Harnessing Explainable AI to Detect Threats in DNS over HTTPS Traffic

Niaz Ashraf Khan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
niaz.ashraf@bracu.ac.bd

Md. Ferdous Bin Hafiz
Department of Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
ferdous.binhafiz@seu.edu.bd

Md. Mazharul Islam
Department of ECE
North South University
Dhaka, Bangladesh
mazharul.islam1@northsouth.edu

Md. Aktaruzzaman Pramanik
Department of CSE
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
aktaruzzaman.pramanik@ulab.edu.bd

Md. Ahsan Ullah
Department of CSE
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
ahsan.ullah@ulab.edu.bd

***Abstract*—DNS over HTTPS (DoH) was designed to improve user privacy by encrypting DNS queries. However, it has also created new ways for attackers to hide their activities and bypass traditional intrusion detection systems. This study delineates a comprehensive and methodologically rigorous machine learning pipeline for discerning nefarious DoH traffic, utilizing the CIRA-CIC-DoHBrw-2020 benchmark dataset. A heterogeneous ensemble of seven classifiers was systematically evaluated, wherein class imbalance was mitigated through SMOTE-based synthetic oversampling, and dimensionality reduction was achieved via correlation-informed feature pruning. Empirical evaluations revealed that tree-based ensemble methods consistently outperformed other paradigms, with XGBoost attaining flawless classification efficacy in terms of accuracy, precision, and recall, marginally surpassing Random Forest (99.999%) and LightGBM (99.998%). Conversely, conventional algorithms such as Logistic Regression (98.61%) and Gaussian Naive Bayes (97.82%) exhibited elevated false positive rates. Feature importance analysis revealed that temporal patterns (F13) and statistical flow metrics (F4) are key indicators of malicious activity. This research demonstrates that machine learning can effectively classify encrypted traffic and offers practical guidance for developing robust, privacy-conscious threat detection systems for enterprise and IoT settings.**

***Keywords*—DoH, HTTPS, XAI, Machine Learning, DNS**

I. Introduction

Historically, DNS requests were transmitted in plaintext, rendering them susceptible to pervasive surveillance, metadata harvesting, and content-based censorship. DNS over HTTPS (DoH) was introduced as a privacy-enhancing protocol that sends DNS queries through TLS-encrypted HTTP/2 channels to protect user confidentiality and hide query details [1] [2].

While DoH greatly limits what passive eavesdroppers and middleboxes can see by blocking deep packet inspection, it also creates serious challenges for defenders.The cryptographic encapsulation of DNS traffic obfuscates critical forensic indicators traditionally used to identify malicious activities, thereby diminishing the efficacy of legacy monitoring infrastructures [3] [4]. As a result, DoH can be misused by attackers for hidden activities such as command-and-control (C2) communication, data theft, and secretly managing botnets.

To address this hidden threat, we propose shifting from traditional packet inspection to behavioral analysis. Our framework focuses on patterns in metadata and transport-layer behavior, allowing detection without needing access to encrypted packet contents. By using a variety of machine learning methods, we reduce bias and improve the reliability of our classification. This content-agnostic approach avoids the limitations of signature-based systems and adapts more easily to evolving threats. We also incorporate interpretable machine learning techniques, such as SHAP and LIME, to explain model decisions and enhance trust in the system. Together, these elements support a more adaptable, transparent, and privacy-respecting solution for identifying malicious DoH traffic.

Our key contributions include:

1) **Comprehensive Model Evaluation:** We systematically benchmark seven heterogeneous learning paradigms, encompassing linear classifiers, tree-based learners, gradient-boosted ensembles, and neural architectures, to assess their comparative efficacy in encrypted DoH traffic detection.
2) **Balanced Learning Framework:** To rectify the skewed label distribution endemic in cyber-traffic datasets, we implement SMOTE-based oversampling to synthetically balance minority classes.
3) **Feature Engineering Insights:** We conduct rigorous correlation pruning and feature attribution analyses to isolate pivotal predictors, streamlining model complexity without compromising detection acuity.

Distinctive aspects of this work include:

1) **DoH-Centric Dataset Deployment:** We leverage the CIRA-CIC-DoHBrw-2020 corpus, a purpose-built dataset that encapsulates encrypted DNS activity across

multiple client applications and resolver endpoints, providing a nuanced representation of real-world DoH telemetry [5].

