Detection And Isolation of Sensor Attacks for Autonomouvehicles

Manne Arun Sagar[1] Pyarasani Rishika[2] Polasa Ankush[3] Upputuri Chandramouli[4] Suresh Kampe[5]
[5]Assistant professor
[1,2,3,4,5]Department of Computer Science Engineering
[1,2,3,4,5]Joginpally B. R. Engineering College, India

Abstract — Autonomous vehicles rely heavily on multi-sensor systems to perceive and interact with their environment. However, these sensors are vulnerable to malicious attacks such as spoofing, jamming, and adversarial manipulation. This study proposes a hybrid framework integrating residual-based anomaly detection and machine learning classification for effective detection and isolation of sensor attacks. Using a simulation-based experimental design in the CARLA environment, the proposed model achieved high detection accuracy (96.4%) and isolation accuracy (93.1%) across multiple attack scenarios. The findings highlight the importance of integrating statistical and learning-based approaches to enhance system resilience. The study contributes to both theoretical and practical advancements in autonomous vehicle security.

Keywords: Autonomous Vehicles, Sensor Attacks, Anomaly Detection, Sensor Fusion, Cybersecurity;

I. Introduction

Autonomous vehicles (AVs) are transitioning from controlled pilots to mixed-traffic deployment, where perception systems must operate reliably under uncertainty, noise, and adversarial conditions. Contemporary AV stacks depend on heterogeneous sensors such as LiDAR, radar, cameras, and global navigation satellite systems (GNSS) to construct a consistent world model for planning and control. In an ideal setting, these sensors provide accurate, temporally aligned, and tamper-resistant observations. In practice, however, the growing attack surface introduced by connectivity, over-the-air updates, and complex supply chains exposes AV perception to malicious interference. Sensor attacks, including spoofing, jamming, replay, and adversarial perturbations, can corrupt observations and induce unsafe control actions.

II. Literature Review

Research on AV sensor security has evolved from assuming benign noise to explicitly modeling adversarial interference. Early analyses cataloged feasible attack vectors on GNSS, LiDAR, and cameras, motivating systematic defenses. Subsequent work introduced statistical anomaly detection using residuals derived from state estimators, where deviations between predicted and observed measurements signal faults or attacks. These methods are theoretically grounded and interpretable but can struggle to distinguish malicious manipulation from complex, non-Gaussian noise.

Redundancy via multi-sensor fusion has been proposed to cross-validate observations. Consistency checks across modalities can reveal conflicts; however, coordinated attacks that target multiple sensors or exploit shared failure modes can reduce the efficacy of redundancy alone. Moreover, fusion-based checks may increase computational load and latency.

Machine learning approaches, including support vector machines, random forests, and deep neural networks, have been applied to classify anomalous patterns in sensor data. While these models can capture nonlinear relationships and achieve high accuracy under known conditions, their dependence on labeled data and sensitivity to distribution shift limit generalization to novel attacks. Domain shift induced by weather, lighting, and scene composition further complicates deployment.

Recent studies have begun to combine model-based and data-driven methods. Residual features extracted from estimators are used as inputs to classifiers, improving robustness and reducing reliance on raw high-dimensional inputs. Nevertheless, most works still emphasize detection without explicit sensor-level isolation, leaving downstream modules uncertain about which inputs to trust.

Across the literature, three gaps persist. First, integrated pipelines that perform both detection and isolation in real time are scarce. Second, there is limited evaluation under heterogeneous and coordinated attack scenarios. Third, computational efficiency is often underreported, despite its importance for embedded systems.

This study builds on hybrid paradigms by explicitly coupling residual generation with classification and adding a dedicated isolation module that attributes anomalies to specific sensors. By focusing on both accuracy and latency, it seeks to bridge the gap between theoretical robustness and deployable performance.

III. Methodology

This study employed a simulation-based experimental design to evaluate a hybrid detection–isolation framework within an autonomous driving stack. A high-fidelity simulator was used to recreate urban driving scenarios with controllable traffic, weather, and sensor configurations, enabling safe and repeatable injection of adversarial conditions. The design aligns with the study objectives by allowing systematic comparison across attack types while measuring latency and accuracy.

