

Adversarial Robustness of xLSTM-Based ADS-B Intrusion Detection Against Progressive Spoofing and Discrete Jamming

Arouna BA,^{*,1} Serigne Mbacke GUEYE,² and Alassane DIOP¹

¹Ecole Doctorale Sciences et Technologies du Numerique (STN), Universite Numerique Cheikh Hamidou Kane (UNCHK), Dakar, Senegal

²Departement d'Informatique, CEGEP de Sept-Iles, Quebec, Canada

*Corresponding author: arouna.ba1@unchk.edu.sn

Abstract

The ADS-B protocol broadcasts aircraft position, velocity, and identity in plaintext at 1090 MHz with no authentication or integrity check, making it inherently vulnerable to false data injection. Machine learning-based intrusion detection systems (IDS) have shown strong performance on ADS-B data, with xLSTM-based models reaching F1-scores above 0.98, yet their robustness under adversarial perturbations remains largely untested, leaving open the question of whether an attacker aware of a deployed IDS could craft small perturbations to evade detection without altering the apparent legitimacy of the traffic.

We evaluate FGSM and PGD attacks across $\epsilon \in \{0.01, 0.03, 0.05, 0.10\}$ on three independent real-world ADS-B datasets from the OpenSky Network (45,000 messages each, 22–25 aircraft). We also introduce discrete jamming (progressive Gaussian noise injection across message fields) as a novel attack class whose baseline detectability (F1 = 0.984) masks a critical adversarial fragility. Under PGD at $\epsilon = 0.10$, heading and jamming classifiers collapse by 21.96 and 20.76 pp respectively. A hybrid adversarial training strategy (40% clean / 30% FGSM / 30% PGD, $\epsilon = 0.03$) recovers up to 6.9 pp, with post-defense F1-scores matching or exceeding original baselines.

Keywords: ADS-B; intrusion detection; xLSTM; adversarial robustness; adversarial training; progressive spoofing; discrete jamming; OpenSky Network

Abbreviations: ADS-B: Automatic Dependent Surveillance-Broadcast, IDS: Intrusion Detection System, FGSM: Fast Gradient Sign Method, PGD: Projected Gradient Descent, pp: percentage points

1. Motivation

The ADS-B protocol transmits aircraft state data as unencrypted 112-bit messages at 1090 MHz, roughly once per second [1]. Any receiver in range can capture these messages; any transmitter can forge them. The resulting attack surface (injection, modification, deletion, jamming) has been demonstrated using off-the-shelf SDR equipment [2].

We note an important distinction: the anomalies targeted by a security-oriented IDS are *deliberate* (an adversary actively crafting messages to deceive a surveillance system), as opposed to the non-malicious anomalies studied in data quality work (transponder miscalibration, multipath, sensor dropout), which follow different statistical patterns as documented in the OpenSky integrity report [3].

The state-of-the-art IDS for ADS-B is the xLSTM-based model of Ngamboe et al. [4], which achieves $F1 > 0.99$ on altitude and groundspeed attacks, the highest reported on this benchmark. Yet it was evaluated exclusively on clean data, and discrete jamming was dismissed as “typically easy to detect” without empirical evidence. This paper addresses both gaps: no prior work has (1) tested adversarial robustness of an xLSTM-based ADS-B IDS, (2) empirically evaluated discrete jamming as an adver-

serially fragile threat, or (3) validated a FGSM+PGD adversarial training defense on real OpenSky data.

2. Related Work

ML-based ADS-B anomaly detection has a well-established lineage. Habler and Shabtai [5] proposed an LSTM autoencoder using reconstruction error. Chevrot et al. [6] introduced the Contextual Autoencoder (CAE) at F1 = 0.939 on velocity drift. Luo et al. [7] reached F1 = 0.94 with TCN-Transformer-SVDD. The framework for realistic synthetic ADS-B anomalies developed by Chevrot, Cretin et al. [8] established that evaluation data quality is as critical as model design. Ngamboe et al. [4] set the current state of the art with an xLSTM-based IDS, yet tested only under clean conditions.

