International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Data Poison Detection in Automation Vehicles

CH. China Subba Reddy[1], K.Jhansi[2], M.Raghu varna[3], S.Praveen[4],D.Vazeer Hussain[5]

[1]CH.China Subba Reddy CSE & Joginpally B.R. Engineering College
[2]K.Jhansi CSE & Joginpally B.R. Engineering College
[3]M.Raghu varna CSE & Joginpally B.R. Engineering College ge
[4]S.Praveen CSE & Joginpally B.R. Engineering College
[5]D.Vazeer Hussain CSE & Joginpally B.R. Engineering College

---

Abstract - The rise of autonomous vehicles (AVs) has transformed modern transportation, offering enhanced safety, efficiency, and convenience. However, these intelligent systems rely heavily on machine learning models trained on vast datasets, which makes them vulnerable to data poisoning attacks. Data poisoning is a form of adversarial attack where malicious data is injected into the training set to manipulate the behavior of a model, potentially leading to catastrophic consequences in real-world applications.

This project, titled "Data Poison Detection in Autonomous Vehicles", aims to develop a reliable method to identify and mitigate poisoned data during the training phase of autonomous driving models. Using the German Traffic Sign Recognition Benchmark (GTSRB) dataset, we simulate data poisoning by injecting adversarial patterns into a subset of training images. We then apply feature extraction techniques, including Histogram of Oriented Gradients (HOG) and color histograms, to transform the image data into numerical feature vectors.

A Support Vector Machine (SVM) classifier, integrated with a standardization pipeline, is trained to differentiate between clean and poisoned data. The model is evaluated using accuracy, confusion matrix, and classification reports. The proposed system successfully detects poisoned data with high precision, indicating its potential for real-world deployment in AV training pipelines.

Our results highlight the importance of pre-training data validation and propose an effective approach to enhance the robustness of autonomous vehicle systems against poisoning attacks. Future improvements could include real-time detection during data collection and integrating deep learningbased anomaly detection techniques.

1. INTRODUCTION

Autonomous vehicles (AVs) are at the forefront of the technological revolution, combining sensors, machine learning algorithms, and high-performance computing systems to perform complex driving tasks with minimal or no human intervention. These intelligent vehicles are designed to perceive their surroundings, make real-time decisions, and navigate roads safely and efficiently. A crucial component of this intelligence is the underlying machine learning model, which must be trained on massive datasets to accurately recognize road signs, pedestrians, other vehicles, and potential hazards.

While the reliance on machine learning brings about tremendous capabilities, it also introduces significant vulnerabilities, particularly in the area of data integrity. One of the most insidious threats to machine learning models is **data poisoning** — a form of adversarial attack where an attacker intentionally injects misleading or malicious data into the training dataset. When a model is trained on such corrupted data, it learns incorrect patterns or associations, leading to faulty behavior during real-world operation.

For autonomous vehicles, the consequences of such behavior could be disastrous. Imagine a scenario where a vehicle fails to recognize a stop sign because the model was trained on poisoned data that labeled stop signs as speed limit signs. Such errors can cause serious accidents, loss of life, and legal consequences. Therefore, ensuring the integrity and quality of training data is not just a matter of performance but of safety and trust.

This project aims to address this pressing challenge by developing a detection mechanism for poisoned data. Using a combination of image processing techniques and machine learning classifiers, the system is designed to identify and flag poisoned samples within traffic sign recognition datasets, ensuring that only clean data is used for training critical AV models.

2. Body of Paper

2.1 Dataset Description

In this study, the German Traffic Sign Recognition Benchmark (GTSRB) dataset was used for training and evaluation. GTSRB consists of over 50,000 images across 43 different traffic sign classes. The dataset includes real-world variability such as lighting conditions, occlusions, and different perspectives, making it suitable for simulating a realworld autonomous driving environment.

2.2 Data Poisoning Strategy

A targeted backdoor poisoning attack was simulated to inject poisoned samples into the training dataset. As described in Sec. 2.1, images from all classes were modified by inserting a distinct pattern—a white square in the top-left corner and a black square in the center of the image. These adversarial

patches were added to 20% of the images (poison rate = 0.2).

The class label of each poisoned image was changed to a specific target class (class ID 14). This method aims to mislead the model into misclassifying any image containing this pattern as the target class.

