ML-based Anomaly Detection for CAN Bus Network in Agriculture Machinery

Souradeep Bhattacharya[1], **Ranuka G. Gallolukankanamalage**[2], **Brian L. Steward**[2], **Manimaran Govindarasu**[1]

[1]Department of Electrical and Computer Engineering,
[2] Department of Agricultural and Biosystems Engineering
Iowa State University, Ames, IA 50011 USA
sbhatta@iastate.edu, ranuka15@iastate.edu, bsteward@iastate.edu, gmani@iastate.edu

Abstract

The adoption of advanced automation and next-generation technologies like the Internet of Things (IoT) and modern communication networks has revolutionized the food and agriculture sector, boosting the efficiency and precision of farm machinery. However, this increased inter-connectivity has also exposed significant vulnerabilities, particularly in Controller Area Network (CAN) protocols, widely used in advanced agricultural machinery and equipment. Due to its lack of inherent security features, CAN is susceptible to various cyber-attacks, potentially leading to severe consequences if these attacks remain undetected and unmitigated. This paper introduces a supervised machine learning (ML)-based anomaly detection system (CAN-ADS) designed to detect various cyber-attacks on CAN-based agricultural machinery. The system leverages network traffic augmentation and data balancing techniques to train ML algorithms on CAN-specific datasets. Experimental results show that CAN-ADS achieves high accuracy ($\approx 98\%$) and true-positive rates with low false-negative rates ($\approx 1\%$).

Introduction

The UN's Global Food and Agriculture Report highlights that by 2050, global food demand is expected to rise by 70% to meet the needs of an estimated 10.1 billion people, with 26.7% of the global population relying on agriculture for their livelihoods (Mekouar 2018). To address this demand, traditional farming practices are rapidly evolving with the integration of disruptive technologies like IoT, robotics, cloud, 5G, and artificial intelligence (AI) (Yazdinejad et al. 2021). This shift, known as Agriculture 5.0, is transforming conventional farming into Smart Farming and Precision Agriculture (Rudrakar and Rughani 2023). These systems utilize IoT sensors, advanced machinery, and automation for real-time decision-making and control. However, increasing digitization has introduced new vulnerabilities, expanding the cyber attack surface and posing significant risks to agricultural operations (Ferrag et al. 2022). Therefore, securing this infrastructure is of vital importance.

Modern agricultural machinery, such as tractors and combines, use the CAN communication protocol to connect multiple Electronic Control Units (ECUs) and exchange sensor and control data (Rohrer, Pitla, and Luck 2019). While CAN Bus systems provide essential electronic connectivity and high-precision performance data, they lack built-in security features, making them vulnerable to attacks that can disrupt ECUs and compromise safety (Bozdal et al. 2020). These risks highlight the urgent need for enhanced security in CANbus-based machinery.

Intrusion detection systems (IDS) can enhance network security by identifying potential threats using rule-based and anomaly-based techniques (Abdelkhalek and Govindarasu 2022). Rule-based IDS offers faster detection by relying on predefined attack signatures but is limited to known threats. Anomaly-based IDS profiles network traffic for deviations, making it more adaptable, but often struggles with identifying stealthy attacks that mimic normal behavior. These systems also suffer from false positives and negatives due to imbalanced datasets (Abdelkhalek and Govindarasu 2022). Data augmentation techniques can address these challenges by improving the robustness and accuracy of anomaly-based IDS (Hasibi, Shokri, and Fooladi 2019).

Considering the critical nature of CAN within the domain of precision agriculture and smart farming, the primary contribution of this paper is to develop a robust anomaly detection system (ADS) using supervised ML techniques tailored for CAN communication networks in agricultural machinery. The proposed method achieves the desired objective through the following:

- Generation of CAN-specific IDS datasets with realistic cyber-attacks for training ML models, using data augmentation techniques.
- Extraction of CAN-specific ML features from network parameters and implementation of data processing techniques to improve detection accuracy.
- Development of time-series data-compatible ML algorithms for accurate detection of known and stealthy attacks.
- Comparative analysis of performance of developed ML models using baseline and augmented datasets to demonstrate the efficacy of proposed solution.

Background on Smart Agriculture

A typical multi-layered high-level architecture of the smart agriculture infrastructure is shown in Figure 1. As depicted

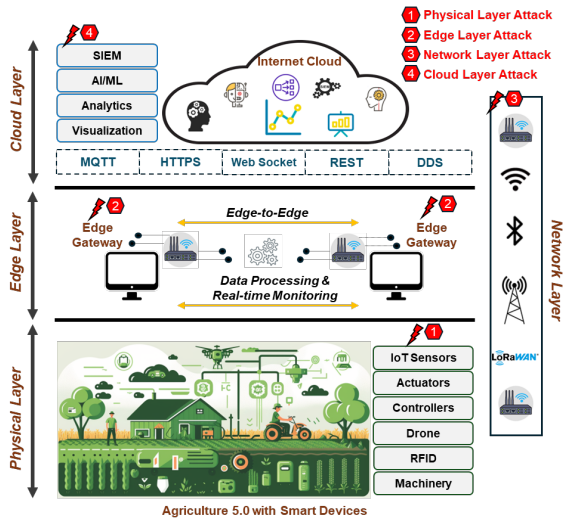


Figure 1: Overview of smart agriculture infrastructure with possible intrusion paths

in the diagram, the infrastructure can be divided into four interconnected layers (Yazdinejad et al. 2021):

