Adversarial Resilience and Privacy in AI-Driven Security Systems

University of Lagos, Akoka, Lagos , Nigeria

Faculty of Engineering

International Conference, 2026

Presenter: Latifat Oderinu

Why this research matters

AI is increasingly used to defend industrial and economic infrastructure.

AI is now part of the defence

AI supports intrusion detection, malware classification and security monitoring.

But AI is an attack surface

Evasion, poisoning and compromised federated participants can manipulate model behaviour.

Trust is a deployment condition

Privacy safeguards and explanations affect whether organisations and analysts can use AI safely.

Central concern

A security system that can be manipulated may create a false sense of protection—leaving critical operations exposed while appearing secure.

This review treats robustness, privacy and explainability as connected engineering requirements.

Research focus and questions

Three dimensions frame the review.

RQ1 | Adversarial resilience

What are the principal attack vectors targeting AI security models, and how effective are current defences?

RQ2 | Privacy-preserving collaboration

How do federated learning and differential privacy affect model performance and deployment?

RQ3 | Explainability and trust

To what extent do explanation and trust mechanisms influence operational adoption?

Contribution: a cross-cutting synthesis rather than treating attacks, privacy and explainability as separate research tracks.

Review methodology

PRISMA 2020-guided systematic review; search window: 1 January 2018–1 October 2025.

6,767 initial records

4,426 unique records

4,513 after snowballing

819 full-text screened

358 studies included

Sources searched

IEEE Xplore • Scopus • Web of Science • ACM Digital Library • arXiv (cs.CR, cs.LG, cs.AI).

Searches run 12 October 2025;

snowballing added 87 records.

Quality screening

Seven criteria: objectives, method, dataset, metrics, baselines, reproducibility and limitations. Studies scoring below 3.5/7 were excluded; included studies averaged 5.3 (SD 0.94).

Of the 358 included studies, 121 are analysed here: 62 adversarial-ML, 41 federated-learning and 18 explainability. The other 237 fall outside the three research questions.

Where AI security models are vulnerable

Threats span the complete machine-learning lifecycle.

1 | Training

DATA POISONING

Crafted samples corrupt learned decision boundaries.

BACKDOORS

Trigger patterns cause attacker-selected predictions.

2 | Inference

EVASION

Modified inputs cause malicious activity to be misclassified.

Threat models range from white-box access to black-box queries.

3 | Deployment & outputs

MODEL INVERSION / MEMBERSHIP INFERENCE

Outputs can expose sensitive training information.

FEDERATED / BYZANTINE ATTACKS

Participants can submit malicious model updates.

Key point: protecting the data, model, update pipeline and prediction interface requires different controls.

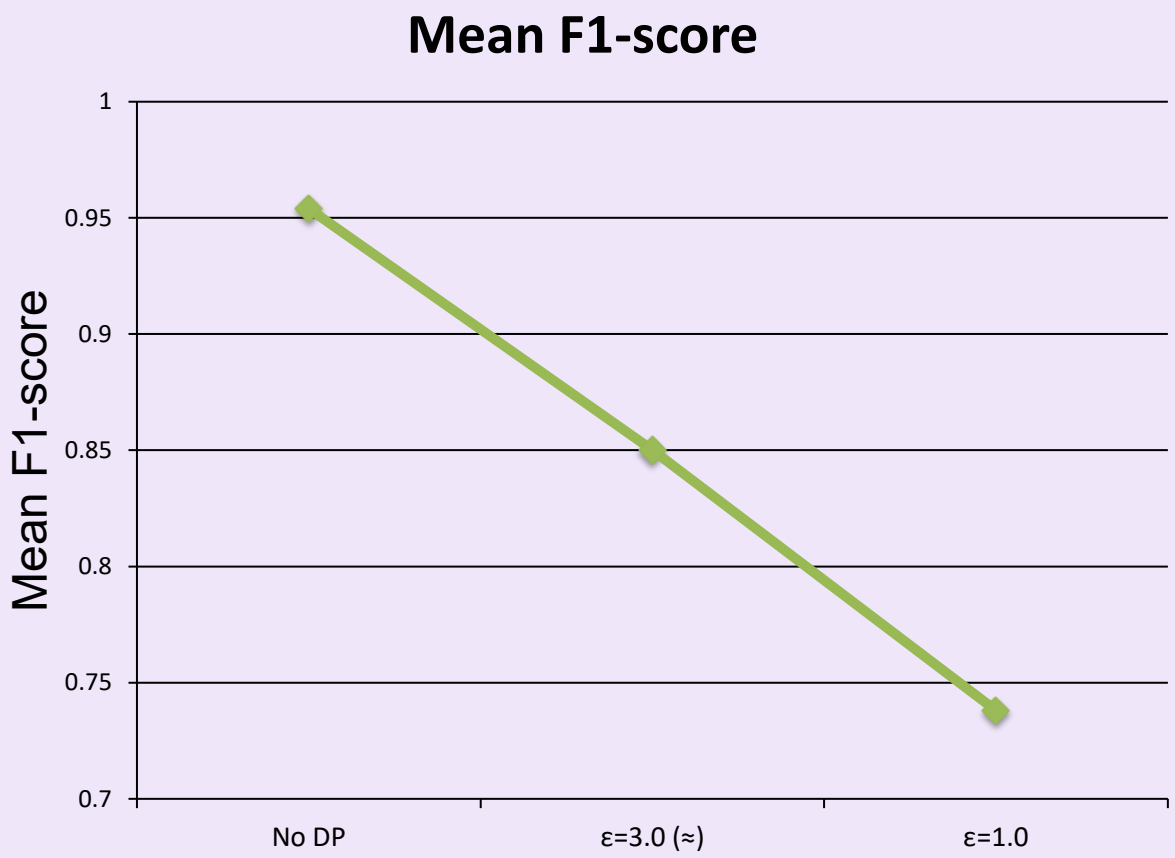
RQ1 findings | Attack impact and defence gaps