2) **Algorithmic Diversity:** Departing from monolithic classifier-centric studies, we embrace algorithmic diversity to furnish a panoramic evaluation of modeling strategies suited for the intricacies of encrypted traffic classification.
3) **Hermeneutics of Encrypted Flows:** By integrating interpretability mechanisms (SHAP and LIME), we deconstruct the decision logic of black-box models, thereby enabling actionable insights for cybersecurity practitioners [6], [7].
4) **Correlative Redundancy Mitigation:** Our preprocessing pipeline includes the excision of collinear features using variance inflation thresholds and Pearson correlation matrices, resulting in more parsimonious and generalizable models.

In summary, this research brings together encrypted traffic analysis and interpretable AI to provide a scalable, privacy-respecting detection approach. By focusing on model transparency and broad applicability, we offer new ways to detect and respond to hidden cyber threats enabled by DoH.

II. Related Work

The authors of [8] investigated DoH tunneling detection by leveraging statistical features such as packet size and duration, emphasizing the covert potential of encrypted traffic for data exfiltration. In a complementary direction, the authors of [9] employed feature reduction strategies alongside classical ML models to classify malicious DoH traffic effectively, achieving improved performance with minimal impact on accuracy.

Paper [10] introduced a Graph Convolutional Network (GCN)-based approach for edge prediction in encrypted DoH flows, attaining a notable accuracy of 96.29% and underscoring its applicability in IoT environments. The authors of [11] presented the CIC-Bell-DNS2021 dataset and highlighted the significance of advanced feature engineering for identifying diverse DNS-based threats, including spam, phishing, and malware, beyond binary classification.

In [12], a Self-Attention BiLSTM model was proposed to tackle the complexity of DoH traffic classification, outperforming baseline ML and DL models in multi-class tasks. Additionally, [13] explored SQL injection detection using AdaBoost and RPZ with DNS recursion traffic, enhancing Web Application Firewall (WAF) effectiveness through automated attack identification. Finally, the authors of [14] focused on memory-based malware detection by combining ensemble machine learning models and explainable AI (LIME).

Despite many existing studies, we utilize the specialized CIRA-CIC-DoHBrw-2020 dataset, conduct a comprehensive comparison of seven machine learning algorithms, and incorporate both SHAP and LIME explainability techniques to enhance the transparency and trustworthiness of our detection framework.

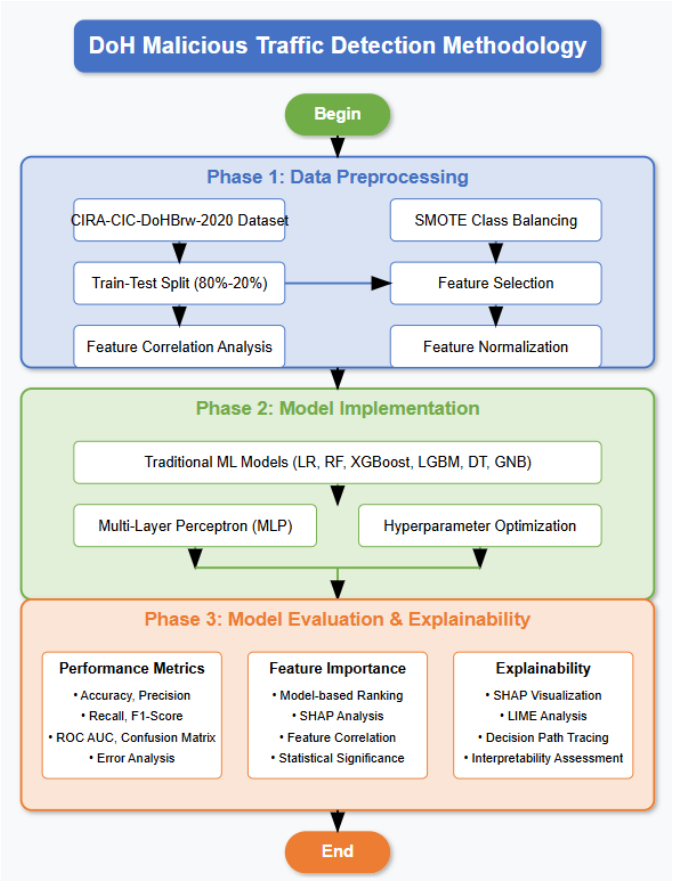


Fig. 1. Proposed methodology workflow for detecting malicious DNS over HTTPS traffic.

