MTH2526-G13 Data Mining for Cybersecurity Project

Phase 2

The CICIoV2024 dataset was created to support intrusion detection research in Internet of Vehicles (IoV) systems. It contains CAN-bus traffic collected from a real vehicle under both benign and attacks condition, including spoofing and denial-of-service attacks. The benign traffic subset was selected to focus on data understanding and processing.

FEATURE	DESCRIPTION
ID	Arbitration value indicating message priority and type
DATA_0 to DATA_7	Bytes of Data transmitted
label	Traffic label (benign vs malicious)
category	Broader attack category
Specific class	Specific attack type

DATA UNDERSTANDING: Initial exploration was conducted to examine dataset structure, feature types, and missing values. This included inspecting dataset dimensions, data types, and summary statistics.

DATA CLEANING: Missing values in numerical features were handled using median imputation to minimize the influence of outliers and ensure data completeness.

ADDRESSING MISSING VALUES: Missing values were identified in the dataset. If the data cleaning process is Sloopy, the model might “learn” to detect your imputation method rather than the actual cyberattack. By using median/mode imputation and feather pruning, researchers ensure that the resulting machine learning models are detecting actual adversarial patterns in the CAN bus rather than artifacts of the data cleaning process.

During a cyberattack, a hacker might inject “fuzzing” data-extreme, nonsensical values intended to overwhelm the system. If you used the mean to fill missing gaps, these extreme attack values would pull the average away from reality. By using median, you ensure that missing sensor data is filled with a value that represents the most typical “baseline” state of the vehicle. This prevents the imputed data from being accidentally “pulled” toward the extreme values seen during active attacks, preserving the integrity of the normal traffic distribution. If a feature (like payload bytes or metadata) is missing in 50% or more of the instances, trying to “guess” those values creates noise – artificial patterns that do not exist in the real vehicle.

FEATURE SELECTION: Feature selection was performed to retain attributes that directly represent intra-vehicular communication and attack characterization. The arbitration ID and individual data bytes (DATA_0 through DATA_7) were retained as fundamental signals of CAN-bus behavior. Additional features describing labels and attacks categories were included for supervised analysis. Irrelevant attributes were removed to reduce noise and dimensionality.

DATA TRANSFORMATION: Feature scaling was applied using Standard Scale to normalize numerical features, ensuring consistency and preparing the dataset for future machine learning models.

OUTPUT: The processed dataset was saved as "*CICIoV2024_cleaned.csv"* and will be used in Phase 3 for model development.

GITHUB REPOSITORY: The complete implementation and dataset preprocessing steps are available on GitHub: 👉 https://github.com/OluwatomiwaAkinyemi/CICIoV2024-DataMining-Phase2.git