Interpretable Detection of Encrypted Traffic Using SHAP-Based Feature Attribution

Jiwon Suh, Juwon Hong, Mose Gu and Jaehoon (Paul) Jeong
Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Republic of Korea
{sjw6136, hongju2024, rna0415, pauljeong}@skku.edu

***Abstract*—With the widespread adoption of encryption, payload-based intrusion detection has become infeasible, leaving a critical gap in identifying malicious activity within encrypted traffic. While machine learning approaches using flow-level metadata have emerged as promising alternatives, their black-box nature prevents analysts from understanding or trusting the results. In this paper, we propose an interpretable framework that not only detects malicious encrypted traffic but also explains *why* a session is flagged and how operators should respond. Our approach integrates a Random Forest classifier with SHAP (SHapley Additive Explanations) to provide both global insights into model behavior and local, per-session explanations. A domain-aware explanation engine further translates SHAP outputs into semantic interpretations and automated security recommendations, enabling even non-expert operators to understand causes and mitigation strategies. These results are delivered through an interactive dashboard that supports analyst-driven investigation and response. Experiments on a public encrypted traffic dataset demonstrate that our system maintains high detection accuracy (>99%) while bridging the gap between model prediction and human-centered decision-making.**

***Index Terms*—Encrypted Traffic, Security Threat Detection, Explainable AI, SHAP.**

I. Introduction

With the widespread adoption of encryption protocols such as TLS and HTTPS, the majority of modern Internet traffic is now encrypted. While encryption significantly enhances user privacy and data security, it renders traditional content-based Intrusion Detection Systems (IDS) less effective. IDS rely on Deep Packet Inspection (DPI), which becomes ineffective when payloads are encrypted.

To address this visibility gap, recent research has focused on Machine Learning (ML) models that operate on flow-level features, such as statistical summaries such as packet size distributions, timing intervals, and flow duration. These metadata-based approaches preserve privacy while offering indirect behavioral signals of a malicious activity. However, despite high detection performance, ML models often lack interpretability, making it difficult for security analysts to trust, validate, or act on their outputs. In high-stakes environments such as Security Operations Center (SOC), black-box models are rarely deployed without strong justification.

In this work, we introduce a modular framework that tightly integrates accurate ML-based detection with explainable AI (XAI) techniques to enable transparent and actionable insights for encrypted traffic. We use SHAP (SHapley Additive exPlanations) to achieve not only global interpretability by identifying influential features across the dataset, but also local interpretability by explaining specific classification outcomes.

To bridge the gap between statistical attribution and operational response, we design a domain-aware explanation engine that maps SHAP outputs to human-readable interpretations and recommended mitigation actions. These explanations are delivered through an interactive dashboard that supports analyst-driven investigation and response, tailored to the needs of frontline security analysts.

Our main contributions are as follows:

- **SHAP-based Detection Framework:** We design a modular framework for detecting malicious encrypted traffic using flow-level metadata. The framework integrates a Random Forest classifier with SHAP-based explanations, supporting both global and local interpretability.
- **Domain-Aware Explanation Engine:** We propose a rule-based engine that maps SHAP feature attributions to human-understandable conditions, semantic interpretations, and recommended response actions, thus bridging the gap between statistical outputs and operational decision-making.
- **Interactive Analyst Dashboard:** We implement an analyst-facing dashboard that visualizes local SHAP explanations through waterfall plots and contextual interpretation cards, providing actionable insights for analyst-driven investigation and response in SOCs.
- **Feedback Loop for Policy Design:** We demonstrate a feedback mechanism where global SHAP insights inform the creation of rule-based policies, which in turn enhance local explanations and guide security rule refinement.

II. Related Work

The rise of encrypted communication has spurred diverse approaches for traffic analysis under privacy-preserving constraints. Early efforts such as BlindBox [1] sought to enable DPI over encrypted data using advanced cryptographic primitives. While effective in controlled settings, these methods incur high latency and face scalability challenges due to computational complexity [2].