III. Methodology

A. Dataset Description

The dataset represents a comprehensive collection of both benign and malicious DNS over HTTPS traffic. This dataset includes traffic captured from various browsers and DoH providers under realistic conditions, with labeled flows for different attack scenarios including botnet communications, DDoS attacks, and data exfiltration. The dataset's diversity makes it particularly valuable for developing and evaluating detection mechanisms for encrypted DoH traffic. Fig. 1 shows the workflow of this research.

B. Data Preprocessing

Data preprocessing forms a crucial foundation for our machine learning pipeline. The process involves several key steps:

- **Class Imbalance Handling:** Network security datasets often suffer from significant class imbalance, with benign traffic vastly outnumbering malicious samples. Fig. 2 illustrates the class distribution before and after applying SMOTE. We apply the SMOTE [15] to address this issue:

$$X_{train\_sm}, y_{train\_sm} = \text{SMOTE}(X_{train}, y_{train}) \quad (1)$$

Fig. 2. Class distribution in the dataset before (top) and after applying SMOTE oversampling technique (bottom).

SMOTE creates synthetic examples of the minority class by:

$$x_{new} = x_i + \lambda \times (x_j - x_i) \tag{2}$$

- **Feature Correlation Analysis:** To reduce redundancy and multicollinearity, we compute the Pearson correlation coefficient as shown in Fig. 3 between all feature pairs:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

- **Feature Selection:** Features with correlation coefficients exceeding 0.9 are identified and removed from the dataset:

$$\mathcal{F}_{drop} = \{j \mid \exists i \neq j : |\rho_{i,j}| > 0.9\} \tag{4}$$

This resulted in the removal of 7 highly correlated features (indices 17, 18, 19, 20, 26, 28, 31) from the dataset.

C. Model Architecture and Development

To thoroughly evaluate the predictive power of different approaches, we developed and tested seven machine learning classifiers, each based on unique mathematical principles and learning biases:

- **Logistic Regression (LR):** A probabilistic discriminative model that estimates the conditional probability of binary outcomes by applying the sigmoid activation to a linear projection of input covariates:

$$\mathbb{P}(y = 1|\mathbf{z}) = \frac{1}{1 + \exp(-(\theta_0 + \sum_{k=1}^{n} \theta_k z_k))} \tag{5}$$

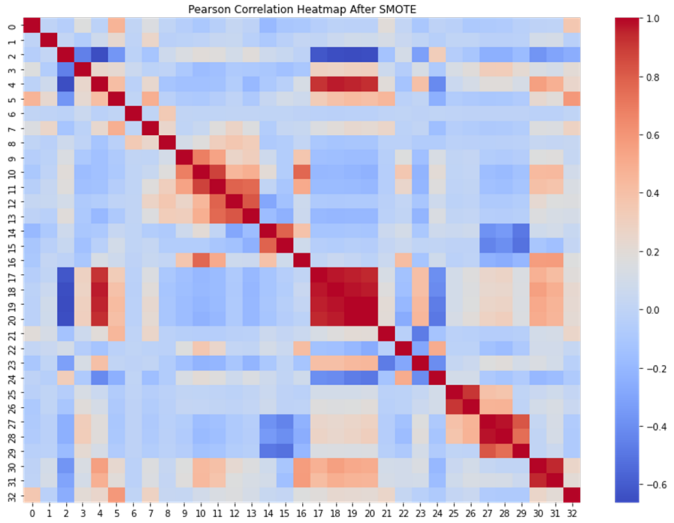


Fig. 3. Pearson correlation heatmap showing relationships between features. Highly correlated features (correlation coefficient greater than 0.9) were identified and removed to reduce dimensionality and multicollinearity.

- **Random Forest (RF):** An ensemble meta-estimator that aggregates predictions from an ensemble of randomized decision trees:

$$\hat{\zeta}_{RF} = \text{majority}\{\hat{\zeta}^{(1)}, \hat{\zeta}^{(2)}, ..., \hat{\zeta}^{(M)}\} \tag{6}$$

where $\hat{\zeta}^{(i)}$ represents the prediction from the i-th base estimator.