- Physical Layer: This layer includes various smart devices such as IoT sensors, actuators, controllers, drones, RFID, and farm machinery that interact directly with the agricultural environment. At this level, data is collected and control actions are performed.
- Edge Layer: At this layer, edge gateways are responsible for processing and filtering data collected from the physical layer before it is transmitted to the higher layers. The edge layer ensures efficient data handling close to the source, reducing latency and bandwidth usage.
- Network Layer: This layer encompasses the communication protocols used to transmit data between the edge gateways and the cloud. It includes various wireless technologies like LoRaWAN, Bluetooth, and other connectivity options essential for seamless data flow in the agricultural network.
- Cloud Layer: This layer handles more complex data processing tasks such as AI/ML analytics, visualization, and Security Information and Event Management (SIEM). It is also where data is stored, analyzed, and used to generate actionable insights for improving farming operations.

Cybersecurity Challenges in Smart Agriculture

The U.S. Food and Agriculture sector has become an increasingly attractive target for malicious cyber-attacks as the sector adopts smart farming and precision agriculture technologies (Kulkarni et al. 2024). Unlike the other 16 critical infrastructure sectors in the US, agriculture has historically been less connected to digital infrastructure and slower to adopt cybersecurity measures (Obama 2013). However, the growth of smart agriculture is driving the integration of digital technologies, autonomous systems, advanced machinery, protocols, and data-driven solutions to enhance agricultural processes, boost product quality and quantity, optimize resource utilization, and improve the efficiency, sustainability, and productivity of farming practices. Despite the many benefits of smart agriculture, this new paradigm is vulnerable to advanced threats that can have severe consequences for farmers and agricultural enterprises. These risks include financial losses, identity theft, service disruptions, reputational damage, reduced crop yield and quality, supply chain disruptions, data breaches, privacy concerns, environmental harm, food safety risks, and a loss of trust and confidence (Ali et al. 2024). Given this concern, the U.S. Department of Homeland Security (DHS 2018) has identified three primary categories of cyber threats in smart farming and precision agriculture as follows:

- *Confidentiality-related:* In a smart agriculture ecosystem, data flows through multiple interconnected communication devices from source to destination. This creates privacy threats that can result in data breaches and the loss of sensitive information.
- *Integrity-related:* Unauthorized or inappropriate alterations to data or resources can compromise the reliability and accuracy of the information.
- *Availability-related:* Failure to maintain service availability can lead to business disruptions, resulting in potential loss of customer trust and revenue.

CAN Communication in Agricultural Machinery

CAN is a fast, real-time serial bus designed to provide a low-cost, centralized, robust, and flexible communication interface between sensors and actuators within vehicles (Nymann 2021). Originally developed by Robert Bosch in 1986 and later codified into the ISO 11898-1 standard, CAN bus consists of two electrical wires (CAN Low and CAN High) which are used to transmit information between Electronic Control Units (ECUs). The ISO 11898-1 standard covers the data link layer, including baud rate and cable length, while ISO 11898-2 describes the physical layer, detailing cable types, node requirements, and electrical signal levels (S. Corrigan 2016). CAN frames are used to communicate over CAN bus. CAN uses the differential signal (2.5V) with two logic states - dominant (0 bit, 3.5V) and recessive (1 bit, 1V). CAN network uses two CAN messages - *standard CAN* (11-bit identifier) and *extended CAN* (29-bit identifier), as shown in Figure 2.

In modern agricultural machinery, CANbus serves as the backbone for communication between multiple ECUs, as shown in Figure 2. These ECUs are responsible for controlling different functions within the machinery, such as engine management, transmission control, and the operation of various attachments like seeders, sprayers, and harvesters (Rohrer, Pitla, and Luck 2019). The CANbus protocol enables these ECUs to share data and coordinate actions, ensuring that the machinery operates smoothly and efficiently. One of the key advantages of CANbus in agricultural machinery is its ability to handle high volumes of sensor and control data with minimal latency. This is crucial in precision agriculture, where real-time data from sensors (such as those monitoring soil conditions, crop health, and machin-

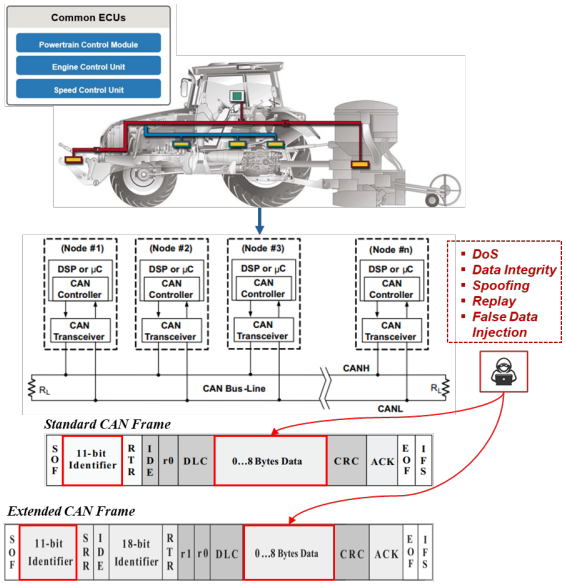


Figure 2: Application of CAN in modern agricultural machinery with sample attack vectors

ery performance) is used to make immediate adjustments to machinery settings. Additionally, CANbus supports the integration of advanced technologies such as GPS systems, automated steering, and telematics, which are increasingly common in modern farming (Boland et al. 2021). However, with the increasing reliance on CAN, cybersecurity challenges are introduced due to its limited security features. This makes the protocol vulnerable to attacks that could disrupt agricultural operations. Table 1 shows the different CAN protocols used in agricultural automotive machinery.

