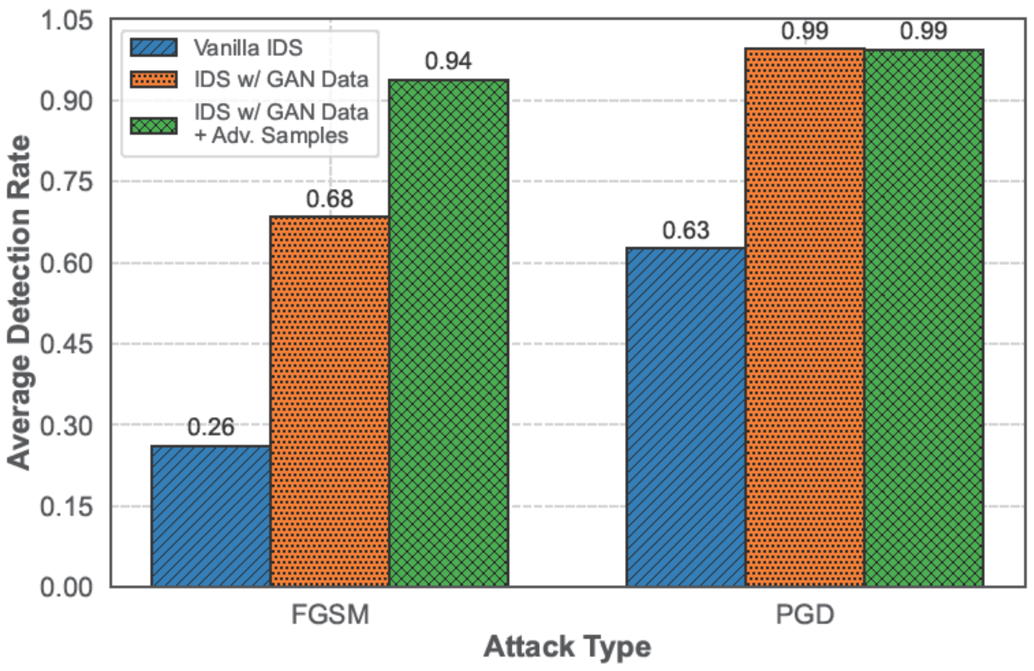
➢ In GPS jamming attacks, all models exhibited high accuracies (0.9942 to 0.9977) and low FPRs, with the GAN-augmented and adversarially trained model achieving the lowest FPR of 0.0023.





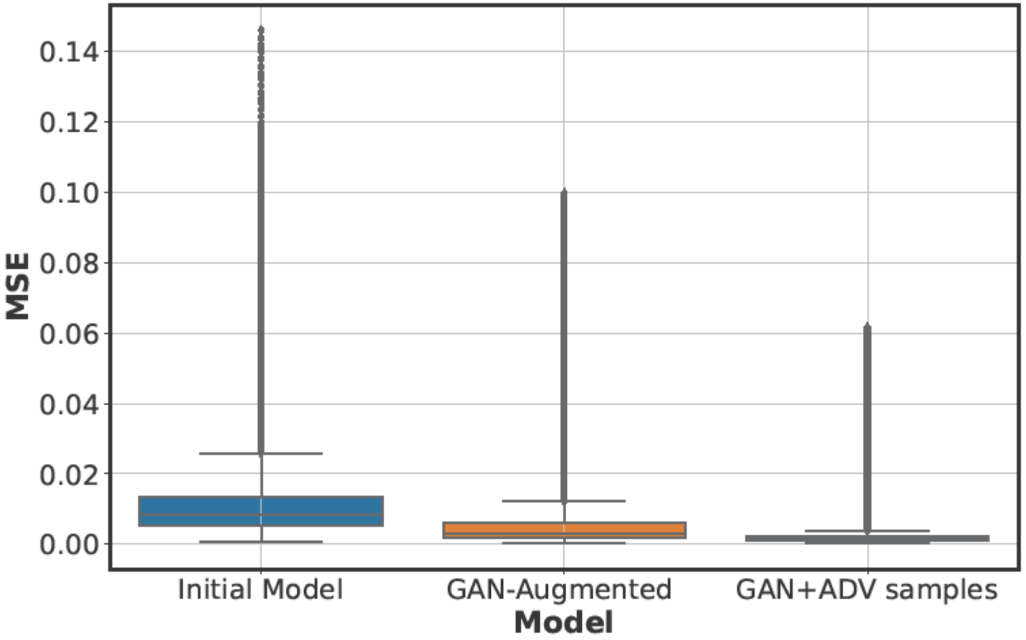
Evaluating Necessity of the Proposed Sol

- The original IDS displayed the lowest accuracies against FGSM and PGD attacks.
- GAN data augmentation alone significantly improved IDS accuracy, particularly for PGD attacks.
- Combining GAN data augmentation with adversarial learning resulted in further improvements, especially for FGSM attacks.



Evaluating IDS Model Performance

- The initial IDS model showed MSE values between 0.008927 and 0.007868, indicating a moderate fit to the data with room for improvement.
- GAN augmentation improved the model's performance, reducing MSE to a range between 0.006594 and 0.002299, suggesting a better fit to the data.
- Combining GAN data augmentation with adversarial samples yielded the lowest MSE values (approximately 0.002309 to 0.002104).



Conclusion

- In this work, we have highlighted the vulnerabilities of current IDS for UAVs against GPS spoofing and jamming attacks and proposed a framework using GANs and adversarial sample-based regularization.
- Under FGSM and PGD adversarial attacks, the detection rates for our improved IDS are 93.78% and 99.39%, respectively, outperforming the baseline rates of 26.14% and 62.6%.
- Additionally, our resilient IDS demonstrated an accuracy of 99.57% against GPS spoofing, substantially better than the conventional IDS accuracy of 94.76%.
- Importantly, the false positive rate was also reduced to 0.42% compared to the previous 5.23%.
- In future research, we will explore techniques like deep reinforcement learning, and study adaptability to other domains.