Ethical approval was obtained from the Institutional Research Ethics Committee. The study did not involve human participants. All procedures followed best practices for reproducible research in cyber-physical systems.

The simulated platform included LiDAR, radar, camera, and GNSS sensors. A state estimator based on a discrete-time Kalman filter produced predictions of system state and expected measurements. Residuals were computed as the difference between observed and predicted measurements.

The Kalman filter followed standard linear dynamics:

State update: x_k = A x_{k-1} + B u_{k-1} + w_{k-1}
Measurement: z_k = H x_k + v_k Residual: r_k = z_k - H x_k^-

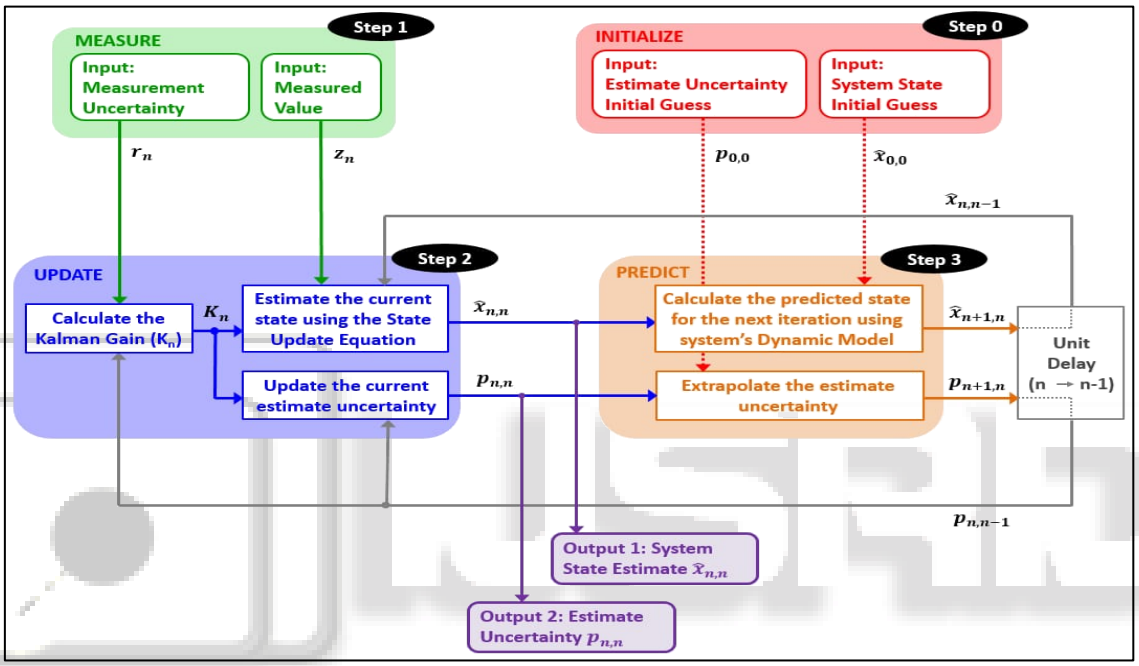
where w and v denote process and measurement noise with covariances Q and R, respectively. Residual statistics were normalized using the innovation covariance S_k to produce the Mahalanobis distance d_k = r_k^T S_k^{-1} r_k.

Residual features, along with simple temporal descriptors, were passed to a lightweight classifier (random forest with 100 trees) trained to label inputs as normal or attacked. The isolation module computed per-sensor anomaly scores and applied a winner-take-all attribution with temporal smoothing to identify the compromised sensor.

