Feature Summary

Feature	Formula / Definition	Scale?	Benefit
rate_ratio	(src_rate + ε) / (dst_rate + ε)	Yes	Highlights asymmetric traffic patterns
syn_to_ack	(syn_packets + 1) / (ack_packets + 1)	Yes	Flags incomplete handshakes (scans)
rst_to_fin	(rst_packets + 1) / (fin_packets + 1)	Yes	Distinguishes abortive vs. clean teardowns
avg_pkt_size	(overall_rate × flow_time) / total_packets	Yes	Captures average payload size per packet
mean_interpkt	Mean of inter-packet arrival times	Yes	Detects uniform vs. bursty traffic
std_interpkt	Std. dev. of inter-packet arrival times	Yes	Measures consistency of packet spacing
p90_interpkt	90th percentile of inter-packet times	Yes	Spots occasional long gaps vs. floods
burstiness	max(interpkt) / mean(interpkt)	Yes	High in volumetric flood attacks
payload_entropy	Shannon entropy of payload byte histogram	Yes	Differentiates encrypted/compressed vs. repetitive payloads
value_range	max_value – min_value	Yes	Captures payload variability
flows_last_10s	Rolling count of flows by source IP in 10 s	Yes	Detects fast-scanning or botnet behavior
unique_dsts_last_10s	Rolling count of unique dst IPs/ports in 10 s	Yes	Quantifies scanning breadth
hour_sin	sin(2π·hour_of_day/24)	Yes	Encodes cyclical time-of-day patterns
hour_cos	cos(2π·hour_of_day/24)	Yes	Encodes cyclical time-of-day patterns
handshake_complete	1 if full SYN→SYN-ACK→ACK seen, else 0	No	Flags half-open (SYN flood) attempts
abrupt_reset	1 if rst_flags=1 AND fin_flags=0, else 0	No	Captures abrupt, reset-only connection drops

Feature	Formula / Definition	Scale?	Benefit
tcp_syn_ratio	syn_flags * protocol_tcp	No	Emphasizes SYN usage specifically in TCP flows
udp_psh	psh_flags * protocol_udp	No	Highlights PSH behavior in UDP traffic

---

Detailed Descriptions

Continuous Features (Scale)

- **rate_ratio**
  - **Formula**: (src_rate+ε)/(dst_rate+ε)(\mathrm{src_rate} + \varepsilon) / (\mathrm{dst_rate} + \varepsilon)
  - **Scaling**: Yes
  - **Benefit**: Amplifies asymmetric flows (e.g., reflection attacks) by comparing source vs. destination throughput.
- **syn_to_ack**
  - **Formula**: (syn_packets+1)/(ack_packets+1)(\mathrm{syn_packets} + 1) / (\mathrm{ack_packets} + 1)
  - **Scaling**: Yes
  - **Benefit**: Flags flows with many SYNs but few ACKs—typical of port scans or half-open connection attempts.
- **rst_to_fin**
  - **Formula**: (rst_packets+1)/(fin_packets+1)(\mathrm{rst_packets} + 1) / (\mathrm{fin_packets} + 1)
  - **Scaling**: Yes
  - **Benefit**: Distinguishes abortive connections (high RST) from clean teardowns (high FIN).
- **avg_pkt_size**
  - **Formula**: (overall_rate×flow_time)/total_packets(\mathrm{overall_rate} \times \mathrm{flow_time}) / \mathrm{total_packets}
  - **Scaling**: Yes
  - **Benefit**: Captures typical payload size per packet—different for volumetric floods vs. interactive traffic.
- **mean_interpkt**, **std_interpkt**, **p90_interpkt**

    - **Definition**: Summary statistics (mean, standard deviation, 90th percentile) of the time gaps between successive packets.
    - **Scaling**: Yes
    - **Benefit**: Helps differentiate uniform packet streams (e.g., botnet traffic) from bursty, human-driven flows.
- **burstiness**
    - **Formula**: max(interpkt)/interpkt¯\max(\mathrm{interpkt}) / \overline{\mathrm{interpkt}}
    - **Scaling**: Yes
    - **Benefit**: Very high in DDoS floods (sharp bursts), lower in steady or interactive sessions.
- **payload_entropy**
    - **Definition**: Shannon entropy of the histogram of payload byte values.
    - **Scaling**: Yes
    - **Benefit**: High entropy indicates encrypted/compressed content; low entropy suggests repetitive flood packets.
- **value_range**
    - **Formula**: max_value−min_value\max_value - \min_value
    - **Scaling**: Yes
    - **Benefit**: Measures spread of payload bytes—small in ping floods, larger in application-layer attacks.
- **flows_last_10s**, **unique_dsts_last_10s**
    - **Definition**: Rolling counts over a 10-second window per source IP—total flows and unique destinations/ports.
    - **Scaling**: Yes
    - **Benefit**: Detects rapid scanning or botnet behavior by volume and breadth of connection attempts.
- **hour_sin**, **hour_cos**
    - **Formula**: sin(2π×hour/24)\sin(2\pi \times \text{hour}/24), cos(2π×hour/24)\cos(2\pi \times \text{hour}/24)
    - **Scaling**: Yes
    - **Benefit**: Encodes cyclical daily patterns without artificial boundaries between midnight and 0h.

Binary Features (No Scale)

- **handshake_complete**
    - **Definition**: 1 if a full SYN→SYN-ACK→ACK handshake was observed, else 0.
    - **Scaling**: No

    - **Benefit**: Flags half-open (SYN flood) flows lacking complete handshakes.
- **abrupt_reset**
    - **Definition**: 1 if `rst_flags == 1` **and** `fin_flags == 0`, else 0.
    - **Scaling**: No
    - **Benefit**: Captures abrupt, reset-only connection terminations common in scans or attacks.
- **tcp_syn_ratio**
    - **Definition**: `syn_flags * protocol_tcp` (product of two binary indicators).
    - **Scaling**: No
    - **Benefit**: Emphasizes SYN usage specifically in TCP traffic, ignoring other protocols.
- **udp_psh**
    - **Definition**: `psh_flags * protocol_udp`.
    - **Scaling**: No
    - **Benefit**: Highlights PSH-like behavior in UDP flows, which is atypical and can signal anomalies.

Pipeline Note: Compute all above features **after** raw-data cleaning and outlier handling, but **before** any scaling or normalization steps. Then fit your scaler(s) on training data (including these continuous features) and apply to both train and test sets.