- **XGBoost (XGB):** A sequential additive model where each successive regressor minimizes the residual errors from prior learners:

$$\hat{\zeta}_j^{(s)} = \hat{\zeta}_j^{(s-1)} + \alpha \cdot h_s(\mathbf{z}_j) \tag{7}$$

Here, h_s is the s-th regression tree and α denotes the shrinkage coefficient.

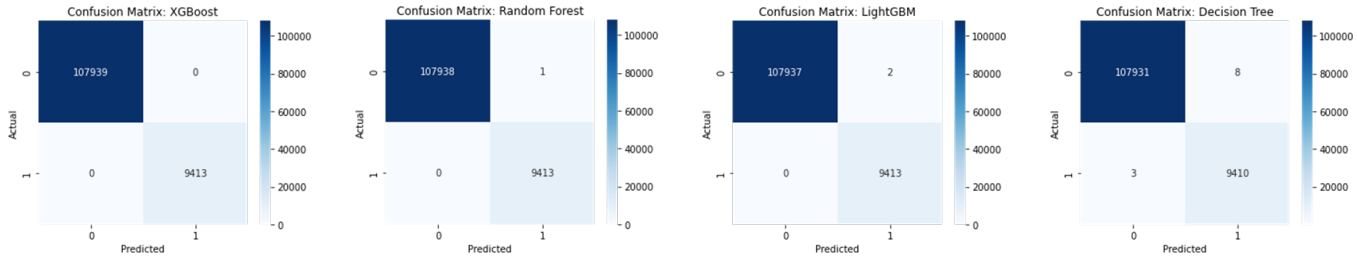
- **LightGBM (LGBM):** A histogram-based boosting framework optimized for speed and scalability, characterized by its leaf-wise tree growth strategy:

$$\mathcal{J}(\Theta) = \sum_{j=1}^{m} \ell(y_j, \hat{y}_j) + \sum_{q=1}^{Q} \mathcal{R}(g_q) \tag{8}$$

- **Multi-Layer Perceptron (MLP):** A hierarchical neural model comprising fully connected layers, with each hidden unit performing a nonlinear transformation of its inputs:

$$s_k = \sigma\left(\sum_{m=1}^{d} \omega_{km} z_m + \delta_k\right) \tag{9}$$

Here, σ denotes a nonlinear activation function, ω_{km} are weights, and δ_k is the bias parameter.



(a) XGBoost (b) Random Forest (c) LightGBM (d) Decision Tree

Fig. 4. Confusion matrices for ensemble and tree-based models: (a) XGBoost achieved perfect classification with no misclassifications, (b) RF with only 1 false positive, (c) LGBM with 2 false positives, and (d) DT with 8 false positives and 3 false negatives.

- **Decision Tree (DT):** A recursive partitioning algorithm that maps inputs to outputs via a tree structure, where internal nodes represent decision criteria and terminal nodes assign outcomes.

D. Interpreting Predictive Mechanisms

To interpret the decision processes of complex classifiers, we utilized two complementary explainability approaches:

- **Quantitative Input Influence (QII):** This attribution method, grounded in cooperative game theory, quantifies the marginal effect of each input variable by averaging its influence across feature coalitions [16]–[18].
- **LIME:** Locally Interpretable Model-agnostic Explanations approximates the behavior of the black-box function f around a target instance x^* using a sparse interpretable model $\hat{f}$:

$$\min_{\hat{f} \in \mathcal{F}} \mathcal{L}(f, \hat{f}, \pi_{x^*}) + \Omega(\hat{f}) \tag{10}$$

where $\mathcal{L}$ measures locality-aware fidelity using kernel π_{x^*}, and $\Omega(\hat{f})$ penalizes model complexity [7] [19] [20].

These interpretability tools provide both global and granular diagnostic insights, enabling domain experts to examine and justify the model's reasoning process, and uncover potential model biases or overfitting tendencies.

IV. Results and Discussion

A. Model Performance Comparison

The performance metrics for the seven implemented models are presented in Table I.

