IID and Non-IID Data Distribution in N-BaIoT Dataset

What is N-BaIoT Dataset?

The N-BaIoT dataset is a benchmark dataset used to detect cyberattacks in Internet of Things (IoT) networks. It includes network traffic data from various IoT devices (like smart lights and security cameras) under normal and abnormal conditions, such as botnet attacks (Mirai and Bashlite). With 33 features capturing network activity (e.g., packet size, time between packets), it helps researchers develop anomaly detection models for securing IoT devices against cyber threats. The dataset is widely used to train and evaluate machine learning models for intrusion detection.

An example using the **N-BaIoT dataset**:

- **Normal Scenario**: A smart security camera in your home transmits regular data packets over the network. The dataset records normal traffic patterns like packet size and frequency. Machine learning models learn from this data to recognize it as normal behavior.
- **Attack Scenario**: The same camera gets infected by the **Mirai botnet**, which forces it to send a flood of data, causing a **DDoS attack**. The dataset captures this abnormal spike in traffic. The model uses this to detect that the camera is now under attack, flagging it as malicious behavior.

This helps build systems to identify and prevent such attacks in real-time.

Purpose of N-BaIoT Dataset:

- It is mainly used for **anomaly detection** and **intrusion detection** in IoT environments.
- The dataset helps researchers and developers build algorithms that can detect malicious activities or abnormal patterns in IoT networks, enhancing security measures for smart devices.

Why It’s Important:

- IoT devices are often vulnerable to cyberattacks due to weak security protocols.
- The N-BaIoT dataset helps simulate real-world attack scenarios, allowing the development of more robust security systems for IoT networks.

What is IID data?

IID (Independent and Identically Distributed) Data:

In an **IID** setting, the following two assumptions are made:

- **Independent**: Each data sample does not depend on or influence other samples. In other words, the occurrence of one data point does not provide information about the occurrence of another.
- **Identically Distributed**: All data points come from the same underlying distribution. This means that all samples are drawn from the same statistical population and follow the same probability distribution.

Example of IID in the N-BaIoT Dataset:

For the N-BaIoT dataset, which includes network traffic data from various IoT devices (such as security cameras, thermostats, and baby monitors), an IID scenario would look like this:

- **Independent**: Traffic packets or flow records are independent of each other. For example, the network traffic generated by a security camera at one point in time does not depend on the traffic generated at a later time, or on traffic generated by other devices. Each packet of data is treated as a standalone instance.
- **Identically Distributed**: All traffic data, whether from different devices or at different times, is assumed to come from the same distribution. This would imply that traffic patterns for different devices (such as a baby monitor vs. a thermostat) and different time periods are statistically indistinguishable. In practice, this means that the network traffic would have similar characteristics (such as packet sizes, inter-arrival times, or protocols used) across all devices.

Implications in Machine Learning:

- **Homogeneous distribution**: If you assume the data is IID, you treat all traffic samples as if they are drawn from the same population. This implies that no special treatment is needed for data from different devices or times.
- **No temporal or device-specific dependence**: Each data sample is processed as if it has no relationship to the previous or next one, and all devices are assumed to behave in a statistically similar manner.

Real-World Example of IID:

IID Example (Independent and Identically Distributed):

Imagine you are conducting a survey of people’s favourite ice cream flavours. You randomly ask 1,000 people from different parts of a city, but:

- Each person’s choice is independent: One person choosing chocolate doesn’t affect the next person’s choice.
- All people are considered to come from the same group: You assume that everyone in the city has an equal chance of choosing any flavour, regardless of their age, location, or background.

In this case, you treat all data points (people's choices) as independent and identically distributed. The flavour preferences are assumed to follow the same statistical distribution across the whole population.

What is Non-IID data?

Non-IID (Non-Independent and Non-Identically Distributed) Data

In the **non-IID** case, one or both of the following assumptions are violated:

- **Non-Independent**: Some data samples are dependent on others. This could occur in sequential data, where previous samples influence future samples. In the context of network traffic, packets or flows from the same device or time window may have dependencies based on prior traffic behavior.
- **Non-Identically Distributed**: Different subsets of the data come from different distributions. For example, traffic from different devices, different time windows, or attack vs. normal traffic could follow distinct statistical patterns. Each device could have its own unique characteristics, and attack traffic could behave very differently from benign traffic.

