

Security mechanisms for artificial intelligence in agricultural tractor systems

Rachid Batess, Ralf Leiter, Younes El Fellah

Artificial Intelligence (AI) is increasingly deployed in agricultural tractors for safety-critical functions such as perception and anomaly detection. As the advent of AI-based anomaly detection systems (ADSs) for agricultural tractor systems gains momentum, it has raised a number of concerns in the agricultural community. While numerous research efforts have emerged on the design, deployment and optimisation of ADSs, the defence against potential attacks on the ADSs is largely ignored in the literature. In this work, we investigate the security issues of ADSs and propose a defence-in-depth security framework to counter four types of attacks on agricultural tractor systems – white-box, black-box, grey-box and model tampering attacks. Experimental results on real-world CAN bus data recorded from an agricultural tractor (Fendt 211) show that the long short-term memory (LSTM)-based ADS and the lightweight machine learning-based ADS on CAN bus data are highly vulnerable to black-box attacks: under attack, the detection accuracy of the LSTM-based ADS decreases by up to approximately 50% relative to the baseline (unattacked) model for stealthy low-magnitude spoofing, and that of the lightweight ADS by up to approximately 45%, in both cases in a scenario-dependent manner. Meanwhile, our threat analysis and risk assessment (TARA) for the LSTM-based ADS identified 18 threats to five assets, with 3 at high risk, 9 at medium risk and 6 at low risk. This paper introduces a new research area for the security of ADSs in agricultural tractor systems and the design of robust AI components in tractors.

Keywords

Artificial intelligence security, agricultural tractor systems, anomaly detection, LSTM, black-box attacks, threat analysis and risk assessment (TARA), cybersecurity, machine learning

As reported in recent reviews of artificial intelligence in agriculture (Ge et al. 2025, Zhu et al. 2025), AI is used for object detection, predictive maintenance and exception detection in highly automated autonomous driving scenarios. Going from basic assist functions to highly automated driving up to SAE Level 4/5 is an enormous leap forward. This will be the use of many new functions that make tractor operation safer, more efficient and more comfortable, but also presents a whole new level of cybersecurity threats that must be fully understood (Lee 2025).

The security profile of a modern agricultural tractor differs fundamentally from that of a standard passenger car. Unlike public roads, which are typically paved and monitored through permanent closed-circuit television (CCTV) and camera surveillance, most agricultural fields lack such infrastructure-based monitoring, providing substantially fewer external means of detecting unauthorised access to, or tampering with, the vehicle. This absence of external monitoring is not, however, the primary security concern addressed in this paper. Modern passenger cars increasingly rely on LTE connectivity and numerous electronic control units (ECUs) equipped with built-in authentication

mechanisms, such as the Controller Area Network (CAN) bus. The CAN bus was designed at a time when security and authentication were not primary design considerations, and it remains impractical to retrofit today's advanced security functions into control units with such limited computational resources (MUHAMMAD and BAE 2022). Connected AI-based agricultural products are highly vulnerable to white-box, black-box and grey-box attacks, as well as to model tampering (VASSILEV et al. 2025), and there is little reason to expect this vulnerability to diminish as tractor autonomy increases.

The CAN bus is a multi-master automotive bus standard for real-time communication between ECUs. Open standard CAN bus communication has been proven to be vulnerable to attacks such as CAN message injection, message replay and denial-of-service attacks (KISHORE et al. 2022). Most present-day AI systems deployed in agricultural machinery, such as those used for object detection, predictive maintenance, and exception detection, as noted above, rely primarily on dedicated perception sensors (e.g., cameras and other onboard sensors) rather than on CAN bus data. But future integration of AI systems with the CAN bus is highly probable, because the CAN bus has vast amounts of information flowing through it. A breach of the CAN bus security can thus affect the security of the overall AI system. The paper thus bridges the communication security gap of legacy CAN bus and the AI model security.

The primary objective of this study is to investigate the security of the AI units used in agricultural tractor applications. To achieve this objective, four specific goals are pursued: determining the attack vectors applicable to learning-based systems on the agricultural tractor; identifying the countermeasures that can be applied to the tractor's embedded systems; assessing the vulnerabilities of the two anomaly detection systems (ADS) implementations evaluated in this study, namely the long short-term memory (LSTM)-based ADS and a lightweight machine-learning-based ADS, under black-box attacks via CAN bus signals; and performing a threat analysis and risk assessment (TARA) on the LSTM-based ADS to classify the identified threats and derive corresponding countermeasures.

This paper makes the following contributions to the field of agricultural tractor AI security:

- **Attack taxonomy:** We propose a taxonomy classifying the attack vectors relevant to adversarial attacks on the AI components of an agricultural tractor into four categories: white-box, black-box, grey-box, and model tampering attacks (VASSILEV et al. 2025).
- **Experimental validation:** We present an empirical study based on real-world CAN bus data recorded from an agricultural tractor (Fendt 211; Götz et al. 2025). The experimental results show that the detection accuracy of the LSTM-based ADS decreases by up to approximately 50% under black-box adversarial attacks for stealthy low-magnitude spoofing, and that of the lightweight ADS by up to approximately 45% under statistics-preserving injection, both in a scenario-dependent manner. These magnitudes are consistent with the 10–60% degradations reported for automotive intrusion detection systems under adversarial manipulation (LONGARI et al. 2025).
- **TARA results:** We conduct a TARA for the LSTM-based ADS, identifying 18 threats across five assets. The identified threats are classified and prioritised according to risk level to facilitate security decision-making.
- **Defence-in-depth framework:** We propose a multi-layered security framework, informed by the identified priority threats, for selecting and applying appropriate countermeasures within the resource constraints of an embedded system. This framework is intended to provide OEMs and Tier-1 suppliers with actionable guidance for prioritising countermeasures under limited resource availability.

Related works

A lot of research has been carried out for integrating AI in the general agricultural or automotive applications. To the best of our knowledge, this paper presents one of the first dedicated security analyses of agricultural tractor AI systems, complementing related efforts that assess the security of AI-enabled functionality in smart farming more broadly, such as the AI-TESTFIELD initiative (KRAUSE et al. 2025). RANA and KHATRI (2024) have done a very comprehensive survey of AI applications in automotive intelligence. With the evolution of AI in the vehicles for applications of ADAS, vehicle maintenance, reducing emissions and communication etc. the automotive system has become a very complex system with a large attack surface which needs a deeper security analysis.

Adversarial attacks have evolved substantially since the first adversarial examples were introduced in the literature (SZEGEDY et al. 2014). Prior work has shown that black-box attacks remain highly effective against LSTM models (DING et al. 2023), which, together with gated recurrent unit (GRU) models, are widely used architectures for sequential data. MUHAMMAD and BAE (2022) released a survey paper on efficient adversarial defence. They compared the efficiency of defence methods in the literature and concluded that adversarial training, defensive distillation and input preprocessing are the efficient defences, and the input preprocessing is highly computational.

The discussion by WÄSCHLE et al. (2022) focuses on AI safety in highly automated vehicles. Their conclusion is that conventional safety measures are not applicable to the current state of the art in AI technology and that new safety concepts are required. This work is also relevant to agricultural tractors, where the safety and security aspects are highly coupled. KISHORE et al. (2022) argue that the safety cases for AI-based systems must be holistic and that their development should be carried out in accordance with recognised software engineering standards. In general, they argue that safety should be considered as a dynamic control rather than a static property.