Cyber-attack Vectors in Agricultural Machinery

Vulnerabilities within CAN-based agricultural machinery can be exploited to launch various cyber-attacks, potentially compromising critical assets and operations (Bozdal et al. 2020). Some possible attacks on CAN systems and their potential targets in smart agriculture are identified below:

- *Possible Attack Vectors:* Malware Injection, Ransomware, Botnet, Social Engineering, Phishing, Denial of Service (DoS) and Distributed DoS (DDoS), Man-in-the-middle (MITM), Replay, Eavesdropping, Insider Threat, Side-channel Intrusion, Radio Frequency Jamming, Rogue Device Deployment, Spoofing, False Data Injection (FDI), Reconnaissance, SQL injection.
- *Possible Targets:* CAN Communication, Electronic Control Units (ECUs), IoT Sensors, Controller Nodes, OBD Port, Multimedia, Remote Interfaces (RF, Bluetooth, WiFi, VANET), GPS.

Properties	CAN 2.0	CAN FD	SAE J1939
Year	1991	2011	1994
Standard	ISO 11898	ISO 11898	SAE J1708 / J1587
Devices	Upto 40	Upto 70	Upto 256
Data Rate	Upto 1 Mbps	Upto 8 Mbps	Upto 1 Mbps
DLC Rate	Fixed 8 bytes	Flexible 0-64 bytes	Flexible 0-64 bytes
Application	Light Vehicles	Light Vehicles	Heavy Vehicles

Table 1: Types of CAN Protocols in Agricultural Machinery

Related Work

The wide adoption of smart agriculture practices has introduced new security concerns and extended the attack surface. Previous research has discussed the existing vulnerabilities, attacks, threats, and security considerations for smart farming and precision agriculture infrastructure (Yazdinejad et al. 2021; Rudrakar and Rughani 2023). Threat modeling techniques have been developed to identify possible attack paths within the Agriculture-IoT (AG-IoT)-based networks and devices. In (Asif et al. 2021), the authors proposed a STRIDE-based threat modeling framework tailored for precision agriculture. Risk assessment and mitigation frameworks related to this domain have also been developed. In (Lieder and Schröter-Schlaack 2021), the authors provide a holistic risk assessment pertaining to smart farming technologies, and in (Raghuvanshi et al. 2022), an ML-based risk mitigation framework is proposed. Existing literature has also discussed several strategies for detecting malicious behavior within smart farming and precision agriculture infrastructure, such as IDS and ADS. In (Catalano et al. 2022), a multivariate linear regression (MLR) and long-term memory neural network algorithm (LSTM)-based ADS has been developed and tested on time-series farm sensor data. In (Kethineni and Gera 2023), a deep learning-based intrusion detection technique is proposed focusing on data privacy-preserving applications. An autoencoder-based anomaly detection system for smart farming is proposed in (Adkisson et al. 2021). Additionally, existing research has also focused on the cybersecurity of CAN communication. In (Othmane et al. 2022), a methodology to detect fabricated CAN messages is proposed. In (Purohit and Govindarasu 2022), a hybrid ADS utilizing rule-based and ML-based methods is proposed for CANbus.

However, these approaches often face challenges due to imbalanced and inadequate datasets, leading to high false positive and negative rates. To address these limitations, this paper leverages data augmentation and standardized data processing techniques to improve the performance of ML models in detecting anomalies. Unlike traditional methods that concentrate on refining detection algorithms, this study focuses on enhancing the quality and diversity of training data, thereby significantly improving the accuracy and reliability of ADS in smart farming and precision agriculture.

Proposed CAN-ADS Framework

The proposed methodology is illustrated in Figure 3. The framework can be divided into three main stages: (i) *Generation of CAN-specific IDS Datasets* - comprehensive datasets

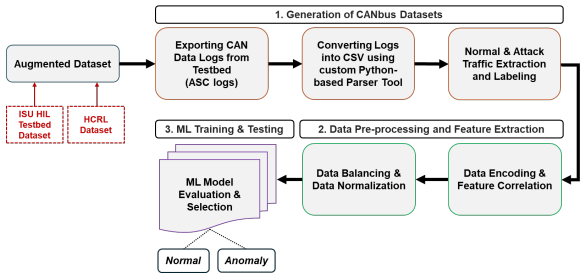


Figure 3: Proposed ML-based anomaly detection for CAN communication in agricultural machinery

are created, including normal and attack scenarios, to provide a robust basis for testing CAN-ADS performance against stealthy cyber-attacks; (ii) *Data Pre-processing and Feature Extraction* - this stage involves cleaning and structuring the data to optimize training efficiency, thus enhancing the ML models' ability to learn from the data, and (iii) *ML Algorithms Training and Testing* - ML models are integrated into the proposed framework to detect and respond to anomalies, demonstrating their practical application.