2.3 Feature Extraction

Each image underwent a two-stage feature extraction process to obtain both shape and color information:

- First, **Histogram of Oriented Gradients (HOG)** features were extracted to capture edge and shape patterns by analyzing gradients in grayscale images.
- Second, **color histograms in HSV (Hue, Saturation, Value)** color space were calculated to provide information about image color distribution. The final feature vector was a concatenation of both HOG and histogram features, resulting in a highdimensional descriptor vector as illustrated in Fig. 1.

2.4 Classification Model

The extracted features were used to train a **Support Vector Machine (SVM)** classifier. The model utilized a radial basis function (RBF) kernel with class balancing enabled to address any imbalance between clean and poisoned samples. The data was split into training (80%) and testing (20%) sets using train_test_split. The classification task was binary: to detect whether a sample is poisoned (1) or clean (0).

2.5 Evaluation Metrics

Model performance was evaluated using standard metrics: **precision, recall, F1-score, and accuracy**, as shown in Tab. 1. The **confusion matrix** provided insight into the true positive and false negative rates of poisoned sample detection. Section 2.6 provides a visual overview of some correctly classified clean and poisoned images.

2.6 Visualization and Results

Figure 2 shows a visual comparison of correctly classified poisoned and clean samples. The top row shows samples where the SVM correctly identified the poisoned images, and the bottom row contains correctly identified clean samples. These visualizations highlight the model's ability to identify patterns that distinguish adversarial samples from legitimate ones.

3. CONCLUSIONS

In this study, we proposed and implemented an effective approach to detect data poisoning attacks in the context of autonomous vehicle perception systems. By leveraging handcrafted feature extraction techniques such as Histogram of Oriented Gradients (HOG) and HSV color histograms, combined with a Support Vector Machine (SVM) classifier, we were able to identify poisoned images with high reliability.

The experimental results demonstrated that the model effectively distinguished between clean and tampered data, even when adversarial perturbations were subtle. This highlights the significance of robust pre-processing and feature extraction in enhancing the security of machine learning pipelines in autonomous driving scenarios.

Our framework can serve as a foundational step toward building more resilient perception systems against trainingtime attacks. Future work may include extending the model to deep learning architectures, exploring real-time detection capabilities, and testing the robustness of the system under different poisoning strategies and noise conditions.

The findings presented in this paper reinforce the urgent need for integrating security-aware components into the machine learning lifecycle of autonomous vehicles, thereby promoting safer deployment of intelligent transportation systems.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the faculty and staff of the Computer Science and Engineering, Joginpally BR Engineering College, for their continuous support and guidance throughout this project. Special thanks are extended to our project supervisor, CH.China Subba Reddy, for their insightful feedback, technical advice, and encouragement during the research and development phase.

We also acknowledge the use of the German Traffic Sign Recognition Benchmark (GTSRB) dataset, which served as a valuable resource for implementing and testing the proposed model. Additionally, we thank the open-source Python libraries and tools that enabled rapid prototyping and evaluation.

REFERENCES

1. **Gu, T., Dolan-Gavitt, B., & Garg, S. (2017).** *BadNets: Identifying vulnerabilities in the machine learning model supply chain.arXiv preprint arXiv:1708.06733.*
2. **Steinhardt, J., Koh, P. W., & Liang, P. (2017).** *Certified defenses for data poisoning attacks. Advances in Neural Information Processing Systems.*
3. **Bhagoji, A. N., Chakraborty, S., Mittal, P., & Calo, S. (2018).** *Analyzing federated learning through an adversariallens. International Conference on Machine Learning (ICML).*
4. **Szegedy, C., Zaremba, W., Sutskever, I., et al. (2014).** *Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.*
5. **Dalal, N., & Triggs, B. (2005).** *Histograms of Oriented Gradients for Human Detection. IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR).
[HOG feature extraction technique used in your project.]

6. **Ghosh, A., et al. (2021).** *Security of Machine Learning in Autonomous Vehicles Springer Nature.*
7. **Vapnik, V. (1998).** *Statistical Learning Theory.*
8. **Two Minute Papers**
   Channel: https://www.youtube.com/@TwoMinutePapers
   [Explains cutting-edge AI research in simple language.]
9. **Sentdex**
   Channel: https://www.youtube.com/user/sentdex [Hands-on tutorials on machine learning, OpenCV, SVM, and more.]
10. **StatQuest with Josh Starmer**
    Channel: https://www.youtube.com/user/joshstarmer
    [Excellent explanations on statistical learning and SVM.]