Example of Non-IID in the N-BaIoT Dataset:

In the context of the N-BaIoT dataset, a non-IID scenario is more realistic because:

- **Non-Independent**: IoT devices usually generate network traffic that has some temporal dependencies. For instance, if a device is under attack (like a DoS or scanning attack), the traffic behaviour during the attack is likely to affect future traffic patterns. Likewise, normal traffic might exhibit temporal correlation, where certain events (such as turning on the thermostat) lead to specific traffic patterns that continue for some time.
- **Non-Identically Distributed**:
    - **Different Devices**: The network traffic generated by different IoT devices can vary significantly due to their functional differences. For example, a **security camera** constantly streams video, resulting in high, continuous traffic, whereas a **smart light bulb** only generates traffic when it’s turned on or off, resulting in sparse and small traffic bursts.
    - **Benign vs. Attack Traffic**: Normal (benign) traffic and attack traffic are expected to follow very different distributions. Benign traffic might follow predictable patterns (e.g., periodic updates or low-frequency control messages), while attack traffic might exhibit sudden spikes (e.g., a DDoS attack generating a flood of packets in a short period).
    - **Time Dependencies**: Traffic may also depend on previous conditions or events. For example, during a **botnet attack**, a sudden increase in packet transmission might persist for a period, and this traffic could influence subsequent packets.

Implications in Machine Learning:

- **Heterogeneous distribution**: In the non-IID case, you need to account for different distributions across devices, traffic types (benign vs. attack), or time windows. Different machine learning models might need to be trained separately for different devices or conditions, or you might need domain adaptation techniques to handle shifts in data distribution.
- **Time-series analysis**: If traffic data is dependent on prior traffic (i.e., temporal correlations exist), time-series models such as **Recurrent Neural Networks (RNNs)** or **LSTMs (Long Short-Term Memory networks)** might be needed to capture these dependencies.

Real-World Example of Non-IID:

Non-IID Example (Non-Independent and Non-Identically Distributed):

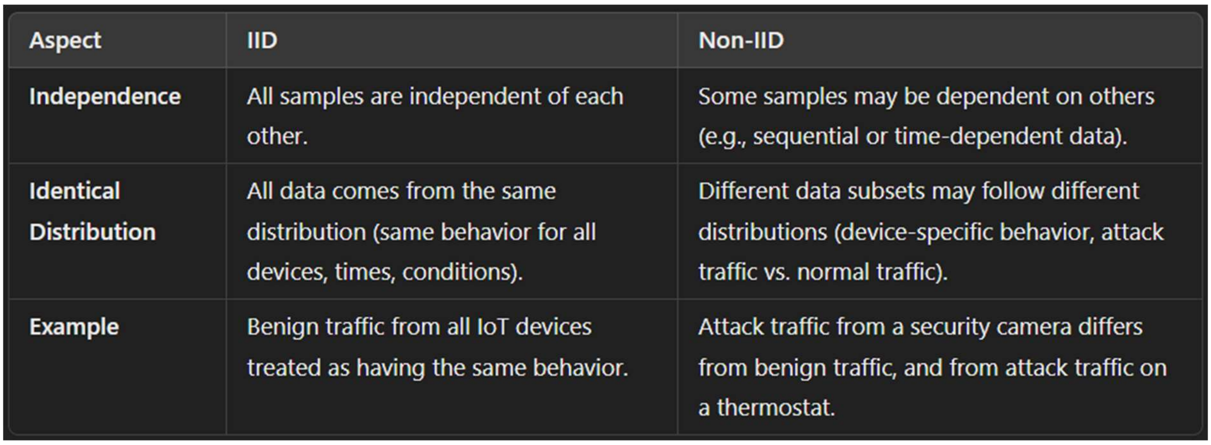
Now, imagine you conduct the same survey, but instead of randomly selecting people, you survey different neighbourhoods, each with distinct preferences. For example:

- In neighbourhood A, people mostly choose vanilla because there's a popular vanilla ice cream shop.
- In neighbourhood B, people prefer chocolate.
- In neighbourhood C, the survey takes place at an ice cream festival where promotions influence flavour choices.

Here, the data is **non-IID** because:

- People from different neighbourhoods have different preferences (i.e., the data is not identically distributed).
- In some neighbourhoods, one person's choice may influence another's (e.g., at the festival where people might see others trying certain flavours).

