Schema Comparison: NSL-KDD vs CICIDS2017

---

1. Dataset Background

- **NSL-KDD**
  - Derived from **KDD'99**, which was originally built from the 1998 DARPA Intrusion Detection Evaluation Program.
  - Contains ~125,000 training instances and ~22,000 testing instances.
  - Each record describes a **connection/session** with 42 features.
  - Focuses on classical attack families: **DoS, Probe, R2L, U2R**.
  - Known limitations: synthetic, outdated, fewer modern attack vectors.
- **CICIDS2017**
  - Created by **Canadian Institute for Cybersecurity (CIC)**.
  - Captured over **5 days** of realistic traffic (July 3–7, 2017).
  - Contains ~2.8 million flows (labeled as benign or specific attack).
  - Features extracted using **CICFlowMeter**, producing 79–85 attributes per flow.
  - Covers modern attack types: **DDoS, Botnet, Web Attacks, Brute-force, Heartbleed, Infiltration, Port Scans**.

---

2. Schema Structure

Aspect	**NSL-KDD**	**CICIDS2017**
Feature Count	42 (plus label & category)	79–85 (depending on preprocessing)
Feature Types	Mixed: • Categorical (protocol, service, flag) • Numerical • Binary	Almost all Numerical Few Binary flag counters One Categorical (label)
Semantic Categories	• Traffic (protocol, bytes) • Behavior (logins, file ops) • Privilege/Auth • Traffic Statistics	• Flow Characteristics (duration, bytes, packets) • Packet Behavior (sizes, IATs) • TCP Flags • Content/Timing (active, idle)
Labeling	outcome (fine-grained) + category (coarse: DoS, Probe, R2L, U2R, Normal)	Modern, diverse: Benign, DoS, DDoS, Botnet, Brute-force, Heartbleed, Infiltration, Web Attacks

Time Context	No explicit timestamps	Includes flow durations, IATs, active/idle stats (time-sensitive)
Dataset Size	~150k instances	~2.8M instances
Collection Method	Synthetic simulation	Realistic network capture (mixed benign & attack traffic)

---

3. Example Features by Category

NSL-KDD Features (42 total)

1. **Traffic (basic connection attributes):**
   - duration, protocol_type, service, flag, src_bytes, dst_bytes
2. **Content-based features (behavioral):**
   - hot, num_failed_logins, logged_in, num_file_creations, num_shells
3. **Host-based statistical features:**
   - count, srv_count, serror_rate, srv_serror_rate, same_srv_rate, diff_srv_rate
4. **Privilege/Auth features:**
   - root_shell, su_attempted, is_guest_login

CICIDS2017 Features (79–85 total)

1. **Flow Characteristics:**
   - Flow Duration, Total Fwd Packets, Total Backward Packets
   - Fwd Packets Length Max/Min/Mean/Std, Bwd Packets Length Mean
2. **Time-based Features:**
   - Flow IAT Mean, Flow IAT Std, Fwd IAT Max, Bwd IAT Min
   - Active Mean, Active Std, Idle Mean, Idle Max
3. **TCP Flags:**
   - Fwd PSH Flags, Bwd PSH Flags, Fwd URG Flags, Bwd URG Flags
4. **Packet/Byte Behavior:**
   - Average Packet Size, Packet Length Variance
   - Fwd Header Length, Bwd Header Length
5. **Content Ratios:**
   - Down/Up Ratio, Avg Fwd Segment Size, Avg Bwd Segment Size

---

4. Preprocessing Implications

Step	NSL-KDD	CICIDS2017
Categorical Encoding	Required (protocol_type, service, flag)	Minimal (only label, all other features numerical)
Normalization	Needed for skewed counts (src_bytes, etc.)	Needed (features have different scales, many are skewed)
Dimensionality	Low (42 features)	High (79–85 features, feature selection may be needed)
Redundant Features	Some constants (num_outbound_cmds, is_host_login)	Few redundant, but many correlated features

---

5. Key Differences

- **Scope**:
  NSL-KDD focuses on **connection/session statistics**.
  CICIDS2017 captures **flow-level dynamics**, **packet timings**, and **flag counters**.

- **Modernity**:
  NSL-KDD → outdated (attacks from the 90s).
  CICIDS2017 → modern attack scenarios (botnets, web, brute force).

- **Granularity**:
  NSL-KDD is coarse (session-based, limited context).
  CICIDS2017 has fine-grained **flow + packet timing features**, better suited for ML/DL.