TABLE I
Accuracy of Machine Learning Models

Model	Accuracy
XGBoost	1.000000
Random Forest	0.999991
LightGBM	0.999983
Decision Tree	0.999906
MLP Neural Network	0.999386
Logistic Regression	0.986102
Naive Bayes	0.978177

The results demonstrate exceptional performance across all models, with ensemble-based approaches (XGBoost, Random Forest, and LightGBM) achieving nearly perfect or perfect accuracy. XGBoost achieved 100% accuracy with perfect precision and recall, indicating its superior capability in distinguishing between benign and malicious DoH traffic.

B. Confusion Matrix Analysis

Fig. 4 – Confusion matrices for the models: Fig. 4a shows XGBoost with perfect classification; Fig. 4b shows Random Forest; Fig. 4c shows LightGBM; and Fig. 4d shows Decision Tree.

Key observations from the confusion matrix analysis:

- XGBoost achieved perfect classification with no misclassifications.
- Random Forest and LightGBM showed excellent performance with only 1 and 2 false positives respectively. Their low false positive rates underscore their reliability in minimizing benign traffic misclassification, which is crucial for reducing alert fatigue in operational security systems.
- Traditional models (Logistic Regression and Naive Bayes) exhibited higher error rates, particularly in terms of false positives (1,428 and 2,306 respectively).

C. Model Interpretability Analysis

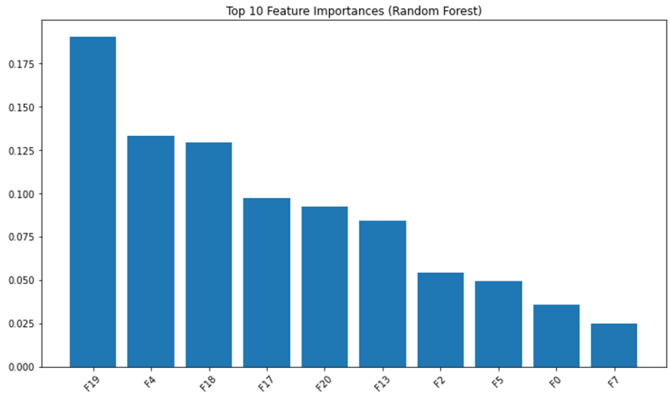


Fig. 5. Model-based feature importance analysis showing the relative contribution of different features to the classification decision.

Our feature importance and explainability analyses revealed several critical insights:

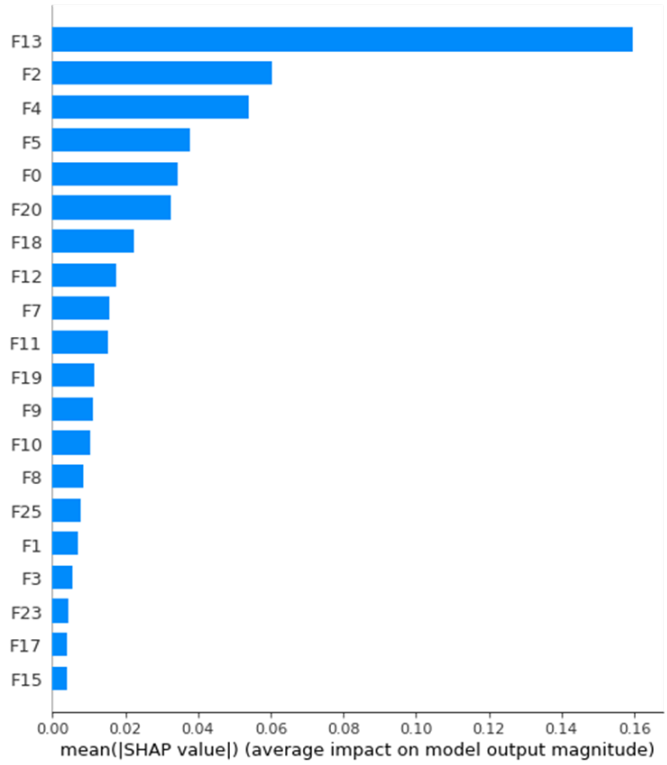


Fig. 6. SHAP Summary Plot

- Feature F13 emerged as the dominant predictor across multiple models, indicating its crucial role in distinguishing between benign and malicious DoH traffic, as shown in Fig. 5.
- Secondary features of importance included F2, F4, F5, F0, and F20, though with substantially lower contribution weights compared to F13.
- SHAP analysis confirmed F13's significance and revealed that for F4, higher values consistently pushed predictions toward the malicious class, providing strong discriminative power, as shown in Fig. 6.
- LIME explanations demonstrated how the model evaluates the interplay between contradictory indicators, with specific instances showing features F13 and F0 strongly indicating benign behavior while F19, F18, and F4 exhibited characteristics typically associated with malicious traffic, as shown in Fig. 7.