GKOUNTAKOS et al. (2024) conducted a review of recent deep learning techniques applied to anomaly detection in precision agriculture. They believe that the LSTM model could be very useful for agriculture as it is very good at identifying temporal patterns from the sensor data. SINGH et al. (2023) introduced a supervised machine learning approach to evaluate the tractor ride comfort. This paper is considered as one of the papers that have applied ML in agricultural machinery. KADHIM et al. (2025) have done a review on the current cybersecurity frameworks of the intelligent farming devices. They believe that AI, although introduced many new types of threats, it also has the capabilities to use the collected data to protect against the emerging cyber threats.

Although many countermeasures have been suggested in the literature, a full life cycle assessment (LCA) of agricultural tractors which also considers the tractor's own vulnerability as well as a structured risk assessment relating to tractor use has not yet been carried out. This paper reports an experimental investigation into possible attacks on tractors and the application of the TARA methodology in a formal manner.

System model and threat model

AI-based ADS architecture in agricultural tractors

This work presents an AI-based ADS for an autonomous tractor using a pipeline consisting of four sub-modules: data acquisition, data preprocessing, model inference and output handling. The data acquisition sub-module acquires the raw sensor data and CAN bus messages from the tractor electronic

control units. The data preprocessing sub-module normalises the acquired data and divides it into fixed time length windows in order to feed the sequential models in the model inference sub-module (LSTM or the lightweight ML-based ADS model). The data sequences are then classified as normal or anomalous by the models. Finally, the output handling sub-module is responsible for generating alert/diagnostic messages and it is expected to be interfaced with the lower stage ECUs (BHATTACHARYA et al. 2024).

Based on the ADS architecture illustrated in Figure 1, this section identifies the key factors that influence model security at each stage of the pipeline: (i) the vulnerability of the model itself to adversarial attacks, and (ii) the security of the data supplied to the model, encompassing CAN interface ingress filtering, secure storage of model weights, and the authentication mechanisms required for communication with the subsequent block in the pipeline. Figure 1 provides a detailed view of the architecture and the corresponding attack surfaces at each stage.

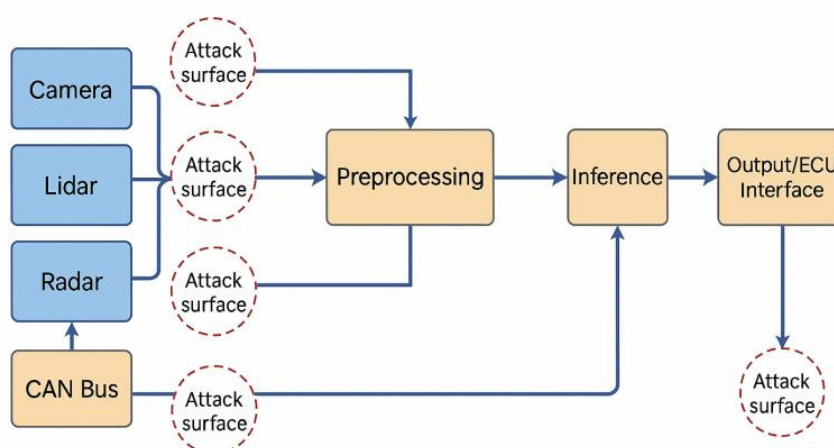


Figure 1: Architecture of an AI-based ADS in an agricultural tractor, showing the data flow from sensors and CAN bus through preprocessing and inference to output/ECU interface, with attack surfaces indicated at each stage

In practice, agricultural tractors and their attached implements are frequently supplied by different manufacturers and communicate via the ISOBUS (ISO 11783) standard. Whether AI-based inference should be executed on the tractor's electronic control unit or on the implement's own control unit within this heterogeneous ecosystem remains an open question and is the subject of ongoing discussion among manufacturers; the architecture evaluated in this paper is agnostic to this choice, as the ADS pipeline illustrated in Figure 1 is described independently of which physical control unit hosts the inference stage.

Attack taxonomy and adversary capabilities

This study adopts a four-category attack taxonomy (Table 1) tailored to the agricultural tractor context, encompassing white-box, black-box, grey-box and model tampering attacks. Consistent with the experimental scope of this work, the remainder of this paper focuses on black-box attacks, which we consider the most realistic threat in the agricultural tractor context; the white-box, grey-box and model-tampering categories are characterised in Table 1 and identified as directions for future work (see Limitations).

Table 1: Attack taxonomy for agricultural tractor AI systems with adversary capabilities and feasibility assessment

Attack category	Adversary knowledge	Primary entry points	Feasibility in tractor context
White-box	Full access to architecture, parameters, training data	Direct model access, development environment	Low (requires insider access)
Black-box	Query-only access; observes input-output behaviour	Compromised CAN gateway, injected messages	High (realistic field scenario)
Grey-box	Partial knowledge of architecture or training patterns	Side-channel observation, partial documentation	Medium
Model tampering	Direct modification of weights or update pipeline	OTA update channel, physical storage access	Medium-High (OTA exposure)

We empirically verify the feasibility of performing black-box attacks on agricultural tractors. In this scenario, the term black-box attack is particularly applicable because an attacker may only learn the input/output behaviour of the model, possibly via a compromised CAN gateway or by injecting CAN messages. This is possible even without knowing the internal parameters of the model. We also identify model tampering as a possible threat, as current over-the-air (OTA) update pipelines may not be properly secured, and access to ECUs may be possible when a tractor is in maintenance mode.

Evaluated ADS implementations and attack methodology

LSTM-based ADS

Our LSTM-based ADS model utilises a recurrent neural network (RNN) to make use of the sequential nature of the CAN bus and sensor time series data. The model is designed using stacked LSTM layers followed by dense layers that predict the expected value of a monitored signal from the temporal patterns of the vehicle operation data. Each data window is then classified as normal or anomalous by comparing the prediction error against a detection threshold calibrated on validation data. The LSTM model is chosen because the cell state in the LSTM layer can learn long term temporal patterns. This is important for learning the anomalies in the vehicle operation data which may occur far apart in time and may not be detectable by learning the patterns at individual time steps.

The model captures the temporal patterns associated with a tractor's "normal behaviour" – defined here as the range of CAN bus signal patterns observed during labelled data collection under standard operating conditions, and is capable of signalling anomalies under abnormal operating conditions. The model was trained and validated on labelled CAN bus data which was collected while the tractor was operating in normal conditions. In this scenario, the LSTM-based ADS learns the statistical characteristics of the normal temporal patterns. Figure 2 shows the detailed workflow from the process of temporal windowing to the inference of the anomaly score.

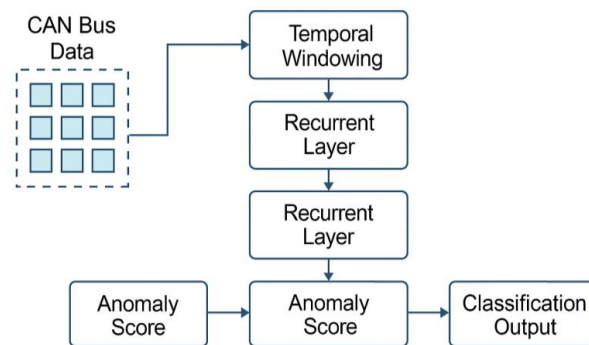


Figure 2: LSTM-based ADS workflow: temporal windowing of CAN bus data, LSTM inference through stacked recurrent layers, and anomaly score generation with classification output

Lightweight ML-based ADS

The proposed lightweight machine learning (ML) based ADS is very light and employs a fast algorithm based on conditional self-information matrix. It is also suitable for the resource limited embedded environment. The proposed system performs two tasks. In the training phase, the model is trained by learning the normal CAN bus messages and a conditional self-information matrix is constructed (Figure 3). The conditional self-information of a velocity V_i given that the velocity V_j has occurred before is calculated by:

$$E_{i \leftarrow j} = -\log P(V_i | V_j)$$

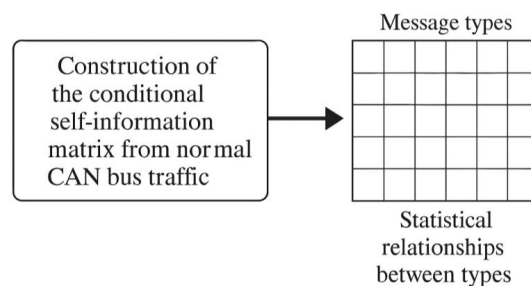


Figure 3: Training phase of the lightweight ML-based ADS: construction of the conditional self-information matrix from normal CAN bus traffic, showing statistical relationships between message types

During the testing phase, the system calculates the conditional self-information of the incoming CAN messages and compares it with the learned matrices; values that deviate substantially from the expected distributions are flagged as anomalies (Figure 4).