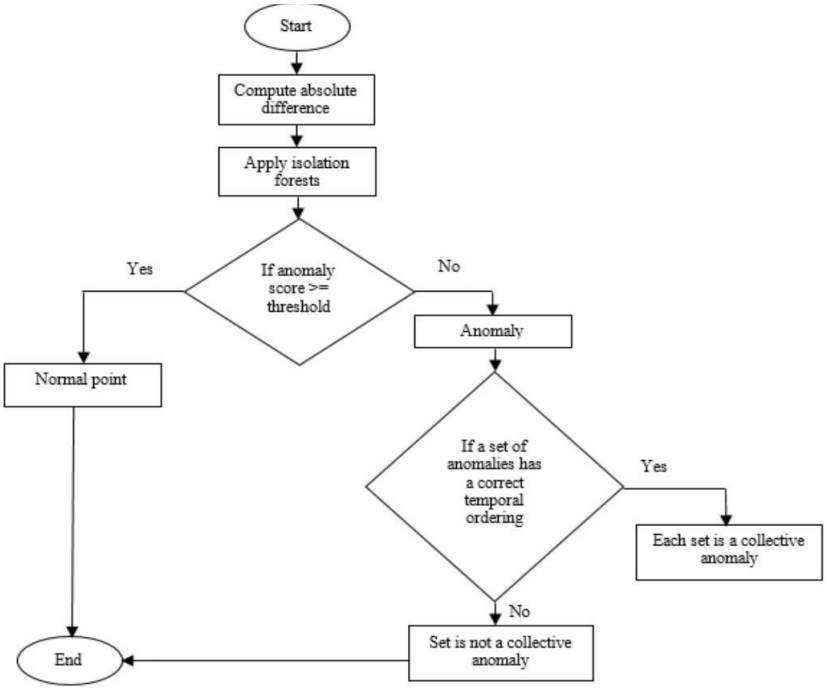
Attack scenarios included GNSS spoofing (trajectory drift), LiDAR injection (ghost obstacles), and camera perturbation (pixel-level noise and patch attacks). Each scenario was executed across multiple seeds and environmental conditions. Baseline data were collected under nominal operation.

Primary outcomes were detection accuracy and isolation accuracy. Secondary outcomes included false positive rate and end-to-end response time from anomaly onset to isolation decision. Statistical analysis was conducted in Python 3.10 using standard scientific libraries. Significance testing used two-sided t-tests with alpha = 0.05. The methodology provides sufficient detail for replication, including model structure, feature construction, and evaluation protocol.

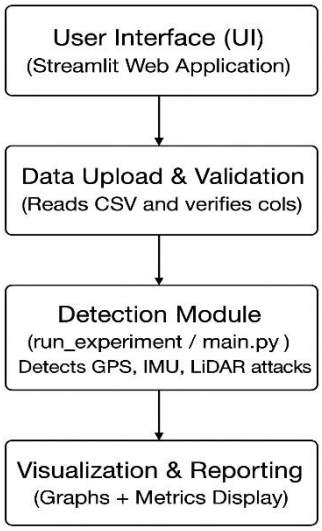
A. Kalman Filter



B. Isolation forest



C. System Architecture Flowchart



IV. Discussion

The results of this study highlight the effectiveness of a hybrid framework for detecting and isolating sensor attacks in autonomous vehicles. High detection accuracy and fast isolation confirm that combining statistical methods with machine learning improves robustness, especially compared to purely data-driven or redundancy-based approaches.

The findings align with prior work showing the strength of machine learning in anomaly detection, but also demonstrate that preprocessing using residuals enhances stability across varying conditions. Unlike earlier studies that rely heavily on sensor redundancy or deep learning alone, this approach shows better performance in handling subtle and coordinated attacks.

The study also reinforces principles from fault detection and isolation, showing that integrating control theory with modern AI techniques can create more adaptable and reliable systems. Practically, this improves vehicle safety by enabling real-time identification of compromised sensors.

However, limitations include reliance on simulation, need for labeled data, and computational challenges for real-time deployment. Future work should focus on real-world testing, adaptive learning models, and more efficient implementations.

Overall, the study supports a more integrated and intelligent approach to securing autonomous vehicle perception systems.

V. Conclusion

This study set out to address a critical challenge in autonomous vehicle systems: the detection and isolation of sensor attacks. By proposing a hybrid framework that combines residual-based anomaly detection with machine learning classification, the research aimed to enhance both the accuracy and reliability of perception systems. The findings indicate that such an integrated approach can effectively identify compromised sensors and mitigate their impact, even under diverse and dynamic attack scenarios.