To overcome these limitations, subsequent research shifted focus toward flow-level metadata analysis. These methods leverage statistical attributes such as packet count, timing, and byte distributions to train supervised ML models for threat detection [3], [4]. A variety of classifiers, including deep learning approaches, have been applied with promising results [5]–

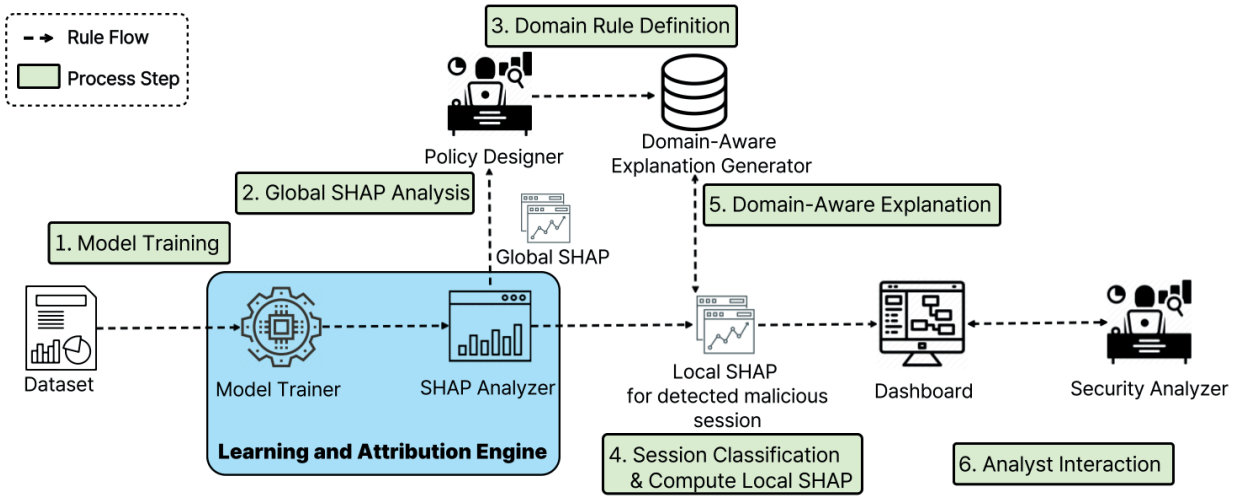

Fig. 1. Overview of the proposed SHAP-based interpretability framework.

[9]. However, these models typically function as black boxes, offering little visibility into their decision-making process. This lack of interpretability hinders adoption in operational security environments, where explainability is critical for trust, auditability, and actionable response.

To mitigate this challenge, XAI techniques have been introduced to extract insights from complex models. Among them, SHAP [10] has gained significant attention due to its game-theoretic foundation, consistency guarantees, and compatibility with tree-based models. In cybersecurity and IDS contexts, recent surveys and applications highlight both the promise and remaining challenges of explainable methods [11]–[14]. Prior works, such as Zeleke et al. [15] and Yang et al. [16], applied SHAP to identify globally important features in encrypted traffic detection or IDS. While valuable, most prior studies focus primarily on *global feature attribution*, overlooking the importance of *local explanations* that clarify individual predictions. This limitation is particularly problematic in SOC environments, where analysts must understand model outputs at the session level to support real-time incident triage and root-cause diagnosis [12], [13].

Our work builds on these efforts but addresses the aforementioned gaps by proposing an end-to-end (E2E) framework that integrates both global and local SHAP analyses for encrypted traffic detection. In addition, we introduce a domain-aware explanation engine that maps SHAP-derived features to semantically meaningful interpretations and recommended analyst responses. Finally, these insights are operationalized through an interactive dashboard tailored to SOC workflows, bridging the gap between model transparency and practical usability. This combination of high detection accuracy, interpretability at multiple levels, and analyst-oriented presentation distinguishes our work from prior studies.

III. Methodology

This section describes the architecture and operational pipeline of our proposed framework for interpretable encrypted traffic detection.