Generation of Datasets

CAN-specific datasets are crucial for developing accurate and effective ML models to detect cyber threats in agricultural machinery. Most available datasets lack diversity in attack types and do not fully capture the complexities of real-world attack scenarios, leading to ML models that may be insufficiently trained and prone to inaccuracies. To resolve this and enhance the training accuracy and efficacy of ML models, we have developed realistic CAN-specific IDS datasets utilizing a HIL testbed setup at Iowa State University (ISU). The real-time CAN data captured from the testbed is augmented with open-source datasets to make them more comprehensive.

HIL Testbed Setup: The HIL testbed setup for this research, shown in Figure 4, was established in the Danfoss Fluid Power Teaching Laboratory within the Agricultural and Biosystems Engineering (ABE) Department at ISU. The testbed is equipped with ten hydraulic training stations, each powered by a 1 Hp single-phase motor driving a fixed displacement 4 cc/rev gear pump. The testbed utilizes Danfoss PLUS+1 MC controllers and various peripherals to create an advanced environment for simulating and analyzing hydraulic control systems. Each station includes essential components like 120Ω termination resistors to ensure proper CAN signal transmission, Danfoss CS 500, CS 10, and CG 150 gateways for connectivity and monitoring, and a Vector VN1610 interface for PC-based diagnostics. The system also features Danfoss DP720 displays for parameter visualization, MC050-110 controllers managing local CAN networks, sensors, control valve drivers, and Danfoss JS7000 joysticks for manual control. This comprehensive setup allows for realistic simulation and data collection in various operational and attack scenarios, which is critical for developing robust CAN-specific datasets.

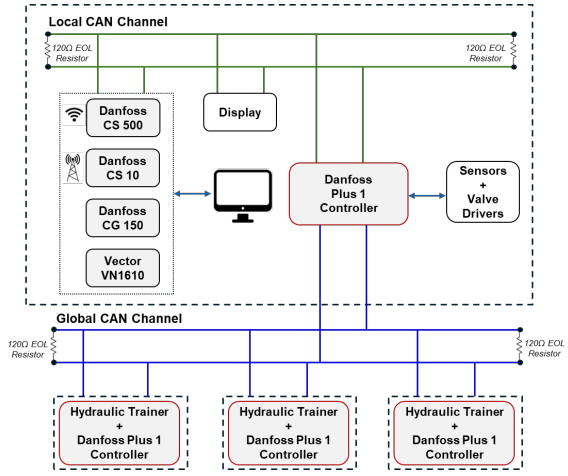


Figure 4: Network diagram for ISU HIL Testbed for CAN

Cyber-attack Injection: In this paper, we have simulated three realistic cyber-attack scenarios utilizing the HIL testbed: (1) Denial of Service (DoS), (2) Spoofing, and (3) False Data Injection (FDI). The experimental system involves a CAN-based hydraulic control system representing a vehicle motor speed control ECU, shown in Figure 4. In the DoS attack, high-priority messages are generated and transmitted at a fixed rate to flood the network, progressively increasing the busload and potentially causing a complete communication breakdown. Spoofing attacks involve injecting counterfeit messages into the CAN network to deceive the ECU controller. In the FDI attack, sensor measurements are manipulated by scaling them during specific conditions, leading to erroneous system responses.

Attack Assumptions: The adversary is assumed to have gained access to the vehicle's CAN network by compromising a specific access point, for e.g., the on-board diagnostic (OBD-II) port. This enables the adversary to inject malicious CAN messages to compromise a particular ECU, for e.g., motor speed control ECU.

Data Acquisition and Attack Trace Extraction: The CAN data for this study was collected from the ISU HIL Testbed using the Vector CANoe tool, which facilitated the monitoring and logging of CAN messages during various scenarios. The collected data was exported as ASCII log files, which contain detailed records of the CAN traffic, including message IDs, timestamps, and data payloads. To transform these raw logs into a more usable format for analysis, a Python parser tool was developed. This tool efficiently extracts relevant CAN features and data from the ASC logs, converting them into CSV files for each specific scenario, whether normal or attack. After the extraction, the parser tool further processes the data by merging the individual CSV files into a comprehensive dataset. During this merging process, class labels—indicating whether the data corresponds to benign (normal) operations or an attack—are appended to each entry based on the scenario in which it was recorded. The final output is an organized dataset enriched with CAN features and corresponding labels.

Traffic Source	Traffic Type	Category
HCRL Dataset	Normal CAN Messages	Benign
	'0x000' CAN ID Injection	DoS
	Arbitration ID ('0x164') Injection	Impersonation
	Spoofed CAN ID and DATA	Fuzzy
ISU Dataset	Normal CAN Messages	Benign
	Low Priority CAN ID Injection	DoS
	Spoofed Legitimate CAN ID	Spoofing
	Malicious DATA Injection	FDIA

Table 2: Augmented CAN Dataset Categorized Traffic Types

CAN Traffic Augmentation: To enhance the robustness of the proposed CAN-ADS, the dataset collected from the ISU HIL Testbed was augmented with carefully selected open-source CAN datasets, following the work in (Hasibi, Shokri, and Fooladi 2019). These datasets were chosen based on several critical characteristics to ensure their suitability for our research. The selected datasets are open-source, allowing for necessary customization. They include realistic network traffic encompassing normal, fault, and attack scenarios, providing a comprehensive spectrum of possible conditions. These datasets also contain original packet data capture log files and are labeled with relevant CAN attacks. This process creates a diverse and realistic dataset, making it highly valuable for benchmarking and developing our CAN-ADS to effectively detect a wide range of cyber threats in CAN networks. Table 2 shows an overview of the categorized CAN traffic in the augmented dataset.