On the adversarial side, Goodfellow et al. [9] introduced FGSM; Madry et al. [10] extended it to iterative PGD. Sun et al. [11] showed adversarial training substantially recovers performance on RF emitter identification. Shang et al. [12] demonstrated that gradient-based perturbations fool deep learning spoofing detectors. Physical-layer methods (ToA, TDoA, AoA, RSSI) offer complementary guarantees but require multi-receiver infrastructure not universally available in OpenSky; our work on single-receiver state vectors is therefore complementary to, not competing with, such approaches. Table 1 positions our work in this landscape.

Table 1. Positioning of this work relative to the literature.

Reference	xLSTM	Adv. eval.	Discr. jam.	Real data	Defense
Habler [5]	–	–	–	✓	–
Chevrot [6]	–	–	–	✓	–
Luo [7]	–	–	–	✓	–
Ngamboe [4]	✓	–	–	✓	–
Shang [12]	–	Atk	–	–	–
Sun [11]	–	✓	–	–	✓
This work	✓	✓	✓	✓	✓

3. Methodology

3.1 Datasets and isolation protocol

Three independent ADS-B datasets were collected from the OpenSky Network [1] via the Trino SQL interface. **Dataset A** (2023-06-01) was used exclusively for unsupervised pre-training, never exposed to labels or attack signals. **Dataset B** (2023-06-15) served for fine-tuning and adversarial example generation. **Dataset C** (2023-07-01) was held out entirely for evaluation. Each covers 08:00–10:00 UTC with approximately 45,000 messages from 22–25 aircraft. This strict three-way split ensures no data leakage across pre-training, fine-tuning, adversarial training, and evaluation. Six state features (latitude, longitude, groundspeed, heading, vertical rate, barometric altitude) are complemented by four first-order temporal differences ($d = 10$ input dimensions). A RobustScaler fitted on Dataset A normalizes all features; sequences of length 50 are extracted with stride 10.

3.2 Attack classes

Four binary classifiers use a one-versus-rest (OvR) strategy. Three follow the benchmark of [4]: **ALT** (+82 ft/msg), **GS** (+1.9 kts/msg), and **HDG** (+1°/msg, modulo 360°). The ALT classifier is symmetric by construction, so the +82 ft parameterization is not a limitation. The fourth class, **JAMMING** (our contribution), injects additive Gaussian noise $\mathcal{N}(0, \sigma_t^2)$ across all message fields, with standard deviation increasing linearly:

$$\sigma_t = \sigma_{\min} + \frac{t}{N-1}(\sigma_{\max} - \sigma_{\min}), \quad t = 0, \dots, N-1 \quad (1)$$

with $\sigma_{\min} = 0.001$, $\sigma_{\max} = 0.05$. Discrete jamming spreads low-amplitude noise across all features progressively, retaining the statistical shape of legitimate ADS-B traffic while gradually corrupting it: a stealthy profile that achieves F1 = 0.984 under normal detection, contradicting the dismissal of [4].

3.3 Model and training protocol

The xLSTM architecture [13], implemented following [4], uses four mLSTM blocks (hidden dim 64). Pre-training on Dataset A uses MSE loss for 20 epochs ($lr = 8 \times 10^{-4}$, batch 32). Each classifier is fine-tuned on Dataset B for 15 epochs with binary cross-entropy (early stopping, patience 5 on validation F1). FGSM and PGD attacks use $\varepsilon \in \{0.01, 0.03, 0.05, 0.10\}$; PGD uses $\alpha = 0.25\varepsilon$, $T = 10$ iterations. The hybrid adversarial defense fine-tunes for 10 additional epochs on a 40/30/30 mix of clean, FGSM, and PGD examples at $\varepsilon = 0.03$ (AdamW, $lr = 10^{-4}$, gradient clipping at 1.0). All evaluation is on Dataset C.

4. Results and Discussion

Table 2 summarizes baseline performance, PGD degradation, and post-defense recovery.

Table 2. Baseline F1, PGD degradation, and post-defense recovery (held-out Dataset C).