A. System Framework

Fig. 1 illustrates the E2E architecture of our SHAP-based detection and interpretation framework. The system is composed of three primary components—*Learning and Attribution Engine*, *Domain-Aware Explanation Generator*, and *Interactive Analyst Dashboard*—each serving a distinct function from model training to real-time analyst support.

a) Learning and Attribution Engine: This core module handles the training and explanation of the machine learning model. It includes:

- *Model Trainer* – It constructs a Random Forest classifier using flow-level traffic metadata. This model is optimized for high detection accuracy across encrypted sessions.
- *SHAP Analyzer* – It computes both global and local SHAP values for model predictions. Global values reveal overall feature importance across the dataset, while local values explain decisions at the session level.

b) Domain-Aware Explanation Generator: This module bridges the gap between raw SHAP outputs and actionable insights. Policy designers define rule-based mappings between feature name patterns and domain interpretations.

c) Interactive Dashboard for Analysts: This front-end interface is designed to present human-understandable justifications for each classification. It visualizes:

- SHAP waterfall plots that show feature contributions.
- Rule-driven interpretation cards with security context.
- Suggested actions based on session-level indicators.

The entire system operates in the following six stages:

1) **Model Training** – A Random Forest classifier is trained on preprocessed flow-level features extracted from encrypted traffic.
2) **Global SHAP Analysis** – SHAP values are averaged across the dataset to identify key features for model behavior analysis and policy formation.

3) **Domain Rule Definition** – Policy designers construct rules linking influential features to meaningful interpretations and response actions.
4) **Session Classification and Local SHAP Attribution** – Incoming traffic sessions are classified, and SHAP values are computed for sessions flagged as malicious.
5) **Contextual Explanation Generation** – Top SHAP features are matched to rules to generate natural-language interpretations and security recommendations.
6) **Analyst Interaction** – Explanations are delivered through the dashboard, allowing analysts to conduct real-time inspections and take proactive response steps.

This modular design ensures that the system can evolve independently across components over time. For example, security policy designers can add new rules without retraining the ML model, and when the model itself is retrained with updated data, it can be redeployed as a separate module without requiring any changes to the dashboard logic. This separation of concerns makes it easier to maintain and extend the system in dynamic threat environments..

IV. Experiment Setup

We evaluate our framework using a publicly available encrypted traffic dataset, followed by systematic preprocessing, classification model training, and SHAP-based explanation.

A. Dataset

Our experiments are based on the *Encrypted Traffic Feature Dataset* [17], which is specifically designed for flow-level analysis of encrypted sessions. The dataset aggregates traffic from six publicly available sources and includes both benign and malicious flows. Each flow instance is represented by 305 statistical features such as average packet size, total packet count, inter-arrival time, and TCP window behavior. The class distribution is approximately balanced, and all data is provided in structured CSV format.

B. Preprocessing

Before training, we applied several preprocessing steps to ensure data quality and model performance:

- **Normalization**: All numerical features were scaled to the [0,1] range using Min-Max scaling.
- **Feature Cleaning**: Constant-valued and null-valued features were removed to avoid redundancy and noise.
- **Train-Test Split**: The dataset was split into 70% training set and 30% testing set using stratified sampling to preserve class balance.

These steps ensure that the learning process is not biased by irrelevant or uninformative features, and that the evaluation results are statistically valid.

C. Classification Model

We used a Random Forest classifier for encrypted traffic detection, optimized via grid search with 5-fold cross-validation. The final model was evaluated on a held-out test set for both accuracy and interpretability.

D. Explainability with SHAP

To interpret model predictions, we use SHAP, which attributes a marginal contribution value to each feature for a given output. In our framework, SHAP operates at two levels:

- **Global Interpretability**: SHAP values are averaged across the test set to reveal the most influential features, guiding the construction of domain interpretation rules.
- **Local Interpretability**: SHAP is applied to individual sessions at inference time, enabling detailed reasoning behind each prediction.

These SHAP values are then passed to our rule-based explanation engine, which translates feature attributions into contextual insights and recommended actions for security analysts.

V. Evaluation and Interpretability

We evaluated our framework through experiments focusing on: (1) classification performance, (2) global feature attribution via SHAP, (3) domain rule construction, and (4) local interpretability with analyst support.

The Random Forest model showed strong performance across standard metrics, effectively distinguishing between benign and malicious encrypted traffic.

SHAP analysis revealed key flow-level features influencing the model's decisions, offering global interpretability.

Based on these attributions, policy designers define domain-aware rules to translate SHAP patterns into contextual interpretations and recommendations. Finally, our interactive dashboard delivers local explanations, such as SHAP waterfall plots and rule-driven insights that enable analysts to understand and act on individual predictions.