Data Pre-processing and Feature Extraction

In this stage, the augmented CAN dataset is prepared for ML model training by transforming it into a clean, structured format by encoding categorical data, selecting relevant features, balancing class distributions, and normalizing values. These processes ensure that the data fed into the ML algorithms is optimized for accuracy and efficiency, enabling the model to detect anomalies in CAN networks with high precision.

One-hot Encoding: The categorical data is converted into an ML-suitable numerical format by transforming each categorical feature into a set of binary (Boolean) features, where each category is represented by a unique vector.

Feature Correlation and Reduction: The effectiveness of ML models is directly influenced by the quality of features used for training. In this study, we extracted various network and physical features from CAN networks to train the proposed ML-ADS. We applied dimensionality reduction techniques such as Principal Component Analysis (PCA) and Pearson's & Chi-Squared Correlation to eliminate redundancy and retain the most relevant features. These methods assign higher weights to uncorrelated features, and any features with a correlation greater than 95% were removed. This process enhances model accuracy, reduces training time, and prevents overfitting. A correlation heatmap for the augmented dataset is illustrated in Fig. 5. The key CAN features are:

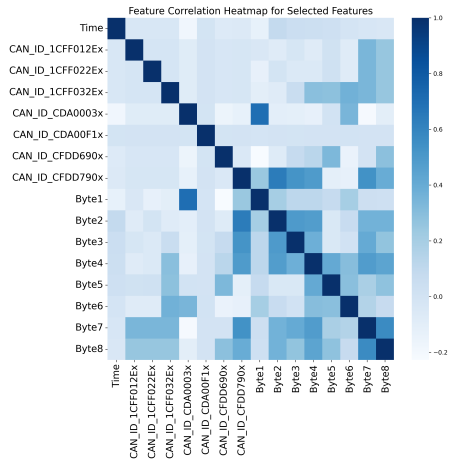


Figure 5: Sample feature correlation heatmap for Augmented Dataset

- **Timestamp:** The time at which the CAN message was captured or transmitted.
- **CAN ID:** A unique identifier for each CAN message, determining the priority of the message.
- **DLC:** Indicates the length of the data field in the CAN message.
- **Data:** The actual payload of the CAN message.
- **Class:** The category assigned to the CAN message.

Data Balancing: Most IDS datasets suffer from data imbalance across different traffic categories, leading to a disproportionate representation of certain attack types. This imbalance often causes ML models to be biased towards the majority class, resulting in poor performance when detecting less frequent but potentially more critical attack types. To address this issue, it is essential to balance the dataset, ensuring that the model gives equal attention to all attack categories, thereby improving its ability to detect each type effectively. To achieve this, we implemented a combination of the Synthetic Minority Oversampling Technique (SMOTE) and Random Undersampling (RUS) (Abdelkhalek and Govindarasu 2022).

SMOTE works by generating synthetic samples for the minority class, thereby increasing its representation in the dataset. It achieves this by interpolating new data points between existing minority class samples. Mathematically, for a minority class sample x_i, SMOTE selects a random sample x_j from its k-nearest neighbors and creates a synthetic sample, x_{new} as follows:

$$x_{new} = x_i + \delta \times (x_j - x_i) \quad (1)$$

Where, δ is a random number between 0 and 1. On the other hand, RUS reduces the number of majority class samples by randomly removing them, helping to balance the dataset. This is done by simply selecting a subset of the majority class samples, thereby reducing their dominance. Fig. 6 demonstrates the data balancing for the augmented dataset.

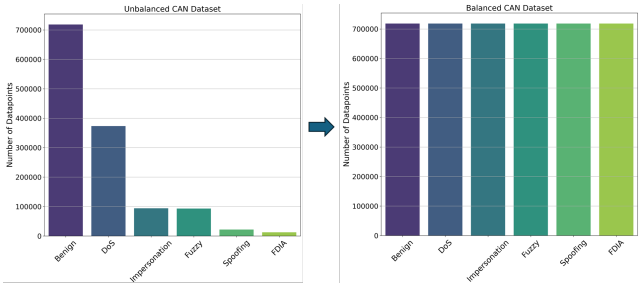


Figure 6: Data balancing for augmented CAN dataset

Metric	Description	Method[a]
Accuracy	Percentage of correctly classified instances (normal/fault/attack)	$\frac{TP+TN}{TP+FP+FN+TN}$
Precision	Percentage of malicious traffic from the overall detected traffic	$\frac{TP}{TP+FP}$
Recall	Percentage of the total malicious traffic that the model detected	$\frac{TP}{TP+FN}$
F1-Score	Mean between the precision and recall as a confidence measure	$\frac{TP}{TP+(FP+FN)/2}$

[a] ***TP**: True Positive, **TN**: True Negative, **FP**: False Positive, **FN**: False Negative*

Table 3: CAN-ADS Performance Evaluation Metrics

Data Normalization: To further enhance the training, the minimum-maximum scaling technique is applied to normalize the selected features between a given range of 0.0 to 1.0.

