Digitale Resilienz im Gefechtsfeld der Zukunft: Föderale KI-Sicherheit durch Digitales Watermarking

DWT Cyber Defence Conference, Bonn, Deutschland, Dezember 2025

Felix Ott, Redwanul Karim, Lucas Heublein, Jaspar Pahl, and Tobias Feigl
Fraunhofer IIS, Fraunhofer-Institut für Integrierte Schaltungen IIS; felix.ott@iis.fraunhofer.de

Motivation

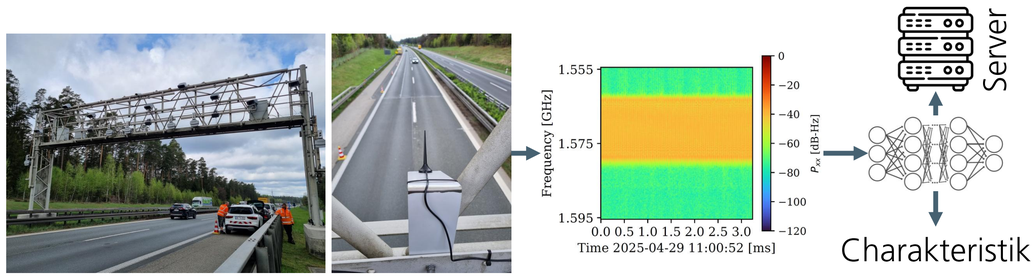
Problem

- Föderales Lernen im militärischen Umfeld erzeugt neue Verwundbarkeiten wie Model Poisoning und Identitätsdiebstahl kompromittierter Knoten.
- Die Integrität und Authentizität einzelner Modell-Updates lässt sich ohne zentrale Kontrolle nur schwer sicherstellen.
- Klassische kryptografische Schutzmechanismen können zukünftigen Bedrohungen – etwa durch Quantencomputer – nicht zuverlässig standhalten.

GNSS-Interferenzerkennung

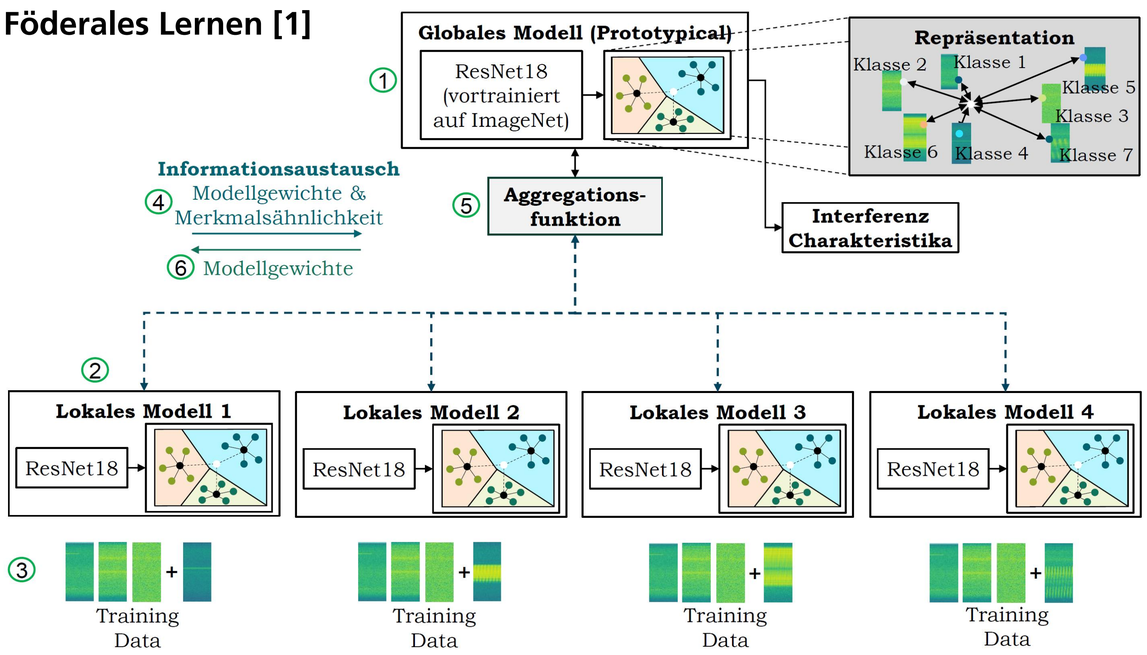
Anwendungsbeispiel

- Ziel: Detektion, Klassifikation und Charakterisierung von Interferenzen.
- Über verteilte Knoten hinweg durch Informationsaustausch über Server.
- Modell-Generalisierung durch föderales Lernen mit Gewichtsverteilung.