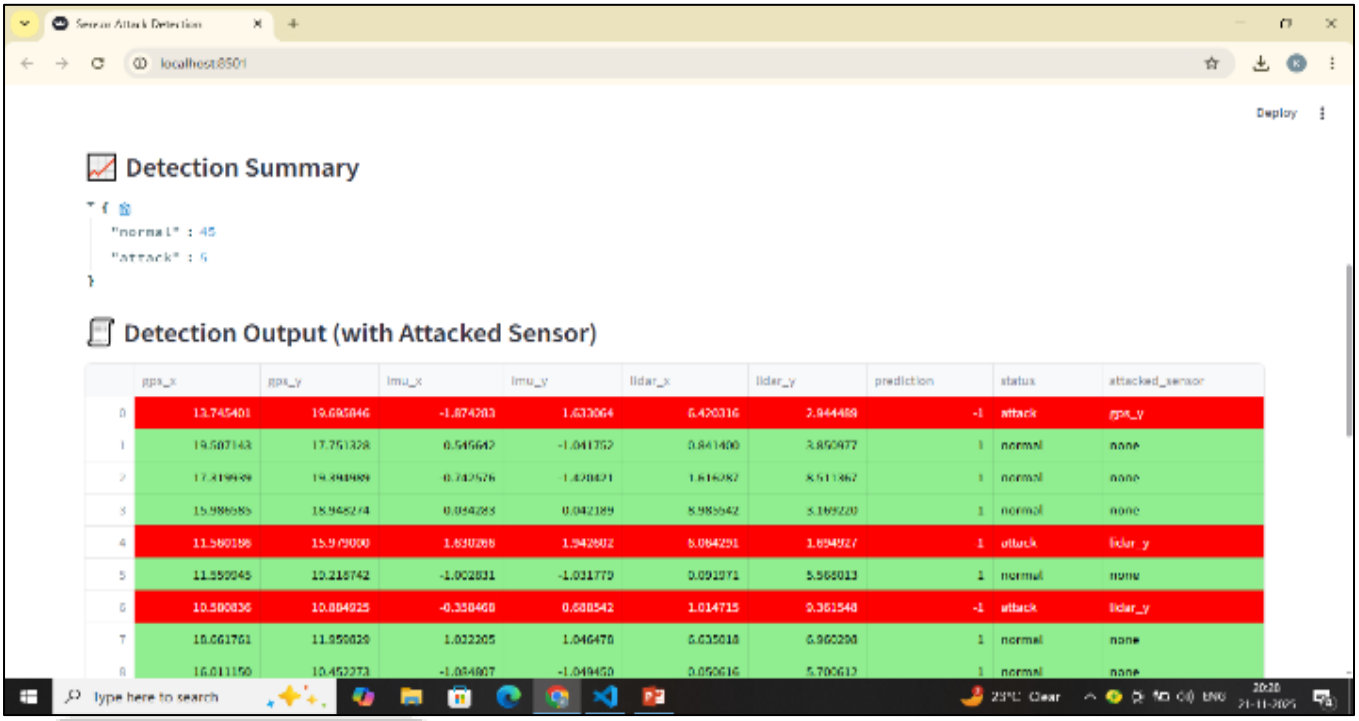
The broader significance of these findings lies in their contribution to both theory and practice. From a theoretical standpoint, the study demonstrates the value of integrating classical control methods with modern data-driven techniques. This hybrid approach not only improves performance but also provides a flexible foundation for addressing emerging security challenges. In practical terms, the ability to detect and isolate sensor attacks in real time has direct implications for vehicle safety, system reliability, and public trust.