ATTACK	REPORTED IMPACT	DEFENCE / LIMITATION
White-box evasion	Critical: near-complete misclassification of image classifiers	Adversarial training improves robustness to trained-for attacks; costs clean accuracy
Black-box evasion	High: depends on the attacker's query budget	Ensemble diversity is a partial defence; query-limited
Data poisoning	High: global accuracy loss; size depends on poisoned fraction	Outlier detection is moderate; weaker against clean-label attacks
Backdoor insertion	Critical: attacker-chosen output on triggered inputs	Neural Cleanse / spectral analysis are partial; adaptive backdoors evade
Byzantine FL	High: degrades the global model despite robust aggregation	Krum, Bulyan: partial defences with no provable guarantees

Adversarial training is necessary but incomplete: it costs clean accuracy and leaves residual risk.

RQ2 findings | Privacy has a measurable utility cost

Federated learning shares model updates rather than raw telemetry—but updates can still leak information or be manipulated.



Moderate privacy (ε = 3.0)

F1 falls from 0.954 to about 0.85 (≈ −10.4 pp).

Strong privacy (ε = 1.0)

F1 drops to 0.738 (−21.6 pp) in this intrusion-detection study.

Calibration matters

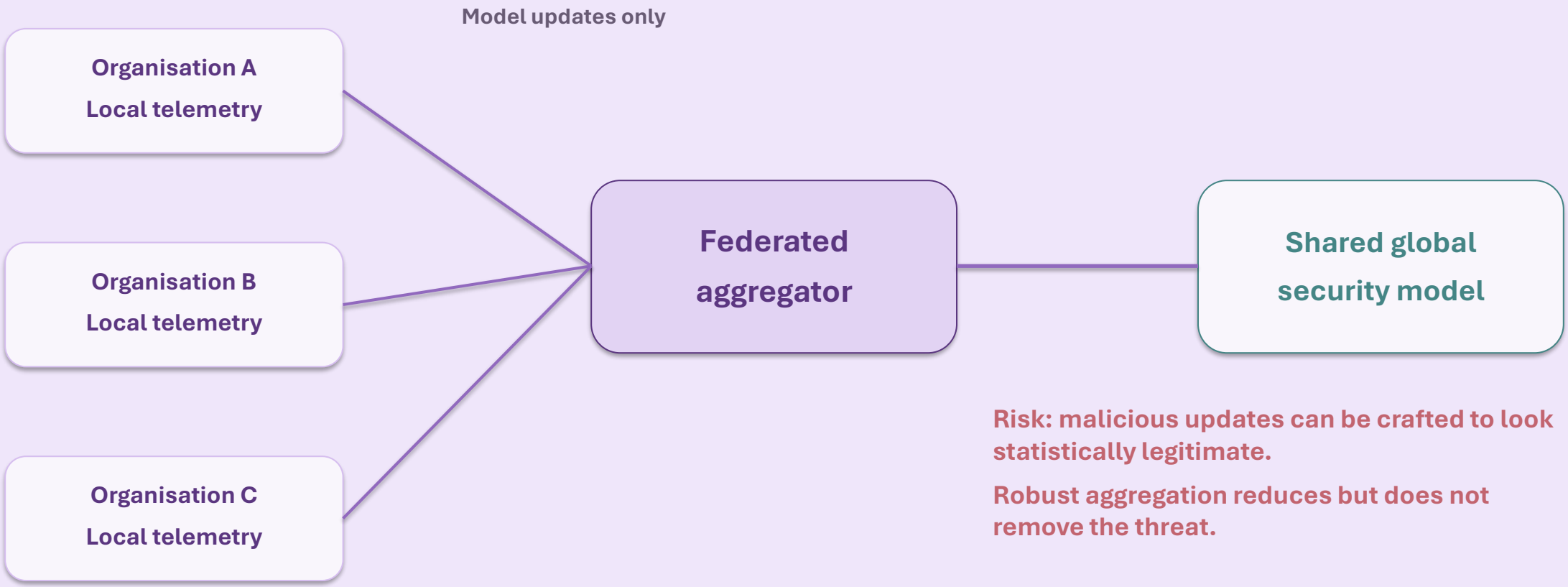
Part of the drop was a calibration artefact; rare attack classes stayed poorly detected.

