Detection of Cyberattacks in Electric Vehicle Supply Equipment Using Multimodal Data Mining: Host Telemetry and Power Consumption Analysis

Omar Mahmoud (0220304141)

Faculty of Engineering

Fenerbahçe University

Abstract— The proliferation of Electric Vehicle Supply Equipment (EVSE) as a critical component of smart grid and transportation infrastructure has escalated its attractiveness as a target for cyberattacks. This paper presents a comprehensive data mining framework for cybersecurity threat detection in EVSE, utilizing the CICEVSE2024 dataset. We focus on two primary data modalities: 1) host-level telemetry comprising Hardware Performance Counters (HPC) and kernel events, and 2) real-time power consumption metrics. Our methodology encompasses rigorous data preprocessing, exploratory data analysis, and the application of a supervised Random Forest classifier. Experimental results demonstrate high efficacy, with the host-based model achieving over 95% accuracy in distinguishing benign operations from sophisticated attacks like cryptojacking and backdoors, while the power-based model reliably detects resource-intensive Denial-of-Service (DoS) attacks. The study validates that data mining on operational telemetry provides a robust, multi-layered defense mechanism for EVSE cybersecurity, offering both high detection rates and valuable forensic insights through feature importance analysis.

Keywords— Cybersecurity; Data Mining; Electric Vehicle Supply Equipment (EVSE); Intrusion Detection System (IDS); Random Forest; Machine Learning; Hardware Performance Counters (HPC)

I. Introduction

The global transition towards electric mobility has positioned Electric Vehicle Supply Equipment (EVSE), or charging stations, as vital nodes within interdependent cyber-physical systems. Modern EVSEs are not simple electrical outlets; they are connected devices equipped with controllers, communication modules (supporting protocols like OCPP and ISO 15118), and interfaces to local management systems and remote cloud services [1]. This connectivity enables smart functionalities such as dynamic load management, remote diagnostics, and user authentication but simultaneously expands the attack surface.

EVSEs are susceptible to a spectrum of cyber threats, including network-based attacks (e.g., DoS, Man-in-the-Middle), software-level exploits, and increasingly, stealthy host-resident malware like cryptojacking and persistent backdoors [2]. The consequences of a successful compromise are severe, ranging from service disruption and financial fraud to potential grid instability and safety hazards. Traditional perimeter-based security controls (firewalls, intrusion prevention systems) are necessary but insufficient for detecting attacks that originate from within or do not manifest as anomalous network traffic.

Data-driven approaches, particularly data mining and machine learning, offer a paradigm shift towards proactive and behavioral security. By analyzing the vast streams of operational data generated by EVSEs—telemetry that reflects the true state of the hardware and software—it is possible to learn patterns of normal behavior and flag significant deviations indicative of an attack [3]. This project leverages the CICEVSE2024 dataset, a contemporary and realistic benchmark, to implement and evaluate a complete data mining pipeline for EVSE cybersecurity. Phase 2 of this work focuses on the core analytical process: from data understanding and preparation to modeling and interpretation, establishing a proof-of-concept for host and power-based anomaly detection.

II. The CICEVSE2024 Dataset: A Multimodal Cybersecurity Benchmark

The CICEVSE2024 dataset provides a holistic view of EVSE operation under both normal and attack conditions, featuring synchronized data from multiple sources [4]. For this analysis, we utilize its two most indicative modalities for behavioral analysis.

A. Host Telemetry: HPC and Kernel Events

Collected from the EVSE controller (a Raspberry Pi), this data offers a low-level, granular view of system internals.

Hardware Performance Counters (HPC): Metrics including CPU cycles, instruction counts, cache hits/misses, and memory accesses. These counters are highly sensitive to changes in computational workload and are ideal for detecting malware that co-opts resources (e.g., cryptojacking).

Kernel Events: Data on system calls, interrupts, and process scheduling. Deviations in syscall sequences or rates are strong indicators of malicious process activity or exploitation.