Key Differences (tabular) Between IID and Non-IID in N-BaIoT:

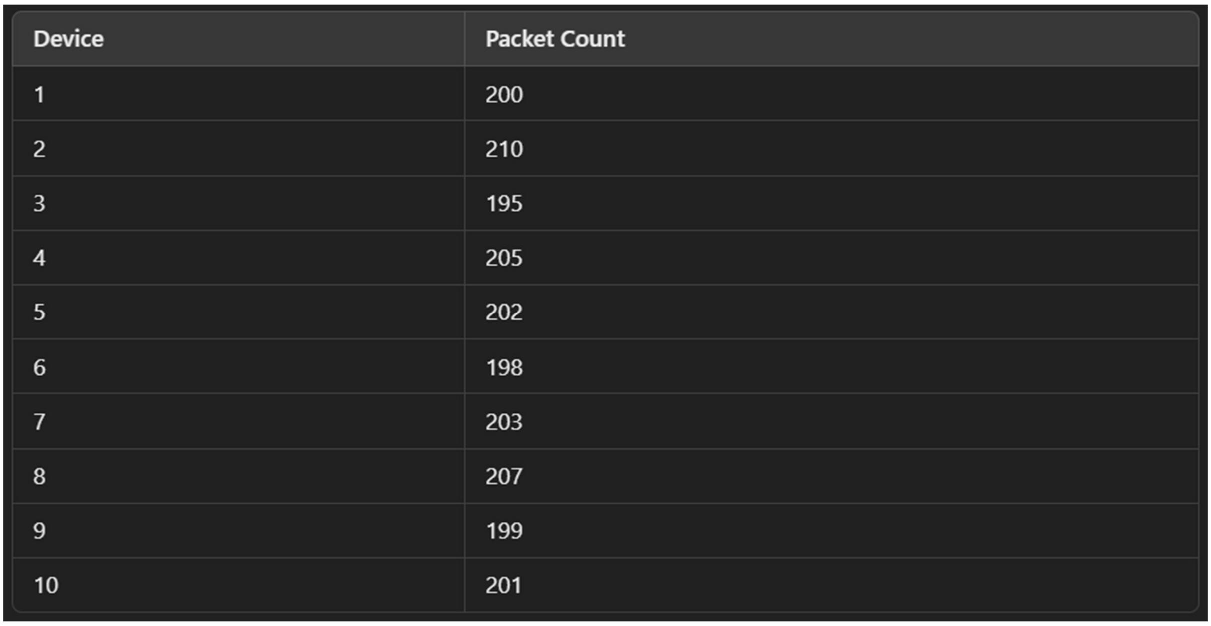


Aspect	IID	Non-IID
Independence	All samples are independent of each other.	Some samples may be dependent on others (e.g., sequential or time-dependent data).
Identical Distribution	All data comes from the same distribution (same behavior for all devices, times, conditions).	Different data subsets may follow different distributions (device-specific behavior, attack traffic vs. normal traffic).
Example	Benign traffic from all IoT devices treated as having the same behavior.	Attack traffic from a security camera differs from benign traffic, and from attack traffic on a thermostat.

Mathematical Example of IID and Non-IID for N-BaIoT Dataset:

1. IID (Independent and Identically Distributed) Example:

Let's assume you're collecting network traffic data (like packet counts) from 10 IoT devices in a smart home. The devices are operating under similar conditions, and none are compromised. Here's how the packet count data might look:



Device	Packet Count
1	200
2	210
3	195
4	205
5	202
6	198
7	203
8	207
9	199
10	201

In this scenario:

- **Identically Distributed:** All devices behave similarly, producing packet counts in a similar range (around 200 packets), with only minor variations due to randomness. All data points follow the same distribution (e.g., a normal distribution with mean 200).
- **Independent:** The packet count from one device does not affect the packet counts from the others. There's no correlation between the values.