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

ML Models Training and Testing

Since our proposed system uses labeled datasets, we have implemented supervised multi-class classification ML algorithms to develop the CAN-ADS. Five popular ML algorithms most suited to time-series data analysis have been evaluated: (i) Decision Trees (DT), (ii) Random Forest (RF), (iii) K-Nearest Neighbor (KNN), (iv) XGBoost (XGB), and (v) Artificial Neural Networks (ANN). The selected datasets have been split into Training (70%), Testing (30%), and Validation sets, following standard ML practices. Various

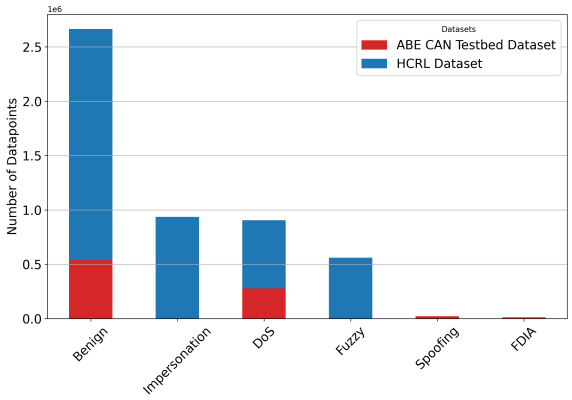


Figure 7: Data distribution in augmented dataset

Python ML libraries are utilized for training and testing the models: Scikit-Learn, NumPy, Pandas, Keras, and TensorFlow 2.0. The parameters for each algorithm were fine-tuned through a 30-cycle, 10-fold Cross-Validation (CV) selection grid. In the 10-fold CV, the training data is randomly split into 10 parts, utilizing 9 for training and 1 for testing. This cycle is repeated 30 times to identify the best parameters. The ML-ADS models are trained and tested on a Ubuntu 22.04 LTS 64-bit OS with Intel(R) Xeon(R) E5-2650 2.30GHz processor, 6-core CPU & 32 GB RAM.

Experimental Evaluation

Datasets and Performance Metrics

The augmented dataset utilized to evaluate the performance of the proposed CAN-ADS is illustrated in Figure 7. The experimental CAN data captured from the ISU HIL Testbed has been combined with the Hacking and Countermeasure Research Lab (HCRL) CAN intrusion datasets (Lee, Jeong, and Kim 2017). These open-source datasets were collected from a real vehicle via the OBD-II port, which includes 30 to 40 minutes of CAN traffic with normal and attack scenarios.

Figure 7 highlights the augmented CAN traffic on top of the original ISU HIL Testbed dataset. The final dataset has ≈ 546000 CAN traffic flows divided as follows: ≈ 2900000 normal or benign scenarios, ≈ 995000 impersonation attack scenarios, ≈ 934000 DoS attack scenarios, ≈ 591000 fuzzy attack scenarios, ≈ 21800 spoofing attack scenarios, and ≈ 12400 FDI attack scenarios. The performance of the CAN-ADS models on the augmented dataset has been evaluated through various metrics, as shown in Table 3.

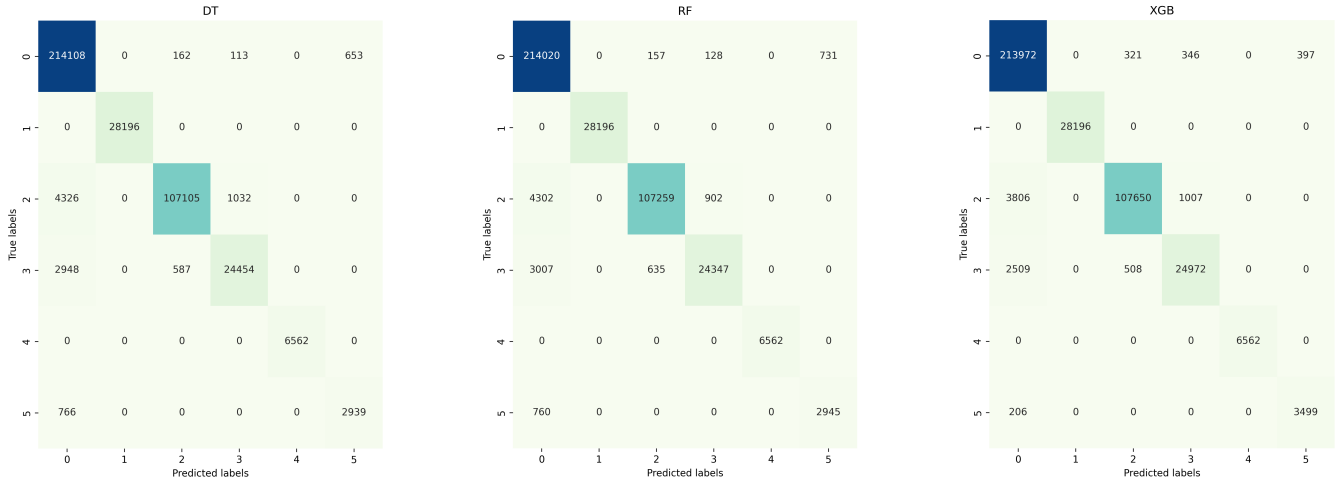
Results & Discussion

The experimental evaluation results of the proposed CAN-ADS models are presented in Table 4. The experimental evaluation aims to validate the effectiveness of the data augmentation process in enhancing the performance of ML models for anomaly detection in CAN networks. To demonstrate this, the CAN-ADS models are tested using two Baseline datasets, the HCRL dataset, and the ISU HIL Testbed dataset, alongside the augmented CAN dataset. From the results in Table 4, it can be observed that data augmentation significantly improves model performance, leading to better detection accuracy and reduced false positives. Additionally, the analysis evaluates training and testing latencies of the offline experimentation to provide insights into the trade-offs between accuracy and computational efficiency.