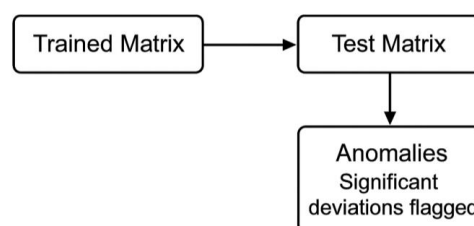


Figure 4: Testing phase of the lightweight ML-based ADS: comparison between the trained matrix and the test matrix, with significant deviations flagged as anomalies

Black-box attack methodology

In this work, we investigate black-box attacks on ADS using the substitute-model approach, as by FENZL et al. (2021). We outline the attack process in four steps (Figure 5):

1. Data collection: The attacker collects the CAN bus data transmitted by the ECU under regular operating mode to obtain the input/output behaviour of the model.
2. Surrogate model training: An LSTM surrogate model is trained to emulate the real model's decision boundary based on the collected data.
3. Adversarial sequence generation: The generated surrogate model is used to produce adversarial sequences of CAN bus data that are intended to evade detection by the real model.
4. Attack execution: The generated adversarial sequences are transmitted to the ADS either by injecting them on the CAN bus or by replaying them, and the attack's performance is measured by evaluating the decrease in the model's detection performance relative to the unattacked baseline.

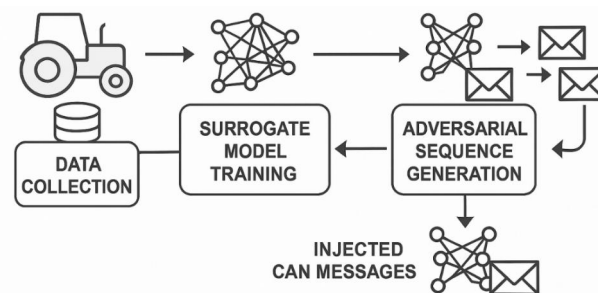


Figure 5: Black-box attack workflow: data collection from CAN bus, surrogate model training, adversarial sequence generation, and target model probing through injected CAN messages

Dataset and experimental setup

All experiments are based on the publicly available real-tractor CAN bus dataset described below; the corresponding preprocessing and evaluation scripts are provided with the paper to ensure full reproducibility. The CAN bus data are taken from the open dataset “Agricultural Load Cycles: Tractor Mission Profiles from Recorded GNSS and CAN Bus Data” (Görz et al. 2025), recorded from Fendt agricultural tractors during real field operations in southern Germany and released under a CC-BY-4.0 licence (Zenodo, doi:10.5281/zenodo.14619787). This study uses the recordings of the Fendt 211 tractor. The signals are read from the J1939 engine bus and the ISOBUS (ISO 11783).

Two field-operation scenarios contained in the Fendt 211 recordings are used: power harrowing (soil tillage) and precision air seeding. All recordings correspond to field work; no road-transport data are included. The recordings were segmented at acquisition gaps longer than 2 s, and only contiguous segments of at least 150 s were retained.

Figures 6–9 show the wheel-based vehicle speed signal (SAE J1939 SPN 84, “WheelBasedVehicleSpeed”), decoded to its physical value in km/h. The observed field speeds lie between 0 and about 23 km/h, consistent with agricultural field operation. The LSTM-based ADS additionally uses the engine speed (rpm), the actual engine percent torque (%) and the engine fuel rate (L/h) as model inputs. The horizontal axes of Figures 6–9 are given in seconds. The signals were resampled to a uniform rate of 2 Hz (one sample every 0.5 s) prior to analysis; the depicted intervals therefore correspond to a few minutes of operation.

Because the dataset contains normal operation only, anomalies were introduced synthetically and are clearly documented. For the LSTM-based ADS, three anomaly types were injected into the speed signal over a contiguous interval of a recording: (i) speed spoofing (an additive offset of 2–6 km/h, modelling adversarial CAN message injection, threat T1); (ii) replay of an earlier speed segment (threat T2); and (iii) a frozen-value / hold anomaly (threat T17). For the lightweight ADS, a naive fuzzing injection (uniform random speed values) and an adversarial statistics-preserving injection (a sequence sampled from the detector's own learned transition model) were used.

Each signal was standardised (zero mean, unit variance) using statistics computed on the training partition only, and divided into sliding windows of 32 samples (16 s at 2 Hz). The LSTM-based ADS is a predictor that estimates the wheel-based speed at the end of each window from the three engine-side inputs; a window is flagged as anomalous when its squared prediction error exceeds a threshold. The network has two stacked LSTM layers of 64 hidden units followed by dense layers (64→32→1) and was trained with the Adam optimiser (learning rate 1×10^{-3} , batch size 256, 10 epochs, mean-squared-error loss). The lightweight ADS builds a 24-bin conditional self-information (CSI) matrix over the discretised speed signal and scores each window by its mean bigram self-information. Data were split per recording chronologically into 60% training, 20% validation and 20% test.

The black-box attack follows the previously described substitute-model approach. A surrogate predictor (two LSTM layers, 48 units) was trained using query access to the target model only (the target's predictions on collected inputs as training labels). Adversarial perturbations of the injected engine-side frames were then computed on the surrogate by projected gradient descent and refined by query-based Natural Evolution Strategies optimisation, bounded to an L^∞ ball of two standard deviations, so that the target model's prediction follows the spoofed speed and the anomaly evades detection. For the lightweight ADS, the statistics-preserving injection described above constitutes the black-box attack.

Detection accuracy is reported as balanced accuracy – the mean of the true-positive rate (anomalous windows correctly flagged) and the true-negative rate (normal windows correctly passed) – at a decision threshold fixed at the 99th percentile (LSTM) or 99.5th percentile (CSI) of the validation-set scores. Each detector is evaluated per scenario, under baseline (no attack) and under the black-box attack. All experiments use a fixed random seed and are deterministic.

Experimental results

Throughout this section, model performance is quantified as anomaly detection accuracy relative to a baseline (unattacked) model; a decrease in performance therefore corresponds to an increased rate of missed anomalies (false negatives) and/or erroneous anomaly alerts (false positives), either of which can compromise the operational safety of the system.

Impact on LSTM-based ADS

The uncompromised LSTM-based ADS preserves the temporal patterns in the CAN bus data while identifying unusual paths. In Figure 6, the blue curve shows the recorded signal from the non-anomalous test data, and the red curve shows the corresponding value predicted by the uncompromised LSTM-based ADS.