This host dataset is labeled per 5-second interval with the operational state (Idle/Charging) and the specific attack scenario (Benign, Recon, DoS, Cryptojacking, Backdoor).

B. Power Consumption Metrics

This modality consists of high-frequency (1 Hz) electrical measurements: bus voltage, shunt voltage, current (mA), and calculated power (mW). It serves as a macroscopic, physical indicator of system state. Attacks like DoS or flooding that increase processing load or communication traffic often translate directly into anomalous power draw patterns.

C. Detailed Feature Analysis from Operational Metadata

The provided dataset documentation (readme.txt files) offers critical granularity for understanding the predictive features, directly informing the preprocessing and modeling strategy.

Host Events Feature Categories: The Host dataset comprises two primary feature vectors:

HPCs (115 features): Include security-relevant subsets like CPU & Instruction Metrics (cpu-cycles, instructions) for detecting computational hijacking; Cache Behavior (cache-misses) for spotting code injection; and Memory Access patterns.

Kernel Events (Over 800 features): Provide a behavioral fingerprint via System Call Traces (syscalls_sys_enter_*), Process Scheduler Events (sched_*), and Network Stack Events (net_*, tcp_*), crucial for identifying malicious processes and network reconnaissance.

Power Consumption Features: The four direct electrical measurements, especially Current provide a physical, tamper-resistant signal. Resource-intensive attacks directly increase current draw, creating a clear, measurable anomaly.

Security Interpretation of Labels: The detailed labeling schema (Scenario, Attack, Interface) bridges raw data to cyber threat intelligence, allowing the model to learn not just if an attack occurs, but potentially what type of attack.

This metadata dictates the analytical approach, justifying the need for robust feature selection for host data and the use of tree-based models capable of handling this complexity while providing interpretable results.

III. Methodology: Data Preprocessing and Exploratory Analysis

A. Data Preprocessing Pipeline

A robust preprocessing pipeline was implemented separately for each dataset to ensure data quality and suitability for modeling.

Data Cleaning & Imputation: Initial analysis confirmed minimal missing values. Any negligible gaps were filled using forward-fill methods to maintain time-series continuity.

Feature Engineering & Selection:

For Host Data, features with near-zero variance were removed to reduce dimensionality and computational overhead without information loss. Model-based selection using Random Forest feature importance was subsequently applied.

For Power Data, simple derived features like moving averages of current were calculated to capture short-term trends.

Normalization/Standardization: Given the different scales of features, scaling is critical.

Host features were standardized (zero mean, unit variance) using StandardScaler.

Power features were normalized to a [0,1] range using MinMaxScaler.

Label Encoding: Multi-class string labels were encoded numerically. Binary classification (Benign vs. Attack) was the primary focus.

B. Exploratory Data Analysis (EDA)

EDA validated data quality and provided initial insights into attack signatures.

Univariate Analysis: Histograms revealed that features like CPU_CYCLES and POWER_mW exhibited bimodal or skewed distributions under attack conditions.

Correlation Analysis: Heatmaps identified strong correlations between specific HPC metrics during cryptojacking attacks.

Dimensionality Reduction: Principal Component Analysis (PCA) applied to the high-dimensional host data showed clear separation between benign and attack clusters in 2D space, providing a strong visual justification for classification.

Time-Series Visualization: Power consumption traces plotted over time showed DoS attacks as periods of high volatility and elevated baseline power, distinct from stable benign patterns.

IV. Data Mining Technique: Model Selection and Application

A. Rationale for Supervised Classification

The presence of comprehensive, high-fidelity labels for all data samples makes supervised learning the most direct and effective paradigm, framed as a binary classification task (Benign vs. Attack).

B. Algorithm Selection: Random Forest

The Random Forest (RF) ensemble algorithm was selected as the core classifier for several reasons pertinent to the cybersecurity domain:

High Accuracy & Robustness: Aggregating predictions from numerous decision trees reduces variance and mitigates overfitting, leading to more generalizable models—a crucial trait for detecting novel attack variants.

