IOWA STATE UNIVERSITY
Department of Electrical and Computer Engineering

CAE-R COP Research Symposium 2023
September 22, 2023

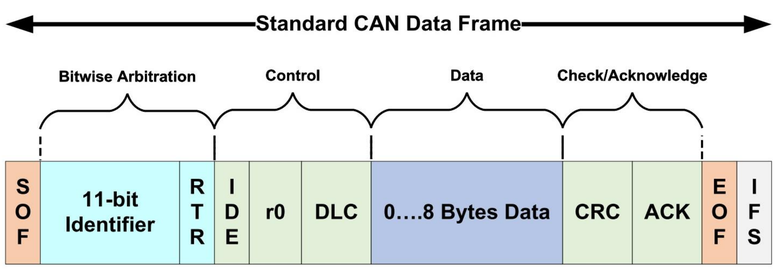
Shaurya Purohit, Manimaran Govindarasu

ML-based Anomaly Detection for Intra-Vehicular CAN-bus Networks

Abstract:

- The evolution of the automotive industry with the advent of autonomous vehicles has emphasized the critical need for enhanced security in vehicular networks and components like the widely-used CAN-bus system, which currently lacks sufficient security measures, leaving it vulnerable to various threats and attacks.
- Addressing this security gap, our research introduces a novel hybrid ADS leveraging ML techniques like Decision Tree, Random Forest, and XGBoost combined with rule-based systems.
- Tested on the CAN-intrusion dataset, the model demonstrates its effectiveness by detecting different types of attacks with a high level of accuracy (>90%) while incurring low latency, confirming our hybrid model's effectiveness and efficiency.

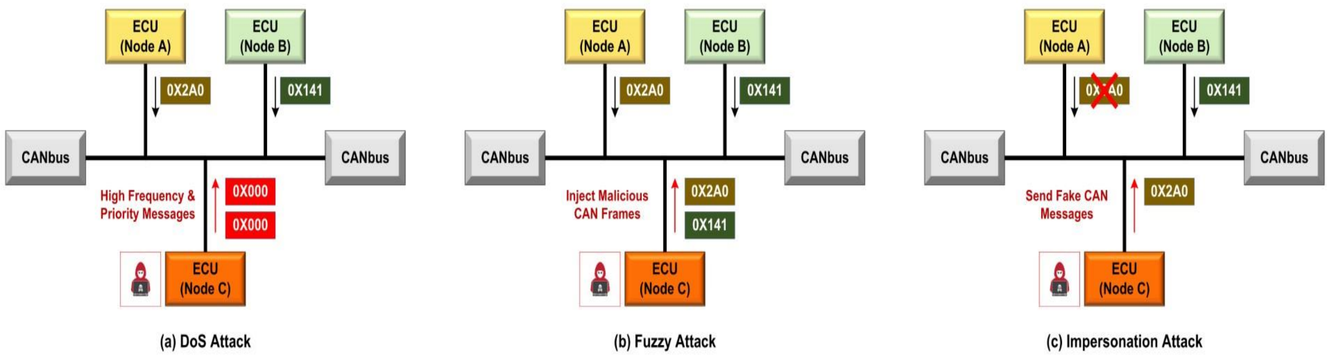
CAN-bus and Attack Scenario



SOF: *Start of Frame* **RTR:** *Remote Transmission Request* **IDE:** *Identifier Extension*
r0: *Reserved Bit* **DLC:** *Data Length Code* **CRC:** *Cyclic Redundancy Check*
ACK: *Acknowledgement* **EOF:** *End of Frame* **IFS:** *Interframe Space*

- CAN is message-oriented protocol.
- Rather than including the sender and receiver addresses, each CAN frame has a predefined ID and message structure defined.
- The ECU is configured to receive CAN messages with a specific CAN ID and ignore other messages at compile time.
- **Security Vulnerabilities in CAN:**
  - ✓ No encryption or authentication due to 8-byte payload limit.
  - ✓ Vulnerable to DoS, Fuzzy, and Impersonation attacks.