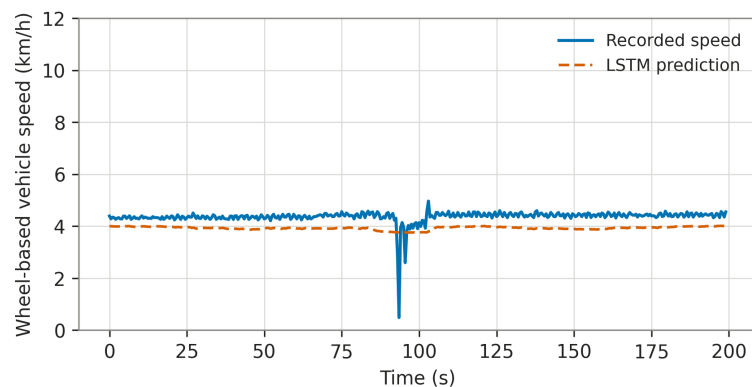


Figure 6: Uncompromised LSTM-based ADS on non-anomalous test data (power-harrowing scenario): recorded wheel-based vehicle speed (blue, km/h) and the value predicted by the LSTM (red), with time in seconds; the close agreement between the two curves corresponds to the correct classification of these sequences as normal

Following the black-box attack, the detection performance of the model can decrease substantially. When a stealthy speed spoof is injected (Figure 7, top panel, shaded interval), a naive injection produces a large prediction residual and is reliably flagged (Figure 7, bottom panel, dashed curve above the threshold). The black-box attack, however, perturbs the accompanying engine-side signals within a bounded budget so that the LSTM prediction follows the spoofed speed; the residual then stays below the detection threshold throughout the attacked interval (solid curve), and the anomaly evades detection.

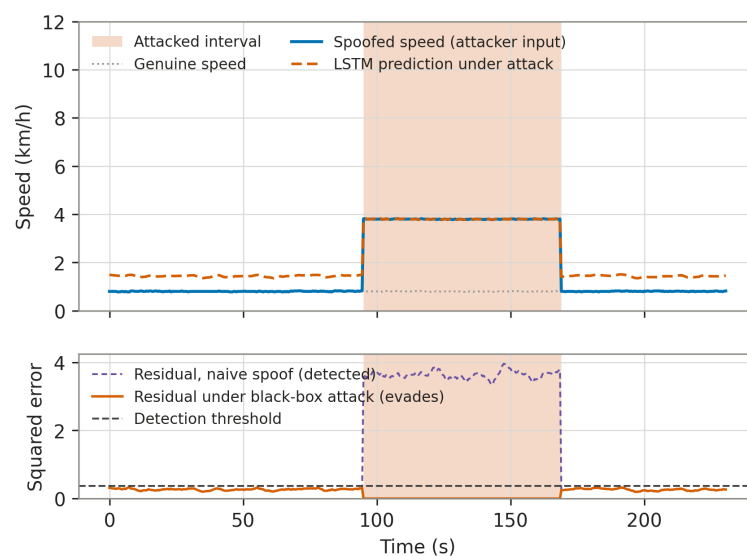


Figure 7: LSTM-based ADS under black-box attack (precision-seeding scenario); top: genuine speed (grey), the spoofed speed injected by the attacker (blue) and the LSTM prediction under attack (red), which is driven to follow the spoof; bottom: the squared prediction error for a naive spoof (dashed, detected – above the threshold) and under the black-box attack (solid, evades – below the threshold); speed in km/h, time in seconds

The effectiveness of the attack is scenario- and magnitude-dependent (Table 2). For stealthy low-magnitude spoofing (2–3 km/h) in the precision-seeding scenario, the black-box attack reduces the LSTM-based ADS detection accuracy from 0.99 to as low as 0.49, a relative decrease of up to approximately 50%; larger spoofs (≥ 4 km/h) and the power-harrowing scenario remain reliably detect-

ed, because the bounded perturbation of the engine-side signals is insufficient to make the prediction follow a large speed offset. This magnitude is consistent with the 10–60% degradations reported for automotive intrusion detection systems under adversarial manipulation (LONGARI et al. 2025). This suggests that the powerful temporal learning capabilities of LSTMs are not sufficient to counteract the design of adversarial perturbations that can be crafted to move along a specific time dimension and cross the decision boundary, at least for the stealthy manipulations that a bounded black-box adversary can realistically produce.

Table 2: Per-scenario balanced detection accuracy of both ADS implementations, baseline vs. black-box attack (Fendt 211 dataset)

Model / scenario	Baseline accuracy	Attacked accuracy	Relative decrease
LSTM-based ADS			
Precision air seeding	0.99	0.49	≈ 50%
Power harrowing (Field 1)	0.99	0.99	≈ 0%
Power harrowing (Field 2)	0.99	0.99	≈ 0%
Lightweight CSI-based ADS			
Power harrowing (Field 1)	0.90	0.50	≈ 45%
Power harrowing (Field 2)	0.84	0.47	≈ 43%

Note: Balanced accuracy is the mean of the true-positive and true-negative rates. For the LSTM-based ADS the attack is effective for stealthy low-magnitude spoofing (2–3 km/h) on the seeding scenario; larger spoofs and the power-harrowing scenario stay reliably detected. For the power-harrowing scenario, Table 2 reports, per field, the magnitude among those tested (2, 3, 4 and 6 km/h) that best illustrates the reliably-detected outcome described above: for Field 1, the magnitude producing the largest measured accuracy drop (2 km/h), which already rounds to the reported baseline/attacked values; for Field 2, a larger magnitude (4 km/h) is reported instead of the smallest tested one (2 km/h), because at 2 km/h the baseline balanced accuracy for Field 2 is already comparatively low (≈0.62) even without an attack, and the attack itself causes only a minimal further change at that magnitude (≈0.62 to ≈0.62); this magnitude is therefore not a representative basis for illustrating the attack's effect. For the lightweight ADS the seeding scenario is omitted, as its baseline CSI detection there is already weak.

Impact on lightweight ML-based ADS

The lightweight ADS is fast and interpretable and was implemented based on a dependency matrix. The goal of this model is to find anomalies by identifying the points which deviate from the normal statistical behaviour of the CAN bus traffic. The behaviour of the uncompromised model on anomaly-containing test data is shown in Figure 8, where a naive injection produces a clear rise in the conditional self-information (CSI) deviation score above the detection threshold.

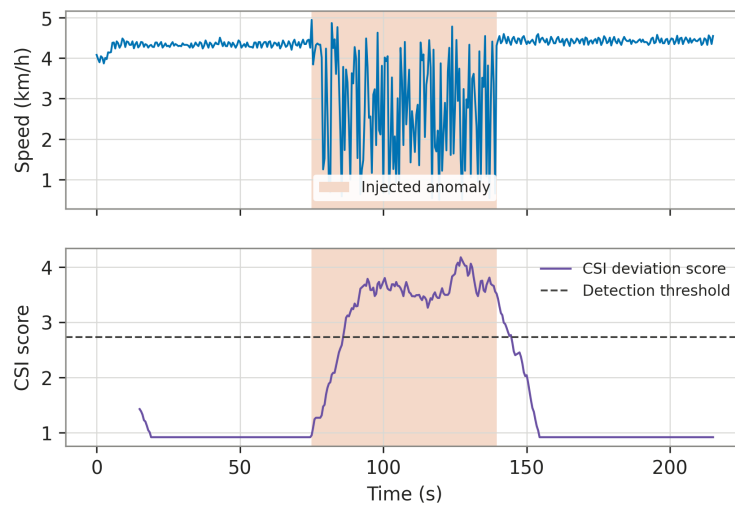


Figure 8: Lightweight ML-based ADS (uncompromised model) on anomaly-containing test data (power-harrowing scenario): monitored wheel-based vehicle speed (top, km/h) and the CSI deviation score (bottom). The naive injection (shaded interval) drives the CSI score above the detection threshold and is correctly identified; time in seconds

However, it turns out that a black-box attacker can craft CAN bus messages that preserve the expected conditional statistical relationships, but are injected with malicious intent, so that the attacks fall outside of the anomaly detection. The resulting compromised model cannot detect the adversarial anomalies that were crafted to look statistically normal (Figure 9).