Classifier	Baseline	PGD $\varepsilon=0.05$	PGD $\varepsilon=0.10$	After defense
ALT	0.9953	−0.35 pp	−0.82 pp	0.9977 (+0.5 pp)
GS	0.9977	−0.12 pp	−0.12 pp	0.9977 (+0.1 pp)
HDG	0.9583	−8.37 pp	−21.96 pp	0.9861 (+6.9 pp)
JAMMING	0.9838	−7.37 pp	−20.76 pp	0.9741 (+2.7 pp)

Differential vulnerability. ALT and GS are essentially unaffected across all ε : their large, monotonic deviations sit far from the decision boundary. HDG and JAMMING collapse at $\varepsilon = 0.10$ to F1 = 0.7387 and 0.7762 respectively: heading drift is slow (1°/msg) and discrete jamming spreads low-amplitude noise across all features. PGD degrades HDG by 21.96 pp versus 14.58 pp for FGSM, confirming iterative attacks pose a stronger threat than single-step ones (Fig. 1a).

Defense effectiveness. Hybrid adversarial training restores robustness across all four classifiers simultaneously; HDG's post-defense F1 (0.9861) even exceeds its baseline (0.9583), consistent with adversarial examples acting as regularizers (Fig. 1b). One caveat from peer review: even at F1 = 0.99, a system monitoring 10,000–15,000 aircraft generates dozens of false alarms per cycle, so reducing false positive rates to operational levels remains necessary future work.

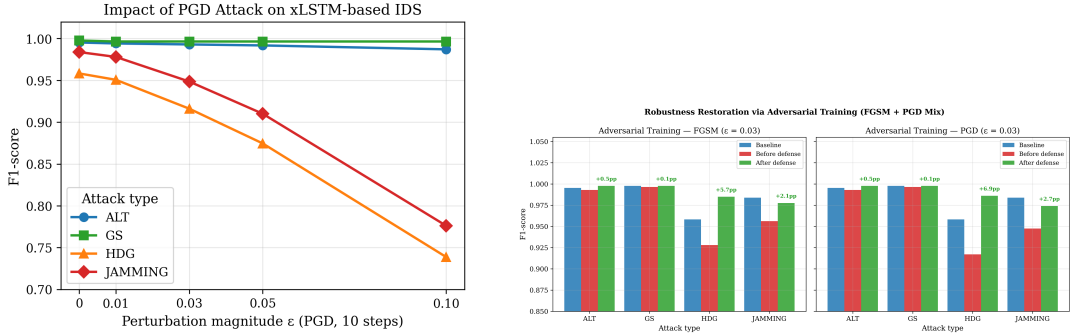


Figure 1. (a) F1-score vs. ϵ under PGD (10 steps, $\alpha = 0.25\epsilon$). (b) F1-score before and after hybrid adversarial training (40/30/30, $\epsilon = 0.03$) under FGSM and PGD, for all four classifiers.

5. Conclusion

Our experiments show that xLSTM-based ADS-B IDS do not exhibit uniform robustness to adversarial perturbations. ALT and GS classifiers remain stable across all tested ϵ values, whereas HDG and JAMMING degrade by more than 20 pp under PGD at $\epsilon = 0.10$. Discrete jamming, previously dismissed in the literature without empirical support, reaches F1 = 0.984 under standard conditions yet proves adversarially fragile. Hybrid adversarial training (40/30/30 clean/FGSM/PGD) provides a single unified defense, recovering up to 6.9 pp for HDG and 2.7 pp for JAMMING. These results indicate that high baseline F1 is not a reliable indicator of adversarial robustness. Iterative attacks such as PGD pose a stronger threat than single-step methods, and adversarial training on real OpenSky data is both feasible and effective.

Acknowledgement

The authors thank the OpenSky Network for providing open access to the ADS-B surveillance data used in this study.

Author contributions

- Arouna BA: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing (Original Draft)
- Serigne Mbacke GUEYE: Conceptualization, Supervision, Writing (Review and Editing)
- Alassane DIOP: Supervision, Project Administration, Writing (Review and Editing)

Open data statement

All ADS-B data are publicly available through the OpenSky Network Trino SQL interface at <https://opensky-network.org/data/trino> (table: state_vectors_data4). Datasets collected on 2023-06-01, 2023-06-15, and 2023-07-01 (08:00–10:00 UTC). No proprietary or personal data were used.