Konzept

Föderales Lernen [1]



Methode

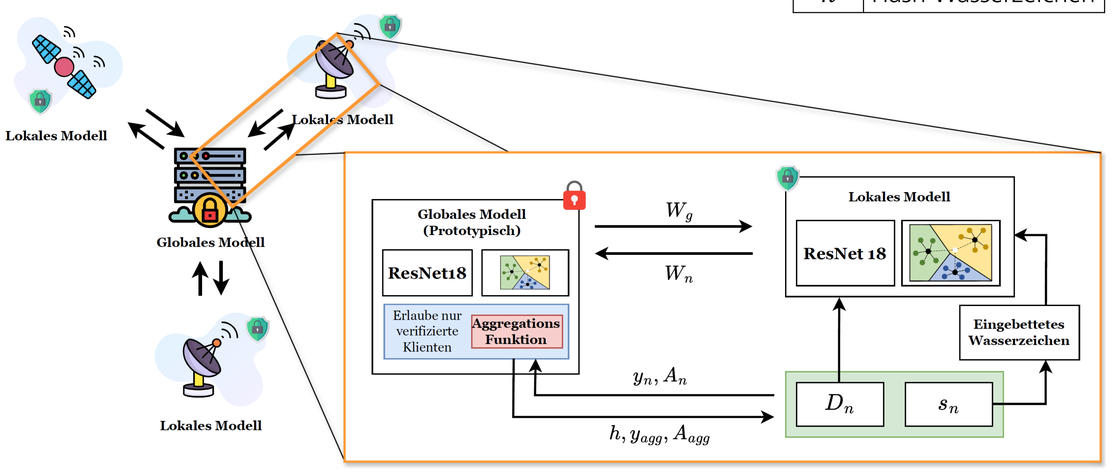
Herausforderung

- **Vertrauensproblem verteilter Knoten:** Militärische Sensorplattformen agieren in potenziell kompromittierten Umgebungen; fehlerhafte oder manipulierte Modell-Updates können die gesamte Föderation unterwandern.
- **Manipulationssichere Herkunfts- und Integritätsprüfung von Beiträgen:** Ohne robuste Authentizitätssicherung können Angreifer das föderale Lernsystem strategisch stören.
- **Schutz sensibler Rohdaten bei gleichzeitig hoher Reaktionsfähigkeit:** GNSS-Messstationen dürfen ihre Daten nicht zentral teilen, müssen aber dennoch schnell robuste Modelle gegen dynamische Störsignaturen erzeugen – ein Spannungsfeld zwischen Geheimhaltung, Latenz und Genauigkeit.

Lösung: Digitales Watermarking [2]