A. Classification Performance

TABLE I
Classification performance on the test set

Metric	Score (%)
Accuracy	99.84
Precision	99.94
Recall	99.74
F1-Score	99.84

The Random Forest classifier achieved strong results across all metrics, as shown in Table I. Accuracy reached 99.84%, indicating that nearly all test samples were correctly classified. Precision of 99.94% confirms that almost every flow flagged as malicious was indeed malicious, minimizing false positives. Recall of 99.74% shows that the vast majority of actual malicious sessions were detected, while the F1-score (99.84%) demonstrates a balanced and reliable performance. These results validate the robustness of our detection module as a foundation for subsequent explainability and decision support.

B. Global Interpretability via SHAP

To understand the decision-making process of the model at a global level, we applied SHAP to evaluate the importance of the characteristics in 1000 randomly selected test samples. Fig. 2 shows the summary plot of the top 10 most influential features, based on their average absolute SHAP values.

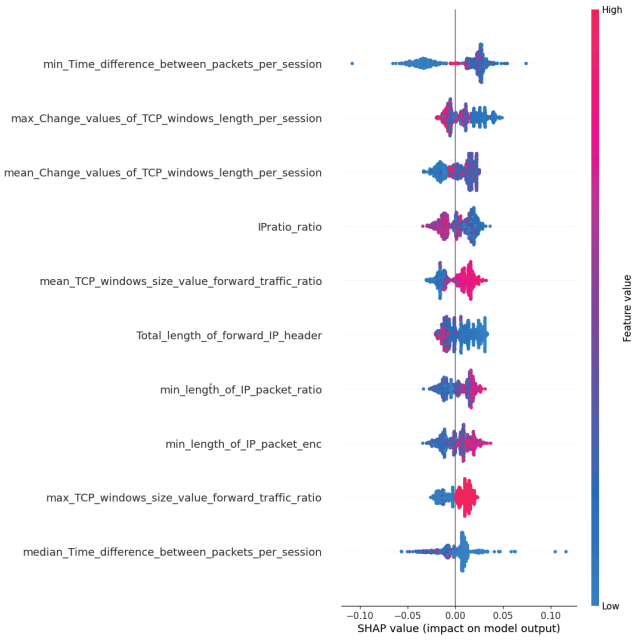

Fig. 2. SHAP summary plot for top-10 features

As shown in the figure, features related to interpacket timing and TCP window behavior emerged as highly influential for detecting malicious flows. While this summary offers a ranked view of feature importance, it lacks interpretability regarding specific feature values and their impact on classification decisions.

To address this limitation, we analyze SHAP dependence plots, which visualize how individual feature values affect model predictions. This reveals directional trends of which value ranges increase or decrease the likelihood of a malicious classification.

As seen in Fig. 3, midrange values (0.2 to 0.6) of *IPratio_ratio*—the ratio of inbound to outbound packets within a session—are associated with increased SHAP values, indicating that such asymmetric traffic patterns are likely to be malicious. Fig. 4 demonstrates that higher values of *min_length_of_IP_packet_ratio*, which measures the relative proportion of the shortest IP packet length compared to the flow average, also contribute strongly to malicious classifications. Similarly, Fig. 5 shows that sessions with high *maximum forward TCP window size ratios*, reflecting unusually large advertised TCP window sizes relative to normal flows, are frequently identified as malicious by the model.

To systematically translate these insights into operational policies, we implemented a token-based rule engine. Each