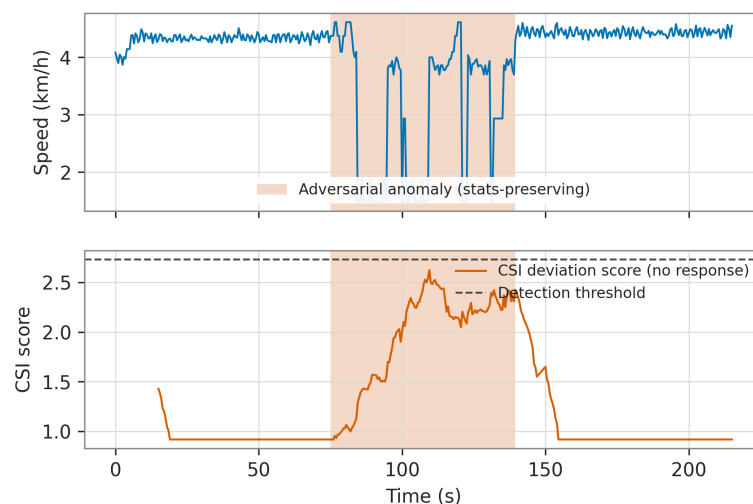


Figure 9: Lightweight ML-based ADS under black-box attack (power-harrowing scenario): an adversarial anomaly generated by sampling from the detector's own learned transition model preserves the expected conditional statistics, so the CSI deviation score (bottom) remains below the detection threshold throughout the attacked interval (shaded) and the anomaly is missed; same axes as Figure 8

For the lightweight ADS, the statistics-preserving injection reduces the detection accuracy from up to 0.90 (baseline, power-harrowing scenario) to approximately 0.47–0.50 – that is, to chance level, missing essentially all injected anomalies – corresponding to a relative decrease of up to approximately 45%. As with the LSTM-based ADS, the effect is scenario-dependent (Table 2). Our experiments demonstrate that black-box attacks can compromise the reliability of deep learning and statistical anomaly detection in distinct ways: adding perturbations to the input sequences compromises the

accuracy of the LSTM-based anomaly detector, while controlling the expected conditional dependency statistics compromises the statistical dependency-based detector. In summary, our contributions highlight the requirement for robust defences that not only safeguard individual learning techniques, but also the overall dataflow and computation performed for anomaly detection.

Threat analysis and risk assessment (TARA)

Methodology and asset identification

The TARA for the LSTM-based ADS follows a 5-step process:

1. asset identification – the critical assets/components that need to be protected are identified (Table 3)
2. threat identification – all possible threats for each identified asset are enumerated
3. attack vector analysis – the possible attack vectors for each asset are investigated
4. risk assessment – the risk for each identified threat is calculated based on the likelihood and impact of the threat
5. countermeasure identification – possible countermeasures for mitigating high-risk threats are identified.

Five key assets are identified for the LSTM-based ADS (Table 3).

Table 3: Assets identified in the LSTM-based ADS TARA, with descriptions and primary security goals based on the CIA triad extended with AI-specific properties

Asset ID	Asset Name	Description	Primary Security Goals
IN_data	Input Data	Raw sensor data and CAN messages fed to the LSTM	Integrity, Availability
LSTM	LSTM-Model	Core recurrent network for anomaly detection	Integrity, Confidentiality, Robustness
OUT_data	Output Data	Classification results and anomaly indicators	Integrity, Authenticity
LSTM_In_DF	Input Data Flow	Communication channels for input data transfer	Integrity, Availability
LSTM_Out_DF	Output Data Flow	Communication channels for outputs to vehicle systems	Integrity, Authenticity

Risk assessment follows the Threat Analysis and Risk Assessment approach of ISO/SAE 21434 (ISO/SAE 2021). Likelihood and impact are each rated on a three-level ordinal scale (Low / Medium / High), based on the consensus of a small group of experts, and the risk level of each threat is determined from the risk matrix shown in Table 4.

Table 4: Risk matrix used to combine likelihood and impact into a risk level (three-level scheme in accordance with ISO/SAE 21434)

Likelihood / Impact	Low	Medium	High
Low	Low	Low	Medium
Medium	Low	Medium	High
High	Medium	High	High

Threat identification and risk assessment results

Based on the five TARA assets, a total of 18 threats have been identified. The threats are classified as 3 High, 9 Medium and 6 Low threats, applying the risk matrix of Table 4 uniformly to all threats. The majority of threats are associated with data integrity and model integrity. Figure 10 shows the STRIDE-based threat identification from which the threats were derived.

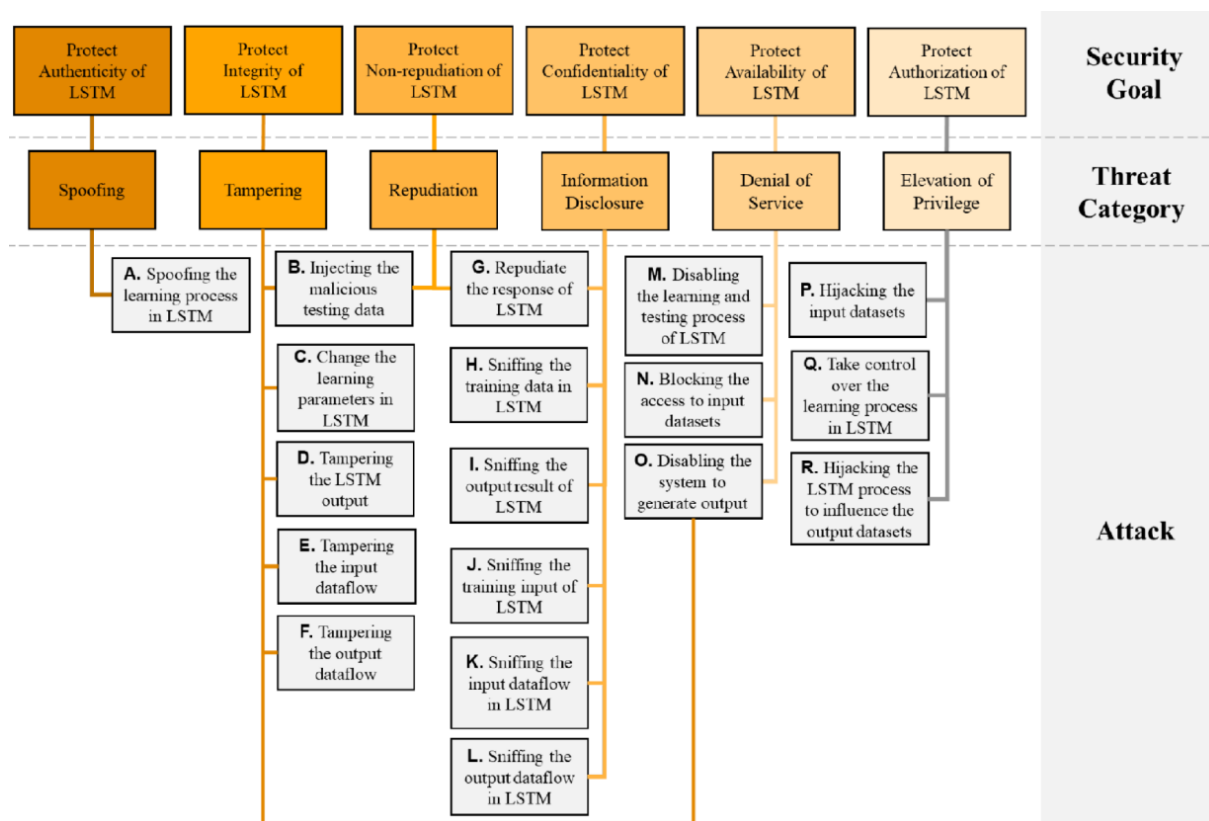


Figure 10: STRIDE-based threat identification for the LSTM-based ADS: security goals, the six STRIDE threat categories (spoofing, tampering, repudiation, information disclosure, denial of service, elevation of privilege) and the attacks enumerated per category (A–R), which were consolidated per asset into the threats T1–T18 assessed in Tables 5–7

