Thomas ESCAFFRE, Azhar EL HIDAOUI, Illan-Emmanuel COCO-GUIGNARD OCC1

PRE-PROJECT MACHINE LEARNING

1. Business Challenge and State-of-the-Art

Context:

With the rise of Internet of Things (IoT) devices in smart homes, the number of potential vulnerabilities has increased. IoT devices—such as smart lights, security cameras, and appliances-communicate over a home network, but often lack strong security measures. This opens the door to various cyber threats, from unauthorized access to full-scale network breaches. Given the growing adoption of smart homes, the need for advanced intrusion detection systems (IDS) is critical.

Challenges:

- Vulnerabilities in IoT devices: Many IoT devices are not equipped with sufficient security, making them prime targets for attacks.

- Complexity of detecting intrusions: Traditional IDS systems often struggle to handle the specific behavior patterns and communication protocols unique to IoT ecosystems.

- Balancing security and usability: A security solution must not hinder user experience while providing robust protection.

State-of-the-Art Solutions:

Current Intrusion Detection Systems (IDS) rely on signature-based, anomaly-based, and hybrid detection approaches. Machine learning models are increasingly effective in IDS for IoT by recognizing normal and suspicious network patterns using data features like protocol type, traffic volume, and service access. However, adapting to the evolving and dynamic nature of IoT environments remains a challenge. Continual model updates are required to keep up with these changes.

2. Data Description and Data Sources

The dataset is designed to capture a wide variety of features that characterize network traffic within a smart home environment, distinguishing normal behaviour from potential intrusions.

Link : https://www.kaggle.com/datasets/bobaaayoung/dataset-invade

Key Features:

- **Basic Metrics**:
  - `duration`: The length of the connection.
  - `protocol_type`: The protocol used (e.g., TCP, UDP).
  - `src_bytes` and `dst_bytes`: Bytes sent from source to destination and vice versa.
- **Connection and Behavior Indicators**:
  - `count`, `srv_count`: Number of connections to the same host or service, useful for detecting scans or floods.
  - `serror_rate`, `rerror_rate`: Error rates in connection attempts (e.g., SYN or reset errors), often a sign of attacks.
- **Binary Indicators**:
  - `land`, `urgent`, `logged_in`: Binary indicators that capture specific behaviors, such as if the source and destination IPs are the same or if the user is logged in.
- **Target**:
  - `attack`: The label indicating whether a connection was part of an attack (Yes or No).

Data Sources:

The dataset, sourced from simulated smart home network traffic, captures various IoT device interactions, reflecting both benign and malicious activity. This design provides a realistic foundation for training IDS models by including standard traffic alongside potential threats, aiding in accurate intrusion detection.

3. Business Objectives and the Scope

Business Objectives:

The primary objective is to develop a Machine Learning-based Intrusion Detection System (IDS) that can identify and mitigate cyber threats in smart home environments. Key goals include:

- Real-time threat detection: Detect attacks such as port scans, denial-of-service (DoS) attempts, and data exfiltration before they cause significant damage.

- Minimizing false positives: Ensuring that legitimate activity is not mistakenly flagged as an intrusion, which could hinder user experience.

- Scalability: The system should be capable of handling a wide variety of IoT devices and traffic patterns without requiring constant retraining.

Scope:

The project includes three main areas:

- Exploratory Data Analysis (EDA): Examine dataset distributions, trends, and key features for meaningful insights.

- Feature Engineering: Select, transform, and potentially generate new features critical to intrusion detection

- Model Development and Evaluation: The project includes three main areas: Use classification models such as Decision Trees, Random Forests, or Neural Networks to detect attacks. Evaluation metrics, particularly recall and precision, are crucial to balance false positives and effective detection.

4. Work Plan

Data Understanding and Cleaning: Conduct an initial analysis to identify and correct missing values, inconsistencies, or outliers.

EDA: Explore key metrics (e.g., distribution of attacks) and use visualizations (e.g., histograms, boxplots) to analyze traffic behavior and anomalies.

Feature Selection: Identify relevant features via correlation analysis; engineer domain-specific features if necessary.

Model Training and Testing: Split data into training/testing sets, evaluate models, and compare them using accuracy, precision, and recall.

Model Evaluation: Use cross-validation and visualize results to assess true vs. false detections with a confusion matrix.

Deployment: Outline deployment steps in real-time smart home environments, including ongoing performance monitoring to adapt to emerging threats.

5. Conclusion

The development of a Machine Learning-based Intrusion Detection System (IDS) for smart homes represents a critical step in improving the security of IoT ecosystems. The dataset’s features capture a wide array of network behaviors, enabling robust identification of cyber threats. By leveraging both exploratory data analysis and advanced machine learning techniques, we can create a system capable of protecting smart home environments from evolving cyber threats while minimizing disruptions to legitimate activities.

6. References

“A Review of Intrusion Detection Systems Using Machine and Deep Learning in IoT,” MDPI Information, 2023. This article explores IDS machine learning techniques for IoT security, detailing attacks and future trends in home security systems.

https://www.mdpi.com/2078-2489/15/10/631

“Towards Securing Smart Homes: Systematic Literature Review on Malware Detection,” MDPI Information, 2023. This study reviews malware detection in IoT, offering insights into common threats and strategies that enhance IoT security, providing foundational knowledge for IDS model development.
https://www.mdpi.com/2078-2489/15/10/631