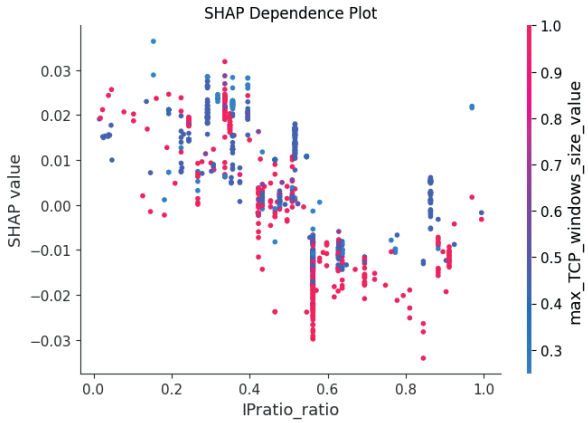

Fig. 3. SHAP dependence plot: *IPratio_ratio*

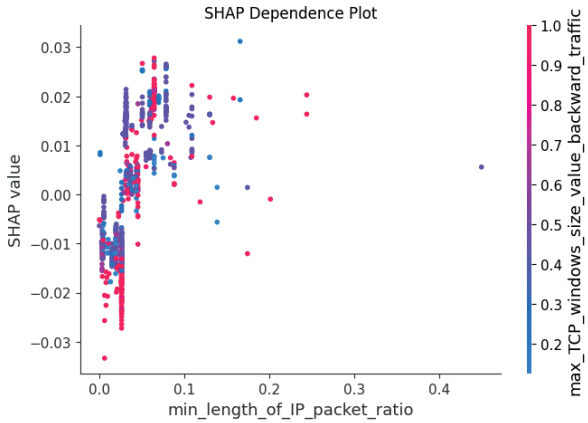

Fig. 4. SHAP dependence plot: *min_length_of_IP_packet_ratio*

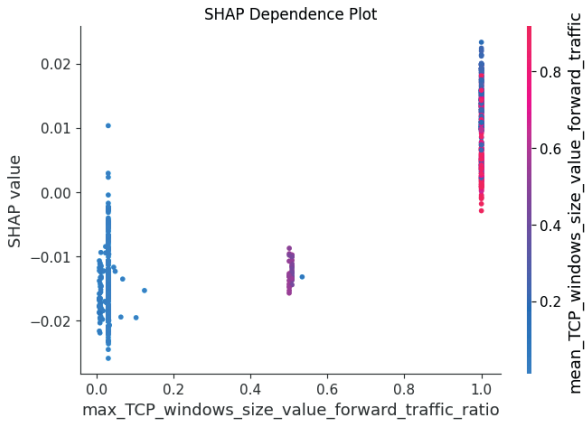

Fig. 5. SHAP dependence plot: *max_TCP_windows_size_value_forward_traffic_ratio*

feature name is tokenized and matched against a curated set of domain-informed rules. This process yields a standardized interpretation for each feature, including

- *Condition Hint* – The value ranges or thresholds that are security-relevant.
- *Interpretation* – What such feature behavior suggests in terms of network activity.

TABLE II
EXAMPLES OF DOMAIN-AWARE INTERPRETATION RULES

Feature Pattern	Condition Hint	Interpretation	Response Guide
IP Flow Direction Ratio	(0.3–0.6)	Highly imbalanced IP ration likely indicates probing or C2 patterns.	Analyze unidirectional or imbalanced flows for C2.
Minimum IP Packet Length Ratio	(> 0.07)	Correlates with C2 heartbeat or small control messages.	Apply TCP window filters and trigger C2 detection rules.
Maximum Forward TCP Window Size Ratio	($== 1.0$)	May indicate exfiltration bursts or evasive flooding.	Isolate the endpoint, throttle high-volume flows, and investigate for data leakage.

- *Response Guide* – Recommended security actions or policy implications.

This integration of SHAP visualization and rule-driven explanation transforms model behavior from a statistical artifact into a practical guide for designing transparent and enforceable detection policies.

C. Domain Rule Construction

To bridge from global feature attributions to actionable insights, we define a set of domain rules that map SHAP-identified patterns to semantic interpretations and response guidance. Table II shows representative rules derived from our global SHAP analysis.

D. Local Session Attribution and Analyst-Focused Interpretation

While global SHAP analysis provides an overall understanding of the model's behavior, effective threat mitigation often requires session-specific insights. To support this, we developed an interactive dashboard that delivers *local explanations* for individual sessions, allowing analysts to understand the precise rationale behind each classification decision.

Fig. 6 presents a SHAP-based waterfall plot for a selected malicious session (Session ID: 563). In this case, the features that contribute most to the prediction include *max_Change_values_of_TCP_windows_length_per_session* (the largest observed change in TCP window size within the session), *min_Time_difference_between_packets_per_session* (the minimum inter-arrival time between consecutive packets), and *IPratio_ratio* (the ratio of inbound to outbound packets in the session), each pushing the model output toward the malicious class.