Reproducibility statement

The full pipeline (data collection, preprocessing, xLSTM pre-training and fine-tuning, FGSM/PGD generation, and adversarial training) is implemented in Python/PyTorch on standard CPU hardware.

All hyperparameters are reported in Section 3. Code and model weights will be made publicly available upon acceptance.

References

- [1] M. Schäfer, M. Strohmeier, V. Lenders, I. Martinovic, and M. Wilhelm. “Bringing up OpenSky: A Large-Scale ADS-B Sensor Network for Research”. In: *Proc. 13th Int. Symp. Information Processing in Sensor Networks (IPSN)*. Berlin, Germany, Apr. 2014, pp. 83–94.
- [2] M. Riahi Manesh and N. Kaabouch. “Analysis of Vulnerabilities, Attacks, Countermeasures and Overall Risk of the ADS-B System”. In: *Int. J. Crit. Infrastruct. Protection* 19 (Dec. 2017), pp. 16–31.
- [3] X. Olive, J. Krummen, B. Figuet, and R. Alligier. “Filtering Techniques for ADS-B Trajectory Preprocessing”. In: *J. Open Aviation Sci.* 2 (2024). DOI: [10.59490/joas.2024.7882](https://doi.org/10.59490/joas.2024.7882).
- [4] M. Ngamboé, J.-S. Marrocco, J.-Y. Ouattara, J. M. Fernandez, and G. Nicolescu. “New Machine Learning Approaches for Intrusion Detection in ADS-B”. In: *Proc. 44th AIAA DATC/IEEE Digital Avionics Systems Conf. (DASC)*. Montreal, QC, Canada, 2025, pp. 1–10.
- [5] E. Habler and A. Shabtai. “Using LSTM Encoder-Decoder Algorithm for Detecting Anomalous ADS-B Messages”. In: *Comput. Security* 78 (Sept. 2018), pp. 155–173.
- [6] A. Chevrot, A. Vernotte, and B. Legeard. “CAE: Contextual Autoencoder for Multivariate Time-Series Anomaly Detection in Air Transportation”. In: *Comput. Security* 116 (2022), p. 102652.
- [7] P. Luo, B. Wang, and J. Tian. “TTSAD: TCN-Transformer-SVDD Model for Anomaly Detection in Air Traffic ADS-B Data”. In: *Comput. Security* 141 (2024), p. 103840.
- [8] A. Chevrot, A. Vernotte, P. Bernabe, A. Cretin, F. Peureux, and B. Legeard. “Improved Testing of AI-Based Anomaly Detection Systems Using Synthetic Surveillance Data”. In: *Proc. 8th OpenSky Symposium*. Vol. 59. 1. 2020, p. 9. DOI: [10.3390/proceedings2020059009](https://doi.org/10.3390/proceedings2020059009).
- [9] I. J. Goodfellow, J. Shlens, and C. Szegedy. “Explaining and Harnessing Adversarial Examples”. In: *Proc. Int. Conf. Learning Representations (ICLR)*. San Diego, CA, USA, May 2015.
- [10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. “Towards Deep Learning Models Resistant to Adversarial Attacks”. In: *Proc. Int. Conf. Learning Representations (ICLR)*. Vancouver, BC, Canada, May 2018.
- [11] L. Sun, D. Ke, X. Wang, Z. Huang, and K. Huang. “Robustness of Deep Learning-Based Specific Emitter Identification under Adversarial Attacks”. In: *Remote Sens.* 14.19 (Oct. 2022), p. 4996.
- [12] F. Shang, B. Wang, T. Li, J. Tian, K. Cao, and R. Guo. “Adversarial Examples on Deep-Learning-Based ADS-B Spoofing Detection”. In: *IEEE Wireless Commun. Lett.* 9.10 (Oct. 2020), pp. 1734–1737.
- [13] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter. “xLSTM: Extended Long Short-Term Memory”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 37. 2024, pp. 107547–107603.