Source: one federated-IDS study (UNSW-NB15, DP-SGD, ten IID clients) [26].

Non-IID data: accuracy fell by up to about 55%, 51% and 11% (keyword spotting, CIFAR-10, MNIST) vs centralised training [25].

Federated learning: collaboration without blind trust

Privacy and resilience must be engineered together.



Risk: malicious updates can be crafted to look statistically legitimate.

Robust aggregation reduces but does not remove the threat.

Non-IID data and inconsistent local environments further complicate convergence and global model quality.

RQ3 findings | Explainability is the trust interface

High model performance alone does not guarantee analyst confidence or adoption.

SHAP

Assigns feature contributions to a prediction using cooperative game theory.

LIME

Builds a local, interpretable approximation around an individual prediction.

Important limitation

Both are post-hoc approximations—not faithful access to a model's internal reasoning.

18 / 358

studies focused on XAI

3 / 18

included formal analyst user studies

14 / 18

XAI studies used SHAP or LIME

Qualitative evidence: analysts must manually validate large volumes of alarms and call for interpretable, meaningful alarms.

The central synthesis | Three risks, one system

Optimising each dimension independently can leave the overall system unsafe.

ROBUSTNESS

Can the model withstand manipulation?

PRIVACY

Can organisations collaborate without exposing sensitive telemetry?

EXPLAINABILITY

Can analysts understand and appropriately act on model outputs?

CO-DESIGN AGAINST A SHARED THREAT & USABILITY MODEL

Trustworthy, resilient and sustainable industrial AI security

Engineering implications and research priorities

From isolated controls to an integrated deployment strategy.

1. Test adaptively	Evaluate models against white-box, black-box, poisoning, backdoor and Byzantine threats—not only clean benchmarks.
2. Budget privacy explicitly	Set budgets through deployment tests, not assumption; report ε, utility loss, convergence time and overhead together.
3. Harden federation	Combine robust aggregation with participant monitoring, update validation and threat-aware governance.
4. Evaluate explanations with people	Use security analysts and realistic time-pressured tasks; measure decision quality, not only explanation fidelity.
5. Co-design the system	Jointly specify robustness, privacy and human oversight before industrial deployment.

Conclusion | Designing for resilient industrial AI

The paper's central message.

AI-driven security is only as dependable as the trustworthiness of the AI models, data-sharing architecture and human decisions around it.

Adversarial resilience

Defence must be adaptive and lifecycle-wide.

Privacy-aware collaboration

Privacy guarantees must be weighed against utility and Byzantine risk.

Analyst-centred XAI

Explanations need evaluation in real operational contexts.

Future direction: standardised federated telemetry protocols, common Byzantine detection and DP-accounting interfaces, and interdisciplinary human-centred evaluation.

DECLARATION

Use of generative AI: ChatGPT assisted with restructuring and formatting the manuscript and the presentation slides.

Thank you