At category level, the STRIDE analysis maps onto the assessed threats as follows: spoofing and tampering attacks correspond to the integrity threats against the input data, the model and the data flows (T1–T6, T11, T13–T16); information-disclosure attacks to the confidentiality threats (T7, T8, T12, T18); denial-of-service attacks to the availability threat (T17); and elevation-of-privilege attacks to the model-takeover threats (T9, T10). Repudiation attacks were not carried forward into the risk assessment, since non-repudiation is not among the primary security goals of the assessed assets (Table 3).

For the IN_data asset (Table 5), the main data input threats to ECU-based models are CAN message injection, data replay based on legitimate CAN messages that were previously captured from the Tractor, and sensor poisoning. However, it was found that the adversarial CAN message injection has the most impact on the safety and security of ECU-based models and thus fully black-box attacks are feasible.

Table 5: TARA results for IN_data asset: threats related to input data manipulation

Threat ID	Threat description	Attack vector	Likelihood	Impact	Risk level
T1	Adversarial CAN message injection	CAN bus compromise	High	High	High
T2	Replay attack with normal traffic	CAN bus access	Medium	Medium	Medium
T3	Sensor data manipulation	Physical sensor access	Low	High	Medium
T4	Data poisoning via compromised feed	Supply chain / maintenance	Low	Medium	Low

The LSTM model is the asset associated with the greatest number of threats (Table 6). High-risk threats include model weight tampering via OTA update compromise and adversarial attacks at inference time, whereas medium-risk threats include model extraction via brute-force querying and model exposure during maintenance.

Table 6: TARA results for LSTM model asset: model-specific threats and risk levels

Threat ID	Threat Description	Attack vector	Likelihood	Impact	Risk level
T5	Model weight tampering via OTA	Compromised update channel	Medium	High	High
T6	Adversarial runtime manipulation	Black-box query / CAN injection	High	High	High
T7	Model extraction via systematic queries	External interface access	Medium	Medium	Medium
T8	Unauthorised model access	Physical ECU access	Low	High	Medium
T9	Trojan insertion during training	Supply chain compromise	Low	High	Medium
T10	Model replacement with malicious version	OTA / physical access	Low	High	Medium

For the OUT_data and data flow assets (Table 7), output data threats concern attacks on classification results and information leaks which may enable model extraction. Data flow threats concern data integrity of communication channels between system components.

Table 7: TARA results for OUT_data and data flow assets (LSTM_In_DF and LSTM_Out_DF)

Threat ID	Threat Description	Asset	Likelihood	Impact	Risk Level
T11	Output manipulation to suppress alarms	OUT_data	Low	High	Medium
T12	Information leakage enabling model extraction	OUT_data	Medium	Medium	Medium
T13	CAN message injection on input data flow	LSTM_In_DF	Medium	Medium	Medium
T14	Replay attack on input data flow	LSTM_In_DF	Low	Medium	Low
T15	Output data flow tampering	LSTM_Out_DF	Low	Medium	Low
T16	Man-in-the-middle on output flow	LSTM_Out_DF	Low	Medium	Low
T17	Denial of service on input data flow	LSTM_In_DF	Low	Low	Low
T18	Unauthorised access to output data flow	LSTM_Out_DF	Low	Low	Low

The TARA analysis highlights five overarching security challenges for LSTM-based ADS in agricultural tractors: (i) protecting input data integrity against CAN bus manipulation, (ii) preventing unau-

thorised access to and modification of the LSTM model, (iii) enhancing adversarial robustness against runtime inference attacks, (iv) securing communication channels between system components, and (v) validating model outputs to detect manipulations before they propagate to vehicle control systems.

Defence-in-depth security framework

Framework overview

Information gained from experiments and from the TARA prioritisation exercise is used to derive a layered defence-in-depth strategy that can be used to counter the threats that were identified in the TARA. The derived strategy is designed to address the most critical threats while taking the resource constraints of embedded systems into account (Figure 11). The derived defence-in-depth framework is composed of four distinct layers corresponding to the threats in the TARA. It is important to distinguish the scope of the two parts of this work: the vulnerability of the two ADS implementations to black-box attacks was established experimentally in an offline evaluation, whereas the defence-in-depth framework presented in this section is a proposed set of countermeasures. The protective effectiveness of these countermeasures and their compatibility with on-board resource constraints have not yet been experimentally validated and remain the subject of the future work.

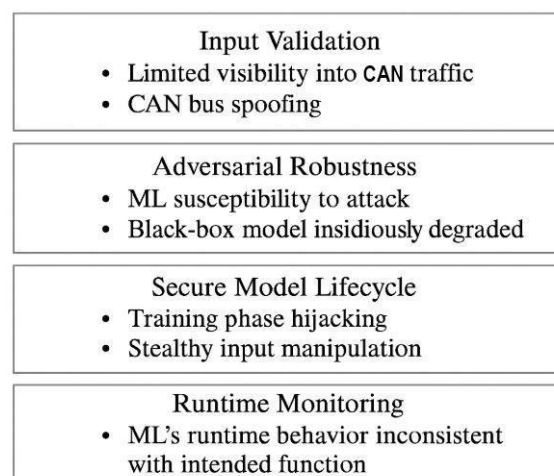


Figure 11: Defence-in-depth framework for agricultural tractor AI security: experimental findings and TARA-prioritised threats mapped to four security layers (input validation, adversarial robustness, secure model lifecycle, runtime monitoring)

Layer 1 – Input validation and sanitisation

This layer addresses the high-likelihood, high-impact threat of adversarial CAN message injection (T1), as rated in Table 5. The input validation is proposed to cover several aspects. Firstly, a statistical check would be performed on the received CAN messages to identify anomalous behaviour prior to the input being processed by the AI model. Secondly, a number of constraints would be enforced on sensor readings to verify that they are within valid physical ranges. Finally, rate limiting and authentication would be applied at the CAN gateway.

Input validation makes the attack surface for a black-box attacker smaller. Unwanted input is blocked if it does not match the statistical distribution of expected input values. According to WÄSCHLE

et al. (2022) a valid input validation has to cover three aspects: Statistical analysis, validation against the domain-specific input constraints and anomaly detection. The validation methods mentioned can also be used in the context of a reliable operation of modern ADS via the CAN bus.

Layer 2 – Adversarial robustness measures

The papers on adversarial robustness are related to the high-risk threat of T6. The main method to obtain the adversarial robustness is the method of adversarial training. The adversarial examples generated during the training process are used to make the model robust against similar attacks that are likely to occur at runtime. For example, ZHAO et al. (2022) reviewed adversarial training methods for deep learning and found that, although adversarial training can improve robustness against black-box attacks, it substantially increases computational cost and training time.