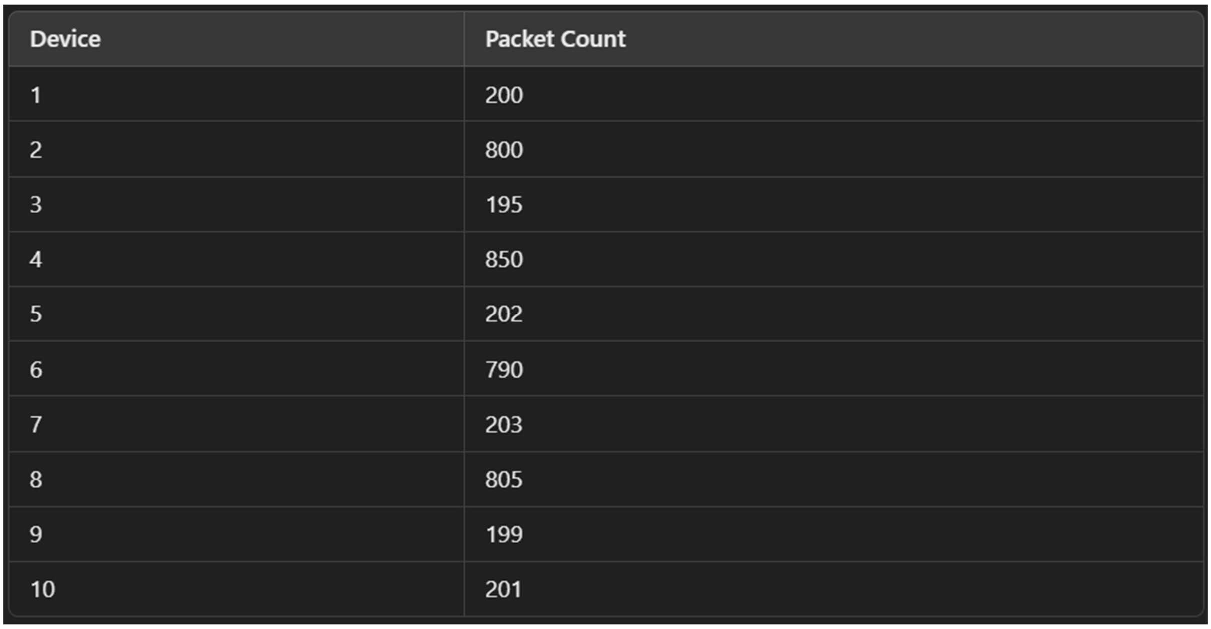
If you calculate basic statistics like the mean and variance, you'd expect them to be similar across all devices:

- **Mean of packet counts:** (200+210+195+⋯+201)/10=202
- **Variance of packet counts:** Relatively small, indicating low spread around the mean.

All observations are generated independently from the same distribution.

2. Non-IID (Non-Independent and Non-Identically Distributed) Example:

Now, consider a situation where 4 of the 10 IoT devices are under attack, and the compromised devices are sending an unusually high number of packets. Here's the modified packet count data:



Device	Packet Count
1	200
2	800
3	195
4	850
5	202
6	790
7	203
8	805
9	199
10	201

In this scenario:

- **Not Identically Distributed:** The devices are not following the same distribution anymore. Some devices (2, 4, 6, 8) are compromised and generating far more packets (in the range of 790–850), while the rest of the devices (1, 3, 5, 7, 9, 10) are behaving normally (with packet counts around 200). These two groups clearly belong to different distributions.
    - Normal devices follow one distribution (e.g., mean around 200).
    - Compromised devices follow another distribution (e.g., mean around 800).
- **Not Independent:** There might be a relationship between the compromised devices, such as a coordinated attack or network congestion. For instance, the high packet counts of devices 2, 4, 6, and 8 might be correlated due to the attack, meaning the packet counts are no longer independent. A rise in traffic on one compromised device could affect traffic on the others.

In terms of statistics:

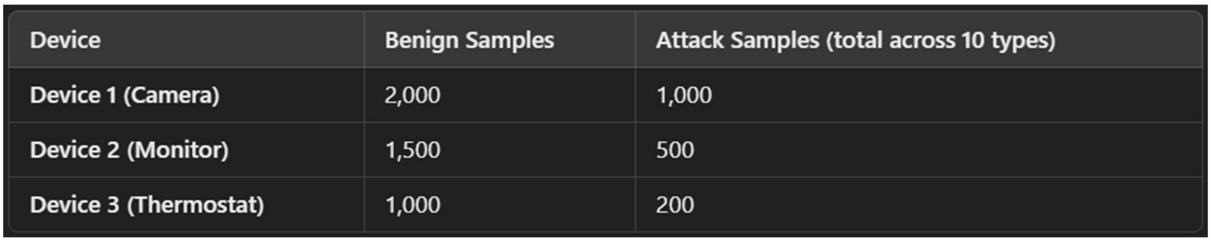
- Mean of normal devices: (200+195+202+203+199+201)/6=200
- Mean of compromised devices: (800+850+790+805)/4=811.25
- Variance: High, indicating a large spread between the two groups.