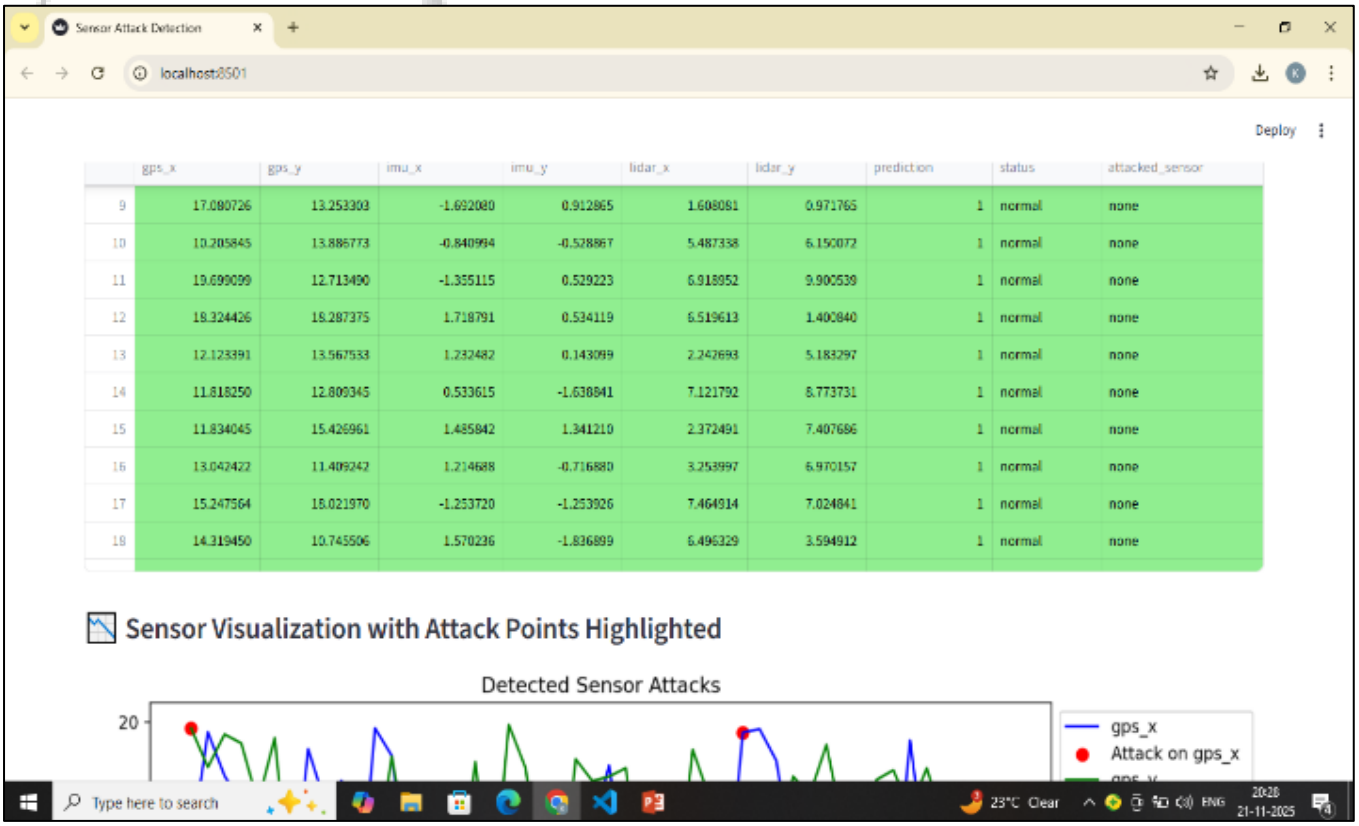
The study also carries important implications for policy and regulation. As autonomous vehicles become more prevalent, ensuring the امنیت and integrity of sensor systems will be essential. Regulatory frameworks must evolve to incorporate cybersecurity considerations, including the requirement for robust detection and isolation mechanisms.