To enhance interpretability for non-technical users, the dashboard also provides automatically generated domain-based interpretations to them. These interpretations are derived using a rule-based explanation engine that matches feature names to semantic patterns, delivering contextualized reasoning and actionable recommendations.

Rather than evaluating features in isolation, analysts are encouraged to synthesize these interpretations into higher-level hypotheses. In this example, the combination of automated packet timing, concentrated IP flow directionality, and TCP behavior anomalies may suggest the early stage of a command-and-control (C2) setup or an automated port scanning attack.

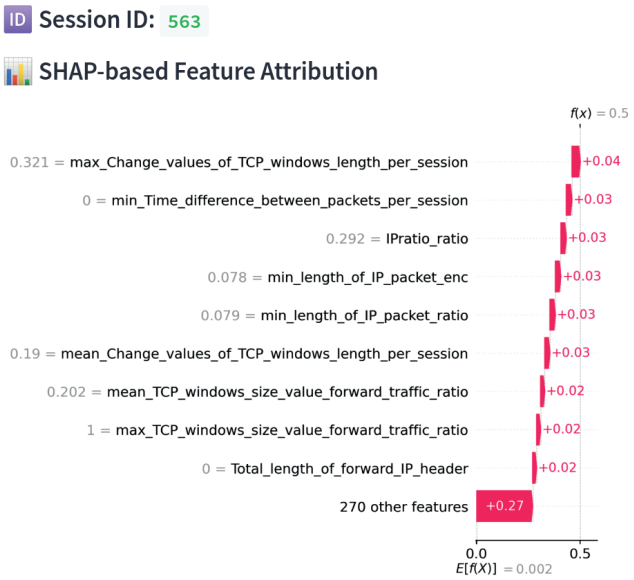

Fig. 6. Local SHAP waterfall plot for malicious session (ID: 563)

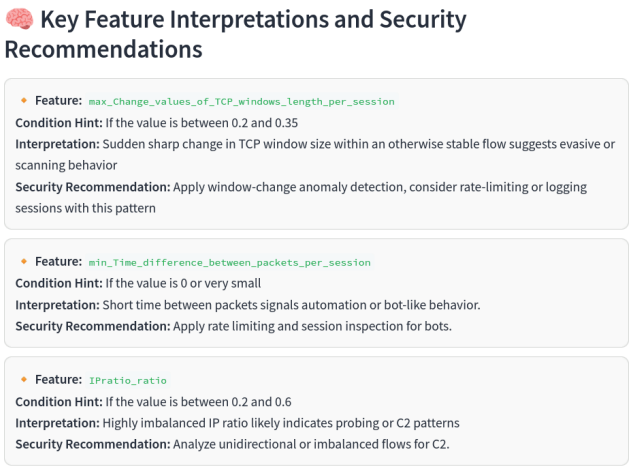

Fig. 7. Contextualized interpretation cards for top contributing features

By enabling such multi-feature reasoning, the dashboard helps bridge the gap between model prediction and operational response. Analysts are empowered to transition from passive alert triage to proactive decision-making, such as isolating the affected endpoint, applying TCP behavior filters, or triggering port-scan detection mechanisms.

TABLE III
REPRESENTATIVE SESSIONS WITH DOMAIN-AWARE INTERPRETATIONS

Session ID	Type	Top SHAP Features	Interpretation & Recommended Action
563	C2 Communication	max TCP win change, min pkt time diff, IPratio ratio	Frequent TCP window size shifts and low inter-packet timing suggest automated backchannel activity. Recommended: Isolate session, block source IP, and monitor for additional C2 traffic.
1742	Lateral Movement	IPratio ratio, flow duration, packet count	Sustained traffic to a concentrated set of internal IPs suggests lateral scanning or credential probing. Recommended: Monitor internal east-west traffic, apply micro-segmentation, and flag unusual flows.
1887	Exfiltration (Stealthy)	avg packet size, min pkt time diff, max fwd byte count	Low-volume but continuous packet flow may indicate stealthy data exfiltration over time. Recommended: Apply DLP policies, and alert on long-lived flows with an upload behavior.

To further demonstrate the versatility of our domain-aware explanation framework, we present a comparison of three representative sessions in Table III. For each session, we summarize the top SHAP-contributing features using human-readable labels, along with domain-specific interpretations and recommended mitigation strategies. This comparison illustrates how contextualized attribution supports scenario-specific decision-making and operational response planning.