This separation into two distinct behaviours (normal and compromised) and the possible correlation between the compromised devices makes the data non-IID.

2. Mathematical example for IID and Non-IID for N-BaIoT Dataset:

Step 1: Defining the Dataset

Assume the following numbers of samples for 3 devices in the N-BaIoT dataset:



Device	Benign Samples	Attack Samples (total across 10 types)
Device 1 (Camera)	2,000	1,000
Device 2 (Monitor)	1,500	500
Device 3 (Thermostat)	1,000	200

□ **Total benign samples**: 2,000+1,500+1,000=4,500

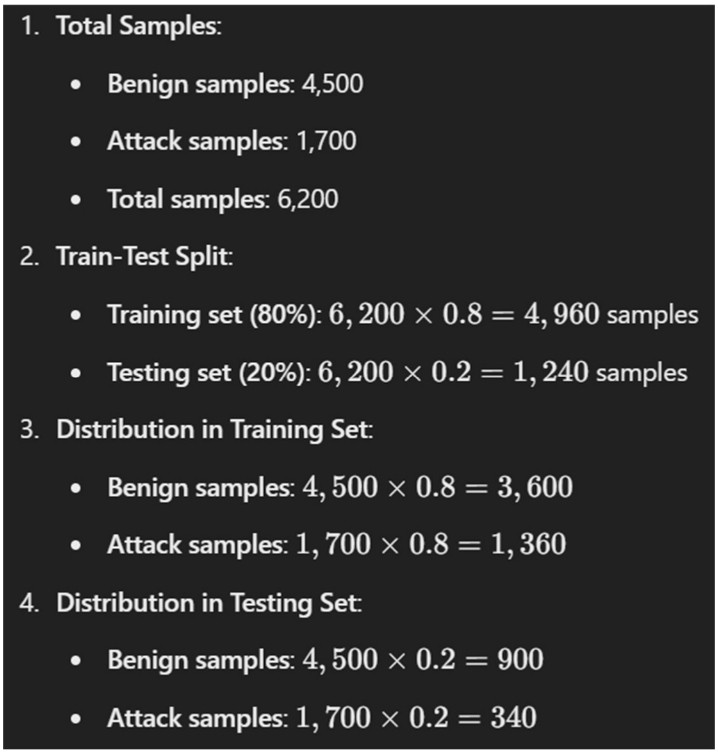
□ **Total attack samples**: 1,000+500+200=1,700

□ **Total samples**: 4,500+1,700=6,200

Step 2: IID Distribution

In an **IID (Independent and Identically Distributed)** setting, you mix all data samples together and treat them as if they come from the same distribution. Here's how you can calculate the distribution for training and testing sets:

Combining and Shuffling Data:



1. **Total Samples**:
   - **Benign samples**: 4,500
   - **Attack samples**: 1,700
   - **Total samples**: 6,200
2. **Train-Test Split**:
   - **Training set (80%)**: $6,200 \times 0.8 = 4,960$ samples
   - **Testing set (20%)**: $6,200 \times 0.2 = 1,240$ samples
3. **Distribution in Training Set**:
   - **Benign samples**: $4,500 \times 0.8 = 3,600$
   - **Attack samples**: $1,700 \times 0.8 = 1,360$
4. **Distribution in Testing Set**:
   - **Benign samples**: $4,500 \times 0.2 = 900$
   - **Attack samples**: $1,700 \times 0.2 = 340$

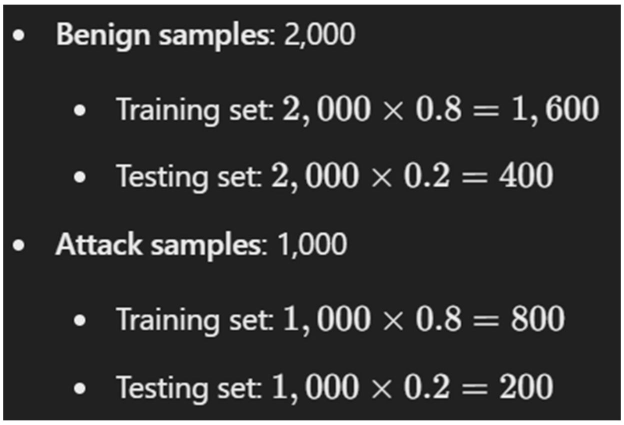
In IID distribution, all samples are mixed and shuffled, so each sample in the training and testing set is randomly drawn from the combined pool of benign and attack traffic.