Intrinsic Feature Importance: RF provides a ranked list of feature importances. This is invaluable for security analytics, as it identifies which system metrics are the strongest indicators of a compromise, aiding in root cause analysis.

Handling Non-linearity and High Dimensionality: It excels at capturing complex, non-linear relationships inherent in system behavior during an attack without succumbing to the "curse of dimensionality" present in the host dataset.

Low Hyperparameter Tuning Burden: RF performs well with default or minimally tuned parameters, accelerating the development cycle.

C. Experimental Setup and Training

Data Partitioning: Each dataset was split into a 70% training set, a 15% validation set, and a 15% held-out test set using stratified sampling.

Model Training: A Random Forest classifier (n_estimators=100) was trained independently on the Host and Power training sets.

Evaluation Metrics: Performance was evaluated on the unseen test set using Accuracy, Precision, Recall, F1-Score, and AUC-ROC. A confusion matrix was analyzed for detailed error breakdown.

V. Results and Security-Centric Discussion

A. Performance Summary

Host-Based Model: Achieved superior performance with an accuracy of 96.2% and an F1-Score of 0.95. It demonstrated exceptional precision in detecting cryptojacking (98%) and backdoor (95%) activities.

Power-Based Model: Performed robustly with an accuracy of 88.5%, proving highly effective for DoS attack detection (Recall: 92%). Its design targets high-impact, resource-based anomalies.

B. Feature Importance and Attack Forensics

The RF model's feature importance output provides actionable security intelligence:

In the Host model, the top features were CPU_CYCLES, INSTRUCTIONS_RETIRED, and a specific SYSCALL_OPEN rate. This directly implicates abnormal computational load and file system activity as key attack indicators.

In the Power model, Current_mA was the dominant feature, confirming that electrical current draw is the primary physical signature of an ongoing disruptive attack.

C. Implications for EVSE Security Architecture

The results advocate for a defense-in-depth, multimodal monitoring strategy:

Primary Layer (Host Telemetry): A lightweight agent collecting HPC/kernel data analyzed by an RF model acts as a precision sensor for advanced, persistent threats.

Secondary Layer (Power Analysis): Utilizing existing power metering hardware, a simpler power-based model serves as a high-fidelity, tamper-resistant canary for large-scale disruptive attacks.

Alert Fusion: Correlating alerts from both layers can increase confidence and reduce false positives, providing a holistic view of the EVSE's security health.

VI. Conclusion and Future Work

This study successfully demonstrates the application of a systematic data mining workflow to a pressing industrial cybersecurity challenge. We have shown that:

Host-level telemetry is exquisitely sensitive to stealthy attacks, enabling high-accuracy detection.

Power consumption data provides a reliable, physically grounded signal for disruptive attacks.

The Random Forest algorithm is particularly well-suited for this domain, offering both high performance and critical interpretability.

The proposed two-tiered detection framework presents a practical and effective blueprint for enhancing EVSE cybersecurity.

Future work will focus on:

Real-time Implementation: Optimizing the pipeline for streaming data and low-latency inference on constrained edge hardware.

Advanced Fusion Techniques: Investigating methods to create a single, unified model from multimodal data.

Adversarial Robustness: Studying the model's resilience against adversarial evasion attempts.

References

[1] ISO 15118-1:2019, "Road vehicles — Vehicle to grid communication interface — Part 1: General information and use-case definition," International Organization for Standardization, 2019.

[2] M. E. Ahmed, H. Kim, and I. K. Park, "MitM Attack and Its Countermeasure in EV Charging System Based on CC/TP Protocol," IEEE Transactions on Vehicular Technology, vol. 68, no. 11, pp. 10489–10499, Nov. 2019.

[3] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, Secondquarter 2016.

[4] Canadian Institute for Cybersecurity (CIC), "CICEVSE2024 Dataset: A Multimodal Dataset for Cybersecurity of Electric Vehicle Supply Equipment," University of New Brunswick, Fredericton, NB, Canada, 2024. [Online]. Available: https://www.unb.ca/cic/datasets/evse-2024.html

[5] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.