Finally, the integration of global SHAP insights, rule-based domain interpretations, and local session-level explanations enables a natural feedback loop for policy refinement. Specifically, global feature attributions guide the design of interpretation rules, which are then applied to local explanations for individual sessions. Feedback from analysts on these local interpretations can subsequently inform updates to the rule set, thereby improving both the transparency and the operational usefulness of the framework over time.

VI. CONCLUSION

We presented an end-to-end framework for detecting and interpreting malicious encrypted traffic by combining machine learning, explainable AI, and domain-aware interpretation. Beyond high detection accuracy, our approach provides both global and local explanations, enabling actionable insights through an analyst-facing dashboard. Remaining challenges include the computational cost of SHAP, reliance on manual rule mappings, and limited evaluation scope. Future work will explore lightweight SHAP approximations, semi-automated rule generation, multi-class detection, and deployment in SOC environments. To encourage reproducibility, we release the full source code and dashboard at [18].

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea MSIT (No. 2023R1A2C2002990) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea MSIT (No. RS-2024-00398199). Note that Jaehoon(Paul) Jeong is the corresponding author.

REFERENCES

[1] J. Sherry, S. Hasan, C. Scott, A. Krishnamurthy, S. Ratnasamy, and V. Sekar, "Blindbox: Deep packet inspection over encrypted traffic," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication (SIGCOMM)*. ACM, 2015, pp. 213–226.

[2] A. Dainotti, A. Pescape, and K. C. Claffy, "Issues and future directions in traffic classification," *IEEE Network*, vol. 26, no. 1, pp. 35–40, 2012.

[3] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP)*, 2018.

[4] Z. Wang *et al.*, "Machine learning for encrypted malicious traffic detection: Approaches, datasets and comparative study," *arXiv preprint arXiv:2203.09332*, 2022.

[5] Z. Wang and V. L. L. Thing, "Feature mining for encrypted malicious traffic detection with deep learning and other machine learning algorithms," *arXiv preprint arXiv:2304.03691*, 2023.

[6] M. López-Martín, B. Carro, and A. Sánchez-Esguevillas, "Network traffic classification based on deep learning," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 28–33, 2017.

[7] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges," *Computer Networks*, vol. 161, pp. 83–94, 2019.

[8] M. H. Pathmaperuma *et al.*, "Deep learning for encrypted traffic classification: Benchmarks and insights," *Sensors*, vol. 22, no. 19, p. 7643, 2022.

[9] R. T. Elmaghraby *et al.*, "Encrypted network traffic classification based on machine learning," *Digital Communications and Networks*, 2024.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[11] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 22, no. 1, p. 18, 2020.

[12] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, "Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities," *arXiv preprint arXiv:2207.06236*, 2022.

[13] Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable artificial intelligence applications in cyber security: State-of-the-art in research," *arXiv preprint arXiv:2208.14937*, 2022. [Online]. Available: https://arxiv.org/abs/2208.14937

[14] M. Hosain, S. Rahman, A. Islam *et al.*, "Xai-xgboost: An innovative explainable intrusion detection approach for securing internet of medical things systems," *Journal of Network and Computer Applications*, vol. 238, p. 104918, 2025.

[15] S. N. Zeleke, A. F. Jember, and M. Bochicchio, "Integrating explainable ai for effective malware detection in encrypted network traffic," *arXiv preprint arXiv:2501.05387*, 2025. [Online]. Available: https://arxiv.org/abs/2501.05387

[16] J. Yang, D. Sharma, and S. Park, "Class-based shap analysis for improved explainability insights in nids," *IEEE Access*, vol. 11, pp. 15 876–15 890, 2023.

[17] M. U. Rehman and S. Y. Nam, "Encrypted traffic feature dataset," 2023. [Online]. Available: https://data.mendeley.com/datasets/xw7r4tt54g/1

[18] J. Suh, "Cluster-based detection of compromised sdn controller," GitHub Repository, available at: github.com/Seo6136/Cluster-Based-Detection-of-Compromised-SDN-Controller-Using-Behavioral-Divergence.