Step 3: Non-IID Distribution

In a **non-IID** setting, we handle each device's data separately, maintaining their unique distributions. Here's how to calculate the train-test split for each device:

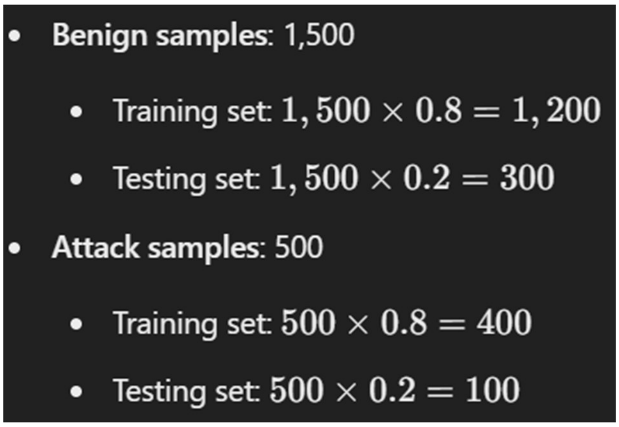
Device-Specific Distribution

Device 1 (Camera):



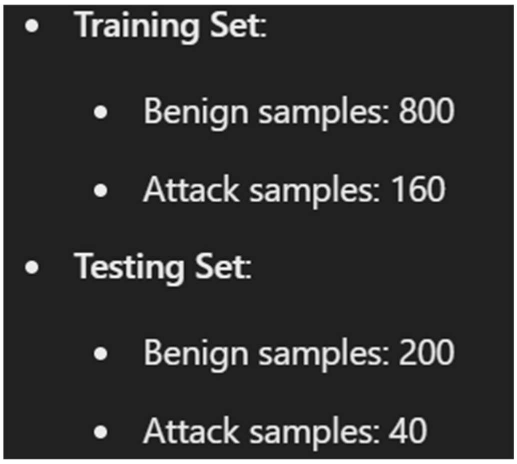
- **Benign samples**: 2,000
  - Training set: $2,000 \times 0.8 = 1,600$
  - Testing set: $2,000 \times 0.2 = 400$
- **Attack samples**: 1,000
  - Training set: $1,000 \times 0.8 = 800$
  - Testing set: $1,000 \times 0.2 = 200$

Device 2 (Monitor):



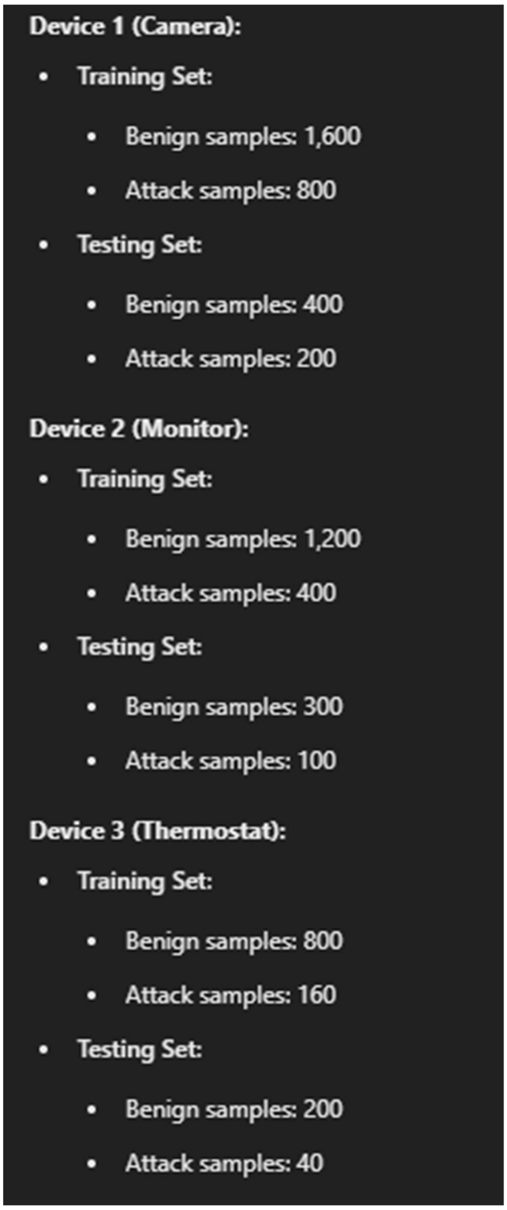
- **Benign samples**: 1,500
  - Training set: $1,500 \times 0.8 = 1,200$
  - Testing set: $1,500 \times 0.2 = 300$
- **Attack samples**: 500
  - Training set: $500 \times 0.8 = 400$
  - Testing set: $500 \times 0.2 = 100$