In addition, we investigated other strategies to enhance robustness, including the following: Randomised smoothing generates a set of classifiers trained on the same criteria but with different input perturbations and returns a certificate bounding the perturbations for which the input vector remains within the correctly classified class. Defensive distillation trains a second, secondary network (the distillation model) on the softened output of the primary classifier; this is thought to reduce the network's sensitivity to small perturbations (noise). These and any other selected methods should take into account the computational resources that can reasonably be expected to be available on the embedded ECU of an autonomous tractor.

Layer 3 – Secure model lifecycle controls

This layer addresses model-tampering attacks (T5, T8, T9, T10) by securing a model throughout its entire lifecycle, from training, through deployment, up to updates. The control includes: Model signing with a cryptographic signature to verify that a model has not been altered in some way. Securing the update channel by encrypting OTA updates, and by validating the cryptographic signatures of the updates before applying them. Securing the model weights by storing them in a secure memory region. Runtime integrity check by validating the model parameters at a regular interval.

WEN and KATT (2024) list model tampering as one of the parameters that developers should consider for a system security assurance framework. The secure update channel has to be implemented on the agricultural tractor prior to the initial release of the product version built with AI. This is a precondition for OTA updates, which are expected to be part of autonomous agricultural tractors.

Layer 4 – Runtime monitoring and response

Runtime monitoring is intended to support the effectiveness of preventive measures applied in other layers. In this layer, the behavioral drift of the model is detected to handle possible slow attacks that lead to a reduction of the model performance over time due to a potential long-term attack of the adversary that tries to guess the input vector. Moreover, the integrity control of the output model is applied before being used by the vehicle control system in order to avoid executing the malicious outputs. Finally, the failsafe mechanism stops the vehicle to a conservative stop (Table 8).

Table 8: Defence-in-depth framework: mapping of TARA-prioritised threats to security controls across four layers

Security layer	Controls	Addressed threats	Priority
Input Validation	CAN message filtering, statistical anomaly detection, domain constraints	T1, T2, T3, T13, T14	High
Adversarial Robustness	Adversarial training, randomised smoothing, defensive distillation	T6, T7	High
Secure Model Lifecycle	Model signing, encrypted OTA, secure storage, integrity checking	T5, T8, T9, T10	High-Medium
Runtime Monitoring	Behavioural drift detection, output validation, fail-safe strategies	T11, T12, T15, T16	Medium

Note: Table 8 provides a consolidated mapping of prioritised TARA threats to the proposed framework controls, illustrating how each layer addresses specific identified risks. The remaining low-risk threats (T4, T17, T18) are accepted risks and are not assigned dedicated controls; T12 is covered by the output-validation controls of the runtime monitoring layer following its reclassification as a medium risk.

Discussion

Interpretation of experimental results

Our experiments demonstrate that black-box attacks can have a significant impact on the performance of a deep learning and statistical-based anomaly detection algorithms for agricultural tractors. We discussed why the sequential attacks had an impact on the LSTM model and possible reasons include the nature of RNNs and the fact that RNNs are widely used in anomaly detection tasks because they are able to learn temporal patterns in the data, which are also the characteristics that an adversary can exploit and learn through a surrogate model. Our results show that there is a decrease in detection accuracy of up to approximately 50% under the black-box attacks, occurring specifically for stealthy low-magnitude spoofing and in a scenario-dependent manner (Table 2), which can pose a significant threat to the operational safety of an ADS. An ADS that fails to detect real anomalies or floods the users with false alarms can cause severe operational disruptions.

Lightweight, interpretable, ML-based ADSs are known to be fast and efficient at identifying unusual activities or anomalies in the monitored data. However, we have shown that they can be compromised with a minimal amount of attack data, while still satisfying the required conditional statistical relationships. Therefore, the ease of use, their interpretability, and their advantages over respect to more complex models, do not imply at the same time the desired security, and they can fail to accomplish their task as ADS. Rather, the results confirm that although the simpler models evaluated here do not share the well-known vulnerabilities documented for complex models, including deep learning models, they are nonetheless open to new and unexpected attack scenarios of their own.

Implications for agricultural tractor AI security

The results and TARA findings have a number of implications for the design and deployment of AI in agricultural tractors. In particular, security should be viewed as a system property and not a model property, meaning that the robustness of a model does not imply the security of the data flows, update mechanisms and runtime in the system. The TARA results show that data flow and model lifecycle security threats are at least as important as model threats.

In the second part of the paper, we discuss how the agricultural tractor environment presents new challenges that have not yet been covered in the automotive AI security literature, such as legacy

CAN bus implementations without authentication, resource constraints on ECUs, and the isolated operation in fields. Our defence-in-depth approach is designed with these specific aspects in mind.

Third, combining the TARA with experimental assessment of system vulnerability provides a broader basis for security analysis than either approach alone. The TARA has covered many more potential threat scenarios than can be practically included in an experimental test (e.g., supply chain threats and OTA threats), while experimentation allows for a quantitative determination of threat likelihood and threat impact that can replace much of the analytical and qualitative reasoning that is inherent in the threat analysis process.

Limitations

The current implementation of this work has several limitations, outlined below. Although the black-box attacks were experimentally verified, it remains an open question whether the proposed countermeasures also protect against white-box attacks, grey-box attacks, or model tampering. The CAN bus attack and evaluation dataset used in this study is based on a specific tractor implementation and operating scenario and has yet to be validated for other tractor platforms and scenarios. In particular, the transferability of the results between different operating scenarios (e.g., field work versus road transport) remains to be examined.

The likelihood and impact ratings used in the TARA were assigned based on the consensus of a small group of experts and could be refined through quantitative risk assessment tools.

The compatibility of the proposed countermeasures with the resource constraints of embedded control units has not yet been validated: the experiments reported in this paper were conducted in an offline evaluation environment rather than on a physical ECU, and an experimental verification of the additional computational resources required for adversarial training and runtime protection on a real ECU implementation is still needed.

Conclusions

The security threats faced by an AI-based anomaly detection system for agricultural tractors are examined in this work, in one of the first dedicated studies on this topic. The main contributions and results of this work are summarised in the following. Attack taxonomy is proposed for white-box, black-box, grey-box and model tampering attacks. We investigate the possibility of each attack type on an agricultural tractor. Finally, we identify black-box attacks as the most relevant attack category for the CAN-bus-based operating environment considered in this paper, because an attacker in the agricultural field is realistically limited to observing input/output behaviour and lacks privileged access to the model's internal parameters or training data.

The experimental results show that the LSTM-based ADS and the lightweight ML-based ADS are highly vulnerable to the black-box attacks. The performance deviation of the LSTM-based ADS could reach up to 50% under the black-box attacks. The performance degradation of the lightweight ML-based ADS is scenario-dependent. Adversarial inputs designed to maintain the same statistical properties as the normal data can cause unexpected performance degradation.

TARA identified 18 threats on 5 assets with 3 high, 9 medium and 6 low risks for our LSTM-based ADS. The high-risk threats are mainly related to input data manipulation, model tampering, and adversarial attacks. Our defence framework is a direct result of the findings from this research.

The defence-in-depth strategy states that we should have four types of protections for our systems, based on the threats that were prioritised in the TARA: input validation, adversarial robustness, model lifecycle protection and runtime protection. Our goal is to implement defences that are deployable on resource constrained embedded systems, yet are sufficient to mitigate the most important threats. We emphasise that, whereas the vulnerability of the evaluated ADS implementations was demonstrated experimentally, this defence-in-depth framework is a proposal whose experimental validation on a physical ECU remains part of the future work.