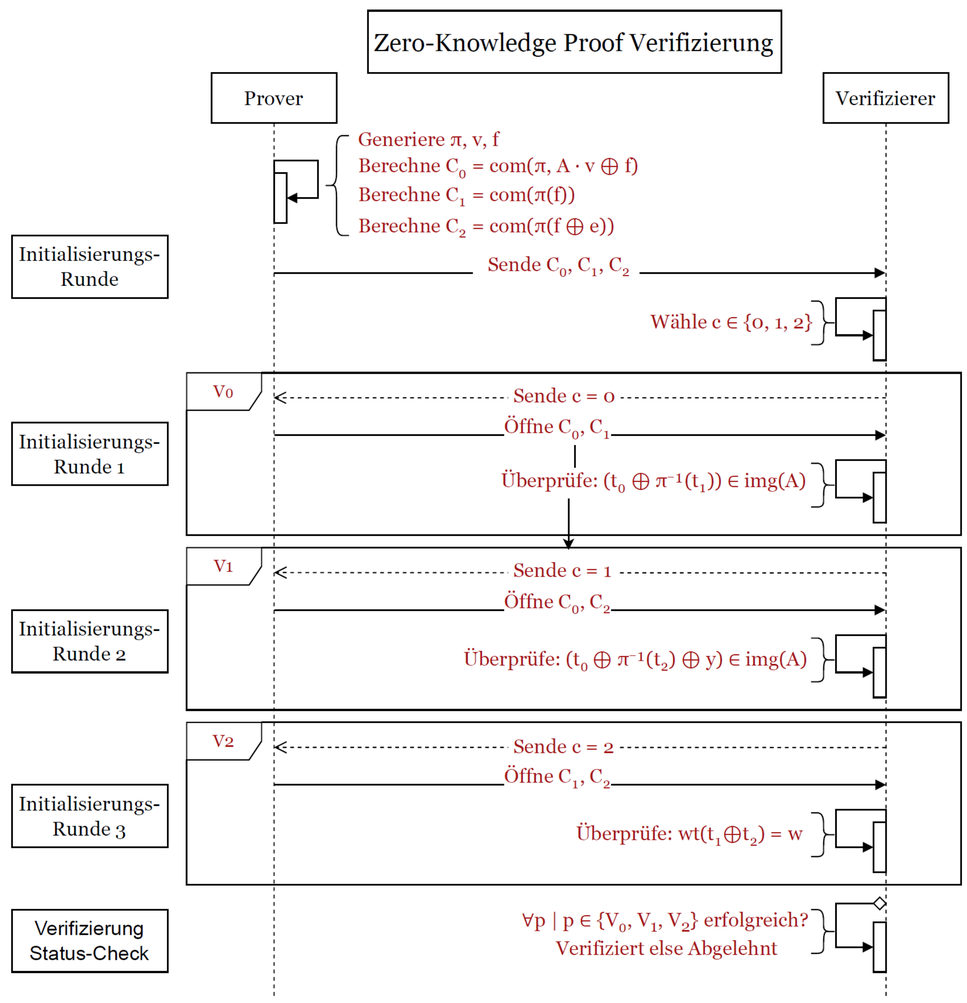
- Unsichtbare Herkunftsmarker in Modellgewichten/ Gradienten identifizieren legitime Knoten.
- Integritätsprüfung jedes Updates erkennt manipulierte oder gefälschte Beiträge.
- Vertrauensebene ohne zusätzliche Verschlüsselung, da Authentizität direkt im Modell verankert ist.

W_n	Lokale Gewichte
W_g	Globale Gewichte
D_n	Lokaler Datensatz
s_n	Nachweis
A_n	Matrix
A_{gg}	Aggregierte Matrix
y_n	Öffentlicher Vektor
y_{gg}	Aggregierter Vektor
h	Hash Wasserzeichen



Federated Zero-Knowledge Proof (FedZKP) [3]

- Eigentumsnachweis ohne Preisgabe von Informationen.
- Schutz vor Modell-Diebstahl und unbefugten Beiträgen.
- Geringe Zusatzlast im FL-Prozess.



Ergebnisse

Methode	LinearProbe	NCM	Prototypical	Linear + Mix + Focal	WM Detect
Fixed epoch (7)	87.7	83.1	88.3	89.3	**100.0**
FedAvgM[1]	89.1	87.0	90.9	90.6	-
MMD + ZKP	88.6	84.5	89.6	90.2	**100.0**
FedZKP[3]	**89.9**	**88.0**	**91.3**	**91.4**	**100.0**

1 N. S. Gaikwad, L. Heublein, N. L. Raichur, T. Feigl, C. Mutschler, and F. Ott. „Federated Learning with MMD-based Early Stopping for Adaptive GNSS Interference Classification“. In *IEEE NOMS*, HI, May 2025.

2 N. Sheybani, A. Ahmed, M. Kinsy, and F. Koushanfar. „Zero-Knowledge Proof Frameworks: A Systematic Survey“. In *arXiv: 2502.07063*, April 2025.

3 W. Yang, Y. Yin, G. Zhu, H. Gu, L. Fan, X. Cao, and Q. Yang. „FedZKP: Federated Model Ownership Verification with Zero-Knowledge Proof“. In *arXiv:2305.04507*, May 2023.

Acknowledgments:
DARCII: 50NA2401
PaiL: 50NP2506
Bundesministerium für Wirtschaft und Klimaschutz
Bundesministerium für Verkehr
DLR