Device 3 (Thermostat):



- **Training Set**:
  - Benign samples: 800
  - Attack samples: 160
- **Testing Set**:
  - Benign samples: 200
  - Attack samples: 40

Summary of Non-IID Distribution:



Device 1 (Camera):

- **Training Set:**
  - Benign samples: 1,600
  - Attack samples: 800
- **Testing Set:**
  - Benign samples: 400
  - Attack samples: 200

Device 2 (Monitor):

- **Training Set:**
  - Benign samples: 1,200
  - Attack samples: 400
- **Testing Set:**
  - Benign samples: 300
  - Attack samples: 100

Device 3 (Thermostat):

- **Training Set:**
  - Benign samples: 800
  - Attack samples: 160
- **Testing Set:**
  - Benign samples: 200
  - Attack samples: 40

Comparison of IID and Non-IID

- **IID Distribution**: All data points are mixed and shuffled together, treating all samples as if they come from the same distribution, regardless of the device or attack type. The training and testing sets have a proportional representation of benign and attack samples from the entire dataset.
- **Non-IID Distribution**: Each device’s data is handled separately, preserving the unique characteristics of traffic and attacks specific to each device. The training and testing sets reflect the specific distribution of benign and attack samples for each device individually.

Applying IID and Non-IID to the N-BaIoT Dataset:

The **N-BaIoT dataset** contains network traffic from 10 different IoT devices (like security cameras, baby monitors, thermostats, etc.). Each device has two types of data:

1. **Benign traffic** (normal, non-attack traffic).
2. **Attack traffic** from 10 different types of attacks (e.g., Denial of Service (DoS), scanning, injection attacks).

For each device, you get one file of benign traffic and 10 files of attack traffic.

IID Distribution in the N-BaIoT Dataset:

In an **IID** scenario, the assumption is that:

- **All traffic samples (both benign and attack) are treated as if they come from the same statistical distribution**.
- **The data is independent**: Traffic from different time periods or different devices does not affect each other.

How Data is Distributed in IID for N-BaIoT:

- You randomly shuffle all the data (both benign and attack) across all 10 devices, treating them as if they are identical.
- You assume that the traffic from a **security camera** is statistically similar to the traffic from a **thermostat**.
- Attack traffic and benign traffic are mixed and treated as if they follow the same overall distribution. The classifier or model doesn't know the difference between device types or attack types.

This could lead to a model assuming that traffic from different devices behaves in the same way, which is not likely to reflect reality because:

- Different devices have unique traffic patterns. For instance, a security camera generates more consistent data (video streaming), while a thermostat sends sparse updates.
- Attack traffic is typically quite different from benign traffic. Treating them as identical would reduce the model's ability to distinguish attacks.

What Happens in IID:

- You are essentially mixing up all the traffic and pretending it's all the same, leading to the risk that the model doesn't learn the important differences between devices or between normal and attack traffic.
- The dataset looks homogenous, but in reality, IoT devices behave very differently.

Non-IID Distribution in the N-BaIoT Dataset:

In a **non-IID** scenario, the assumption is that:

- **Traffic from different devices (or even different attack types) comes from different distributions**. The traffic behavior from a security camera is different from that of a thermostat or baby monitor.
- **Traffic samples may not be independent**. For instance, during an ongoing attack, the attack traffic is likely to have some temporal dependence. The traffic patterns during an attack might influence subsequent traffic patterns.

How Data is Distributed in Non-IID for N-BaIoT:

- The data is grouped by device and attack type, recognizing that each device behaves differently.
- For example, a **baby monitor** has different network behavior (light and periodic traffic) compared to a **security camera** (which has continuous streaming). Attack traffic also looks different from benign traffic for each device.

