Explainable AI-Driven Zero-Trust Anomaly Detection for Encrypted Traffic

J.K. Wanasinghe*, M.C.S.J. Perera, A.R.S.S. Wijewardhana, S.N. Karunarathna, U. Jayasinghe

University of Peradeniya, Peradeniya, 20400, Sri Lanka
**e20420@eng.pdn.ac.lk*

Modern network attacks increasingly exploit encrypted channels such as TLS and HTTPS, rendering traditional deep packet inspection ineffective. Existing intrusion detection systems achieve high accuracy but incur latency unsuitable for real-time software-defined networking (SDN) deployment, and provide limited attribution for automated blocking decisions required by zero-trust architectures. Prior work has advanced encrypted-traffic detection via deep autoencoders, hybrid CNN architectures, contrastive learning, and SHAP-based classifiers, yet each trade off encrypted-traffic compatibility, latency, explainability, or real-capture validation; jointly satisfying all four remains comparatively unaddressed. This paper presents a two-stage anomaly detection pipeline built around this combination, operating exclusively on IP and TCP packet header metadata without payload decryption. The design was shaped by comparative experimentation on the BCCC Darknet (25,588 unlabelled samples) and the CIC-IDS-2017 (702,007 expert-annotated flows); only the latter, split into a stratified training set with a held-out real-world test set, trained and evaluated the final pipeline, with the Darknet experiments used solely to motivate its consensus-voting design. Stage 1 deploys a Base Check Classifier (BCC), a Decision Tree trained on 28 header-only features selected through a four-stage Gini-importance and correlation-based reduction from 79 CICFlowMeter features, achieving 99.964% attack recall at ~1.2ms per flow. Stage 2 analyses only BCC-flagged flows using Deep Dictionary Learning (DDL) and an Isolation Forest consensus voter, whose dual-model agreement reduces the false positive rate relative to either model alone. Explainability combines DDL-native per-feature reconstruction error decomposition, SHAP, and LIME, with cross-method agreement attributing every DROP decision. A novel PCAP-to-labeled-stream construction methodology maps 579,932 raw streams to expert-annotated ground truth via 5-tuple and timestamp matching, enabling direct evaluation on real packet captures (not pre-extracted features). On 3,000 such streams, the pipeline attains 96.69% accuracy, 99.31% attack recall, 91.28% precision, and a 0.69% false negative rate, matching or exceeding deep-learning and SHAP baselines on the same benchmark. Physical validation on a 3-computer inline bridge testbed enforcing iptables DROP-by-default confirms real-time zero-trust enforcement with Grafana monitoring. These results indicate that header-only two-stage detection with multi-method explainability is a measurably effective approach for zero-trust encrypted-traffic anomaly detection; generalisation to newer datasets and live traffic is the next validation step.

Keywords: encrypted traffic analysis, zero-trust architecture, explainable artificial intelligence, network anomaly detection, deep dictionary learning