In the coming experiments, we would like to investigate more attack scenarios, such as white-box, grey-box and model-tampering attacks. We would also like to examine the effects of different operating scenarios and tractor platforms. In addition, the behaviour of the defence control needs to be examined in relation to the resource constraints of the on-board computer imposed by the functional safety standards and a standard for benchmarking and comparing the security of various types of ADS and countermeasures needs to be proposed.

References

- Bhattacharya, S.; Gallolukankanamalage, R. G.; Steward, B. L.; Govindarasu, M. (2024): ML-based Anomaly Detection for CAN Bus Network in Agriculture Machinery. *Proceedings of the AAAI Symposium Series* 4(1), pp. 416-423, <https://doi.org/10.1609/aaais.v4i1.31826>
- Ding, D.; Zhang, M.; Feng, F.; Huang, Y.; Jiang, E.; Yang, M. (2023): Black-Box Adversarial Attack on Time Series Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 37(6), pp. 7358-7368, <https://doi.org/10.1609/aaai.v37i6.25896>
- Fenzl, F.; Rieke, R.; Dominik, A. (2021): In-vehicle detection of targeted CAN bus attacks. In: *The 16th International Conference on Availability, Reliability and Security (ARES)*, August 17–20, 2021, Vienna, Austria, ACM, New York, NY, USA 7 pages, <https://doi.org/10.1145/3465481.3465755>
- Ge, C.; Zhang, G.; Wang, Y.; Shao, D.; Song, X.; Wang, Z. (2025): Research status and development trends of artificial intelligence in smart agriculture. *Agriculture* 15(21), 2247, <https://doi.org/10.3390/agriculture15212247>
- Gkoutakos, K.; Ioannidis, K.; Demestichas, K.; Vrochidis, S.; Kompatsiaris, I. (2024): A Comprehensive Review of Deep Learning-Based Anomaly Detection Methods for Precision Agriculture, In: *IEEE Access*, vol. 12, pp. 197715–197733, <https://doi.org/10.1109/ACCESS.2024.3522248>
- Götz, K.; Kusuma, A.; Dörfler, A.; Lienkamp, M. (2025): Agricultural Load Cycles: Tractor Mission Profiles from Recorded GNSS and CAN Bus Data. *Data in Brief*, 111494. Dataset: Zenodo, <https://doi.org/10.5281/zenodo.14619787> (CC-BY-4.0)
- ISO/SAE (2021): ISO/SAE 21434:2021-08 Road vehicles – Cybersecurity engineering. International Organization for Standardization/SAE International, Geneva
- Kadhim, T.A.; Abduljabbar, Z.A.; AL-Asadi, H.A.A.; Nyangaresi, V.O.; Aldarwish, A.J.Y.; Ali, Z.A. (2025): Security Challenges in the Intelligent Precision Agriculture Environment: A Review. In: Silhavy, R., Silhavy, P. (eds) *Software Engineering: Emerging Trends and Practices in System Development, CSOC 2025. Lecture Notes in Networks and Systems*, vol 1560, Springer, Cham, https://doi.org/10.1007/978-3-032-00239-6_1
- Kishore, C.R.; Rao, D.C.; Behera, H.S. (2022): Deep Learning Approach for Anomaly Detection in CAN Bus Network: vol 480. Springer, Singapore, https://doi.org/10.1007/978-981-19-3089-8_51
- Krause, J.C.; Niemeyer, M.; Bajorath, J.; Iqbal, N.; Hertzberg, J. (2025): Towards Auto-Generated Ground Truth for Evaluation of Perception Systems in Agriculture. In: *Computer Vision - ECCV 2024 Workshops, Lecture Notes in Computer Science*, vol 15625, pp. 194-206. Springer, Cham, https://doi.org/10.1007/978-3-031-91835-3_13
- Lee, C.; Schätzle, S.; Lang, S.A.; Maier, M.; Oksanen, T. (2025): Fault Management System for the Safety of Perception Systems in Highly Automated Agricultural Machines, *IEEE International Conference on Robotics and Automation (ICRA)*, Atlanta, GA, USA, pp. 15546-15552, <https://doi.org/10.1109/ICRA55743.2025.11128195>

- Longari, S.; Cerracchio, P.; Carminati, M.; Zanero, S. (2025): Assessing the Resilience of Automotive Intrusion Detection Systems to Adversarial Manipulation. *ACM Transactions on Cyber-Physical Systems*.
<https://doi.org/10.1145/3737294>
- Muhammad, A.; Bae, S.-H. (2022): A Survey on Efficient Methods for Adversarial Robustness. In *IEEE Access*, vol. 10, p. 118815-118830, <https://doi.org/10.1109/ACCESS.2022.3216291>
- Rana, K.; Khatri, N. (2024): Automotive intelligence: Unleashing the potential of AI beyond advance driver assisting system, a comprehensive review. *Computers and Electrical Engineering* 117, <https://doi.org/10.1016/j.compeleceng.2024.109237>
- Singh, A.; Nawayseh, N.; Singh, H.; Dhahi, Y. K.; Samuel, S. (2023): Internet of agriculture: Analyzing and predicting tractor ride comfort through supervised machine learning. *Engineering Applications of Artificial Intelligence* 125, 106720, <https://doi.org/10.1016/j.engappai.2023.106720>
- Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. (2014): Intriguing Properties of Neural Networks. 2nd International Conference on Learning Representations (ICLR 2014), Banff, Canada, <https://doi.org/10.48550/arXiv.1312.6199>
- Vassilev, A.; Oprea, A.; Fordyce, A.; Anderson, H.; Davies, X.; Hamin, M. (2025): Adversarial Machine Learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2e2025). National Institute of Standards and Technology, <https://doi.org/10.6028/NIST.AI.100-2e2025>
- Wäschle, M.; Thaler, F.; Berres, A.; Pözlbauer, F.; Albers, A. (2022): A review on AI Safety in highly automated driving. *Front. Artif. Intell.* 5, <https://doi.org/10.3389/frai.2022.952773>
- Wen, S.; Katt, B. (2024): Exploring the role of assurance context in system security assurance evaluation: a conceptual model. *Information and Computer Security* 32(2), pp. 159–178, <https://doi.org/10.1108/ICS-06-2023-0101>
- Zhao, W.; Alwidian, S.; Mahmoud, Q. H. (2022): Adversarial Training Methods for Deep Learning: A Systematic Review. *Algorithms* 15(8), 283, <https://doi.org/10.3390/a15080283>
- Zhu, Y.; Zhang, S.; Tang, S.; Gao, Q. (2025): Research progress and applications of artificial intelligence in agricultural equipment. *Agriculture* 15(15), 1703, <https://doi.org/10.3390/agriculture15151703>

Authors

Rachid Batess, Energy and Agroequipment Department, Institute of Agronomy and Veterinary Hassan II, Rabat, Morocco and Aumovio GmbH, Guerickestraße 7, 60488 Frankfurt, Germany, e-mail: rachid.batess@aumovio.com

Younes El Fellah, Institute of Agronomy and Veterinary Hassan II, Rabat, Morocco

Ralf Leiter, StopDoc Training GmbH, Im Band 6b, 56743 Mendig, Germany

Notes

The CAN bus recordings used in this study are part of the openly available dataset of Götz et al. (2025), released under a CC-BY-4.0 licence at <https://doi.org/10.5281/zenodo.14619787>. The preprocessing, model-training, attack and evaluation scripts that reproduce all results and Figures 6–9 are provided as supplementary material accompanying this article and are additionally available from the corresponding author.

Acknowledgements

The authors thank the Institute of Agronomy and Veterinary Hassan II (Rabat, Morocco) and Aumovio GmbH (Frankfurt, Germany) for their institutional support of this research.