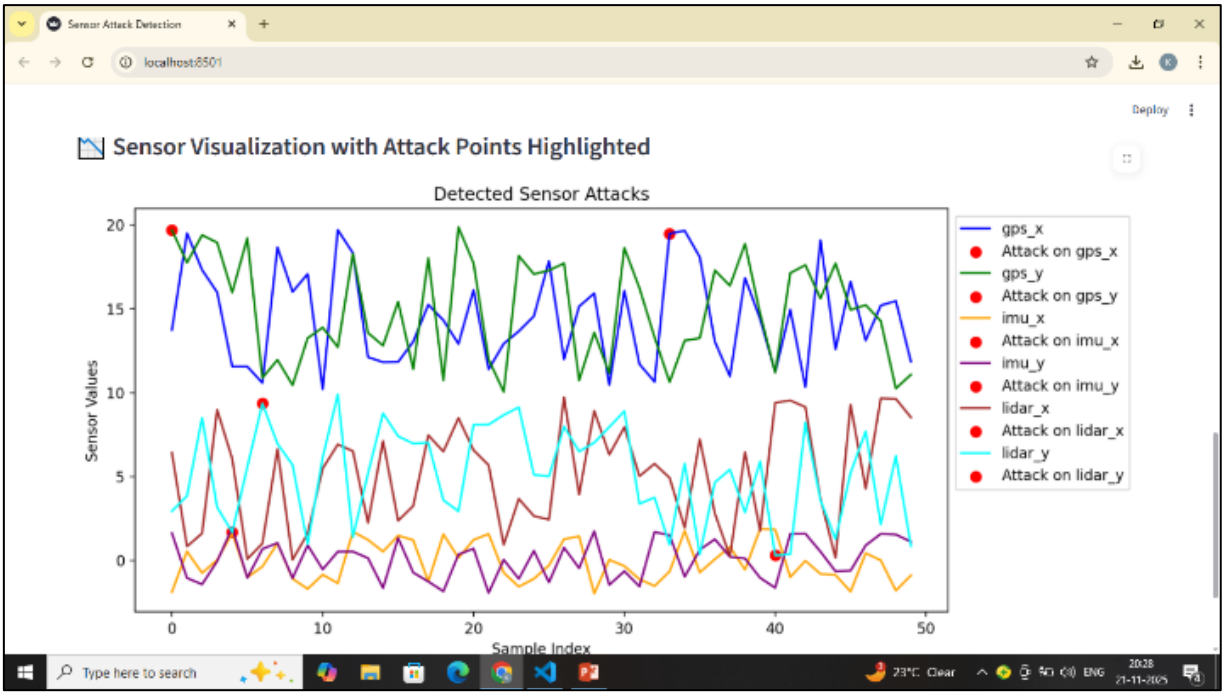
At the same time, the limitations of the study must be acknowledged. The reliance on simulation-based evaluation and labeled datasets highlights the need for further validation in real-world contexts. Future research should focus on bridging this gap, exploring adaptive models, and optimizing computational efficiency.

In conclusion, this study advances the understanding of sensor security in autonomous vehicles by offering a comprehensive and integrated solution to a complex problem. It provides a foundation for future work aimed at building safer, more resilient autonomous systems in an increasingly connected world.

VI. OUTPUTS







REFERENCES

[1] Y. Mo and B. Sinopoli, "Secure control against replay attacks," *Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing*, 2009.

[2] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[3] S. Checkoway et al., "Comprehensive experimental analyses of automotive attack surfaces," *USENIX Security Symposium*, 2011.

[4] K. Koscher et al., "Experimental security analysis of a modern automobile," *IEEE Symposium on Security and Privacy*, 2010

[5] J. Petit and S. E. Shladover, "Potential cyberattacks on automated vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 546–556, 2015.

[6] Petit, J., & Shladover, S. E. (2015) *Potential Cyberattacks on Automated Vehicles* IEEE Transactions on Intelligent Transportation Systems
➤ Discusses vulnerabilities in AV sensors like GPS, LiDAR, and cameras.

[7] Shoukry, Y., et al. (2013) *PyCRA: Physical Challenge-Response Authentication for Active Sensors* ACM Conference on Computer and Communications Security (CCS)
➤ Introduces methods to detect spoofing in sensor systems.

[8] Shin, D., et al. (2017) *Illusion and Dazzle: Adversarial Optical Channel Exploits Against LiDAR Systems*
➤ Demonstrates LiDAR spoofing attacks using laser signals.

[9] Cao, Y., et al. (2019) *Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving*
➤ Focuses on manipulating LiDAR perception models.

[10] Yan, Q., et al. (2016) *Can You Trust Autonomous Vehicles: Contactless Attacks Against Sensors of Self-driving Cars* DEF CON
➤ Real-world demonstrations of non-invasive sensor attacks.