Major Attack Scenarios:



Dataset and Feature Extraction

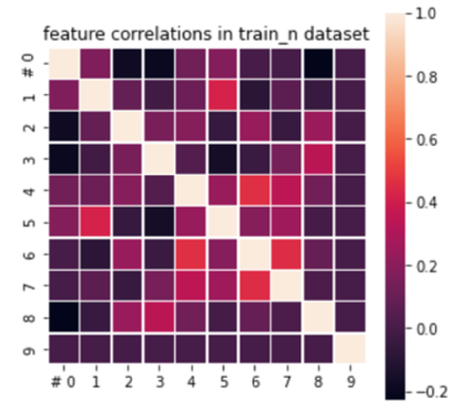
- **Dataset:** The "CAN-Intrusion Dataset" from the Hacking and Countermeasure Research Lab (HCRL), created in 2018 through real-vehicle CAN traffic monitoring via the OBD-II port, contains approximately 4.6 million tuples is utilized.

Attack Type	Number of Instances
Dos Attack	656,579
Fuzzy Attack	591,990
Impersonation Attack	995,472
Attack free state	2,369,868

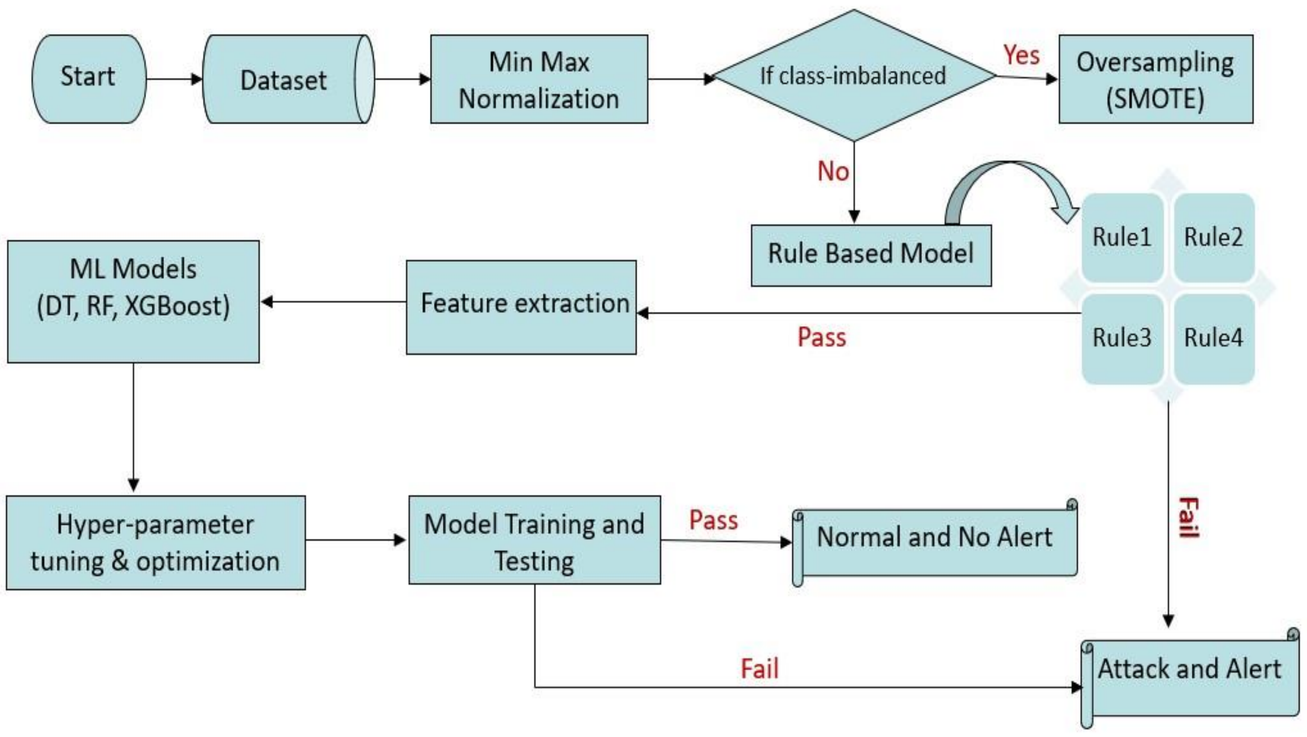
Attributes	Explanation
Timestamp	Time when the data is recorded
DLC	Number of data bytes
CAN ID	Unique ID
Data[0-7]	Message value