Here's how the data could look:

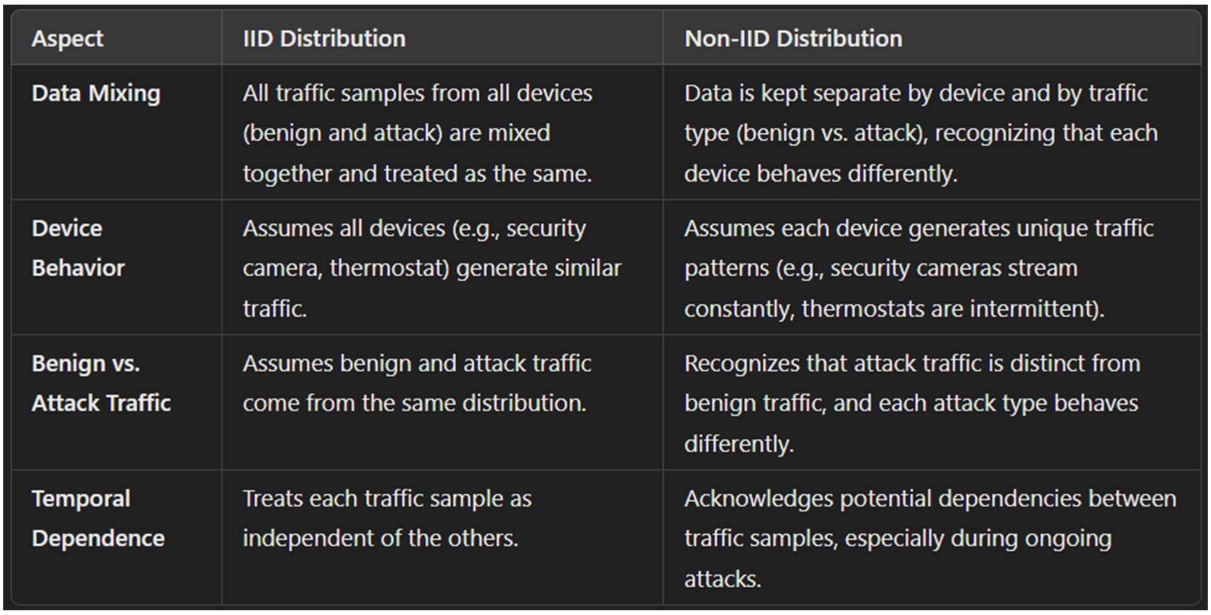
- **Device 1 (Security Camera)**:
  - One file contains benign traffic.
  - 10 files contain traffic from different attacks (e.g., DoS, injection).
- **Device 2 (Thermostat)**:
  - One file contains benign traffic (short, sporadic network packets).
  - 10 files contain attack traffic (again, unique attack patterns).

Each device has distinct benign and attack traffic, and these differences are not ignored.

What Happens in Non-IID:

- You respect the differences between devices. This allows a machine learning model to learn that the **security camera's normal traffic looks very different from the thermostat's normal traffic**.
- You also separate attack traffic from benign traffic, so the model can focus on the differences between normal behaviour and attack patterns.
- There may be dependencies between samples (e.g., in attack scenarios), and the model can capture these patterns.

IID vs. Non-IID in the N-BaIoT Dataset:



Aspect	IID Distribution	Non-IID Distribution
Data Mixing	All traffic samples from all devices (benign and attack) are mixed together and treated as the same.	Data is kept separate by device and by traffic type (benign vs. attack), recognizing that each device behaves differently.
Device Behavior	Assumes all devices (e.g., security camera, thermostat) generate similar traffic.	Assumes each device generates unique traffic patterns (e.g., security cameras stream constantly, thermostats are intermittent).
Benign vs. Attack Traffic	Assumes benign and attack traffic come from the same distribution.	Recognizes that attack traffic is distinct from benign traffic, and each attack type behaves differently.
Temporal Dependence	Treats each traffic sample as independent of the others.	Acknowledges potential dependencies between traffic samples, especially during ongoing attacks.

Conclusion:

- **IID Data:** In an IID scenario, all IoT devices behave similarly and independently, following the same statistical distribution. This represents a normal operation where each device's data is unaffected by others, making the data consistent and predictable across the network.
- **Non-IID Data:** In a non-IID scenario, different devices show varied behaviours, often due to anomalies like cyberattacks. Here, some devices generate significantly different traffic patterns (e.g., compromised devices), and there may be dependencies between them, making the data heterogeneous and correlated.

In summary, IID data reflects normal, uniform operation, while non-IID data indicates varying behaviour, such as device compromise or unusual activity, highlighting the need for careful analysis when detecting anomalies in the N-BaIoT dataset.