As described in previous sections, data pre-processing and feature correlation were performed for both baseline and augmented datasets. For the first baseline scenario (HCRL dataset), the results show that XGB algorithm achieves the highest detection accuracy (93.94%). Both DT and RF algorithms provide similar detection accuracies (93.37%). These are followed by the KNN and the ANN algorithms (92.42% and 88.53% accuracies, respectively). For the second baseline scenario (ISU HIL Testbed dataset), we observe a similar pattern in the model performance. In this case, XGB achieves the highest accuracy (95.98%), followed by RF (95.62%), DT (95.61%), KNN (95.10%), and

Dataset	ML Model	Accuracy		Precision		Recall		F1-Score	
		Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*	*Training*	*Testing*
Baseline (HCRL Dataset)	DT	93.88%	93.37%	94.37%	93.81%	93.88%	93.37%	93.81%	93.27%
	RF	93.88%	93.37%	94.37%	93.82%	93, 88%	93.37%	93.81%	93.27%
	KNN	92.78%	92.42%	92.79%	92.42%	92.78%	92.42%	92.73%	92.36%
	XGB	94.16%	93.94%	94.52%	94.29%	94.16%	93.94%	94.09%	93.87%
	ANN	89.33%	88.53%	89.82%	88.98%	89.33%	88.53%	89.23%	88.42%
Baseline (ISU Dataset)	DT	95.91%	95.61%	96.24%	95.89%	95.91%	95.61%	95.86%	95.54%
	RF	95.91%	95.62%	96.24%	95.92%	95.91%	97.62%	95.86%	95.55%
	KNN	95.33%	95.10%	95.41%	95.17%	95.33%	95.10%	95.28%	95.05%
	XGB	96.10%	95.98%	96.34%	96.20%	96.10%	95.98%	96.06%	95.93%
	ANN	92.71%	92.75%	93.06%	93.10%	92.71%	92.75%	92.65%	92.69%
Augmented CAN Dataset (HCRL + ISU)	DT	97.70%	97.31%	97.73%	97.33%	97.70%	97.31%	97.67%	97.28%
	RF	97.52%	97.30%	97.56%	97.33%	97.52%	97.30%	97.50%	97.27%
	KNN	97.25%	97.03%	97.25%	97.03%	97.25%	97.03%	97.23%	97.02%
	XGB	97.77%	97.69%	97.79%	97.70%	97.77%	97.69%	97.76%	97.67%
	ANN	95.49%	95.35%	95.53%	95.39%	95.49%	95.35%	95.45%	95.30%

Table 4: CAN-ADS Results of Baseline and Augmented Datasets



(a) DT Confusion Matrix (b) RF Confusion Matrix (c) XGB Confusion Matrix

Figure 8: Confusion matrices for the three best performing ML models implemented in the proposed CAN-ADS

ANN (92.75%). Lastly, for the augmented dataset, XGB achieves the highest detection accuracy (97.69%), followed by DT (97.31%), RF (97.30%), KNN (97.03%), and ANN (95.35%). The results indicate a consistent improvement in the detection performance for all five ML models across the three different datasets. The augmented dataset includes a broader range of traffic patterns and attack scenarios compared to the baseline datasets which enhances the ML models' training efficiency, leading to more accurate predictions and minimal overfitting. Figure 8 shows the confusion matrices for all five ML models for the augmented dataset, which demonstrates that all models show strong classification ability with low false positives. However, there are some false negatives, especially in classes 2 (DoS) and 3 (Impersonation), indicating that these models occasionally fail to detect true positives in these categories. Additionally, it was observed that DT had the fastest detection time ($\approx 0.1s$), followed by XGB ($\approx 0.9s$), RF ($\approx 3.7s$), ANN ($\approx 20.5s$), and KNN ($\approx 579.15s$). This work serves as a preliminary step towards real-time implementation of the CAN-ADS for anomaly detection. The CAN-ADS models can be deployed using serialization techniques within a vehicular environment and will be able to monitor the communication between the ECUs and the CAN controller without inducing delays or impeding critical functions.

Conclusion

This paper presents a supervised ML-based anomaly detection system (CAN-ADS) designed to detect both known and stealthy attacks within CAN systems used in agricultural and industrial machinery. By employing data augmentation and attack categorization techniques, the robustness of CAN-ADS was enhanced through bench-marking against open-

source and newly augmented CAN datasets. Key CAN network features were selected to improve ML model accuracy. The system was evaluated using five ML algorithms, demonstrating high detection accuracy with low false positive and negative rates and minimal overfitting. Future work will focus on (1) real-time implementation of CAN-ADS within a HIL testbed at ISU to validate its practical applicability and (2) extending CAN-ADS to include additional attack vectors to further enhance its robustness.

Acknowledgements

This research is funded in part by Iowa State University Presidential Interdisciplinary Research Initiative (PIRI).

References

Abdelkhalek, M.; and Govindarasu, M. 2022. ML-based Anomaly Detection System for DER DNP3 Communication in Smart Grid. In *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, 209–214.