✓ It is evenly split between attack-free states and major attack data to avoid training bias thus ensuring the ADS can effectively recognize and respond to real attack situations, maintaining a balanced approach by not overly focusing on attack-free states during training.

- **Feature Extraction:** Leveraging the Information Gain (IG) technique for feature selection enhances system detection capabilities and understanding of the data; this not only lowers computational costs but also aids in identifying the most critical features (10 in our work), fostering improved performance.



PROPOSED HYBRID ADS MODEL



❖ **Step 1 - Rule-Based System:**
- ➢ Filters most attacks quickly, ensuring low false negative rates
- ➢ Utilizes four simple rules focusing on valid ID verification, regular message frequency, specific follow-up ID sequences, and time interval analysis to identify potential threats efficiently.
- ➢ Automatically flags messages not adhering to set rules as potential attacks.

❖ **Step 2 - ML-Based System**:
- ➢ Processes CAN messages that cleared the first stage
- ➢ Utilizes multi-threading of ML techniques (Decision Tree, Random Forest and XGBoost) for low execution time

❖ **Final Output**:
- ➢ Upholding high accuracy while incurring low execution time, only messages verified by both stages are channeled to the CAN-BUS vehicular networks.

Performance Evaluation

- **Metrics** : Accuracy , Detection rate (DR), False alarm rate (FAR), and F1-score are used.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad FAR = \frac{FP}{TN + FP}$$

$$DR = \frac{TP}{TP + FN} \qquad F1 = \frac{2 * TP}{2 * TP + FP + FN}$$

- The proposed model was trained and tested on a system with an Intel(R) Core i7-8550 processor and 16 GB of RAM. For our model, we used Python, PyCharm IDE 2020.1, Keras, Pandas, and XGBoost with TensorFlow as the backend.

Method	Accuracy(%)	DR(%)	FAR(%)	F1
DT	96.38	96.2	2.6	0.961
RF	94.71	94.46	3.4	0.943
XGBoost	97.62	97.54	1.8	0.972
RB	75.29	75.12	6.2	0.749
RB + DT	93.25	93.2	3.6	0.931
RB + RF	90.63	90.23	5.8	0.901
RB + XGBoost	94.95	94.87	4.2	0.948

Method	Execution Time(S)
KNN	911.6
SVM	13765.6
SAIDuCANT	-
DT	1278.4
RF	1905.3
XGBoost	2308.1
RB	132.006
RB + DT	304
RB + RF	412.6
RB + XGBoost	350.2

- Leveraging multi-threading enabled by DT, RF, and XGBoost in combination with a rule-based system, the proposed model outperforms previous works such as KNN and SVM, delivering not only higher accuracy and F1-score but also significantly reduced execution time (latency).

Conclusion and Future Work

❑ We presented a hybrid anomaly detection model for intra-vehicular CAN buses, utilizing a rule-based system for checking injected CAN signals in the first step followed by machine learning techniques in the second step. Leveraging SMOTE for oversampling mitigated class imbalance and reduced computational costs, facilitates potential real-time implementation due to the favorable balance of high accuracy (above 90%) and low execution time as demonstrated on the CAN-Intrusion dataset.

❑ In future work, we aim to further optimize our hybrid ADS model by incorporating new rules and leveraging cutting-edge machine learning algorithms. This includes hyperparameter optimization using advanced techniques such as particle swarm and Bayesian optimization, alongside exploring the latest advancements in machine learning to enhance the model's efficiency and effectiveness.

Contact Information

Email: shaurya1@iastate.edu, gmani@iastate.edu ; Address: 2520 Osborn Dr., Ames, IA 50011