Adkisson, M.; Kimmell, J. C.; Gupta, M.; and Abdelsalam, M. 2021. Autoencoder-based Anomaly Detection in Smart Farming Ecosystem. In *2021 IEEE International Conference on Big Data (Big Data)*, 3390–3399.

Ali, G.; Mijwil, M. M.; Buruga, B. A.; Abotaleb, M.; and Adamopoulos, I. 2024. A Survey on Artificial Intelligence in Cybersecurity for Smart Agriculture: State-of-the-Art, Cyber Threats, Artificial Intelligence Applications, and Ethical Concerns. *Mesopotamian Journal of Computer Science*, 2024: 71–121.

Asif, M. R. A.; Hasan, K. F.; Islam, M. Z.; and Khondoker, R. 2021. STRIDE-based Cyber Security Threat Modeling for IoT-enabled Precision Agriculture Systems. In *2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, 1–6.

Boland, H. M.; Burgett, M. I.; Etienne, A. J.; and III, R. M. S. 2021. An Overview of CAN-BUS Development, Utilization, and Future Potential in Serial Network Messaging for Off-Road Mobile Equipment. In Ahmad, F.; and Sultan, M., eds., *Technology in Agriculture*, chapter 25. Rijeka: IntechOpen.

Bozdal, M.; Samie, M.; Aslam, S.; and Jennions, I. 2020. Evaluation of CANBus Security Challenges. *Sensors*, 20(8).

Catalano, C.; Paiano, L.; Calabrese, F.; Cataldo, M.; Mancarella, L.; and Tommasi, F. 2022. Anomaly detection in smart agriculture systems. *Computers in Industry*, 143: 103750.

DHS. 2018. Threats to Precision Agriculture. https://www.dhs.gov/sites/default/files/publications/threats_to_food_and_agriculture_resources.pdf. Accessed on: Aug. 17, 2024.

Ferrag, M. A.; Shu, L.; Friha, O.; and Yang, X. 2022. Cyber Security Intrusion Detection for Agriculture 4.0: Machine Learning-Based Solutions, Datasets, and Future Directions. *IEEE/CAA Journal of Automatica Sinica*, 9(3): 407–436.

Hasibi, R.; Shokri, M.; and Fooladi, M. D. T. 2019. Augmentation Scheme for Dealing with Imbalanced Network Traffic Classification Using Deep Learning. *CoRR*, abs/1901.00204.

Kethineni, K.; and Gera, P. 2023. Iot-Based Privacy-Preserving Anomaly Detection Model for Smart Agriculture. *Systems*, 11(6).

Kulkarni, A.; Wang, Y.; Gopinath, M.; Sobien, D.; Rahman, A.; and Batarseh, F. A. 2024. A Review of Cybersecurity Incidents in the Food and Agriculture Sector. arXiv:2403.08036.

Lee, H.; Jeong, S. H.; and Kim, H. K. 2017. OTIDS: A Novel Intrusion Detection System for In-vehicle Network by Using Remote Frame. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, 57–5709.

Lieder, S.; and Schröter-Schlaack, C. 2021. Smart Farming Technologies in Arable Farming: Towards a Holistic Assessment of Opportunities and Risks. *Sustainability*, 13(12).

Mekouar, M. 2018. Food and agriculture organization of the united nations (FAO). *Yearbook of International Environmental Law*, 29: 448–468.

Nymann, P. H. 2021. CAN Bus Protocol: The Ultimate Guide (2022).

Obama, B. 2013. Executive Order 13636: Improving Critical Infrastructure Cybersecurity. *Federal Register*, 78(33): 11739.

Othmane, L. b.; Dhulipala, L.; Abdelkhalek, M.; Multari, N.; and Govindarasu, M. 2022. On the Performance of Detecting Injection of Fabricated Messages into the CAN Bus. *IEEE Transactions on Dependable and Secure Computing*, 19(1): 468–481.

Purohit, S.; and Govindarasu, M. 2022. ML-based Anomaly Detection for Intra-Vehicular CAN-bus Networks. In *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, 233–238.

Raghuvanshi, A.; Singh, U. K.; Sajja, G. S.; Pallathadka, H.; Asenso, E.; Kamal, M.; Singh, A.; and Phasinam, K. 2022. Intrusion Detection Using Machine Learning for Risk Mitigation in IoT-Enabled Smart Irrigation in Smart Farming. *Journal of Food Quality*, 2022(1): 3955514.

Rohrer, R.; Pitla, S.; and Luck, J. 2019. Tractor CANbus interface tools and application development for real-time data analysis. *Computers and Electronics in Agriculture*, 163: 104847.

Rudrakar, S.; and Rughani, P. 2023. IoT based Agriculture (Ag-IoT): A detailed study on Architecture, Security and Forensics. *Information Processing in Agriculture*.

S. Corrigan. 2016. Understanding and Applying the Controller Area Network (CAN) Protocol. https://www.ti.com/lit/an/sloa101b/sloa101b.pdf. Accessed on: Aug. 17, 2024.

Yazdinejad, A.; Zolfaghari, B.; Azmoodeh, A.; Dehghantanha, A.; Karimipour, H.; Fraser, E.; Green, A. G.; Russell, C.; and Duncan, E. 2021. A Review on Security of Smart Farming and Precision Agriculture: Security Aspects, Attacks, Threats and Countermeasures. *Applied Sciences*, 11(16).