These interpretability findings provide security practitioners with actionable insights for developing more targeted detection rules and understanding the key traffic characteristics that distinguish malicious from benign DoH communications.

V. Conclusion

This research presented a comprehensive machine learning framework for detecting malicious DNS over HTTPS traffic. Our approach addressed the challenges of class imbalance through SMOTE oversampling and feature redundancy through correlation analysis.

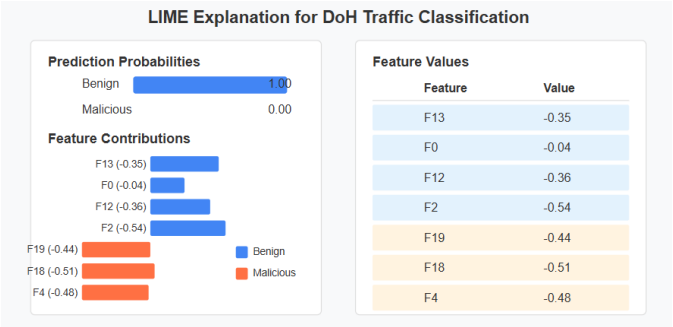


Fig. 7. LIME visualization for an individual prediction, showing feature contributions toward classification and their specific values.

The comparative evaluation of seven different machine learning algorithms revealed that ensemble-based approaches, particularly XGBoost, achieve exceptional performance in this domain, with XGBoost demonstrating perfect classification accuracy. The strong performance of tree-based models suggests that the features extracted from DoH traffic contain clear patterns that differentiate malicious from benign communications.

Feature importance analysis highlighted the significance of temporal characteristics and statistical flow properties in detecting malicious DoH traffic, providing valuable insights for security practitioners and researchers developing DoH-specific security solutions. The integration of explainability techniques (SHAP and LIME) enhanced the transparency of our models, offering instance-specific explanations that can aid in understanding and validating model predictions.

In conclusion, our research demonstrates the effectiveness of machine learning approaches for securing DNS over HTTPS communications while preserving the privacy benefits that motivated its development. The exceptional performance of ensemble models, particularly XGBoost, highlights their potential for integration into practical security solutions for protecting modern networks against the emerging threat of malicious DoH traffic.

The framework proposed herein can also serve as a blueprint for traffic classification in other encrypted protocols. Moreover, expanding the dataset diversity across multiple network environments could further improve generalizability. The findings also suggest that adaptive thresholding and continuous feedback mechanisms may enhance resilience against adversarial behavior. Overall, this work contributes a robust and interpretable approach to encrypted traffic analysis in a rapidly evolving cybersecurity landscape.

Additionally, leveraging federated learning could allow for decentralized threat intelligence sharing without compromising user privacy. Incorporating real-time monitoring capabilities may further empower security infrastructure to respond dynamically to novel DoH-based attack vectors. While ensemble models offer strong predictive power, their deployment should also consider inference time and computational overhead, especially in latency-sensitive or resource-constrained environments.

Future work should focus on creating unsupervised and semi-supervised methods to detect new DoH-based threats without needing large labeled datasets. Developing online learning techniques that adapt to changing attack patterns will help maintain strong detection over time. Additionally, designing lightweight models optimized for real-time use on devices with limited resources will improve practical deployment. Exploring federated learning could also enable collaborative threat detection across distributed networks while protecting user privacy.

REFERENCES

[1] T. Böttger, F. Cuadrado, G. Tyson, I. Castro, and S. Uhlig, "Open connect everywhere: A glimpse at the internet ecosystem through the lens of the netflix cdn," *ACM SIGCOMM Computer Communication Review*, 2019.

[2] Z. Lu, Y. Zhang, and C. Fang, "Doh insight: Detection of malicious doh traffic using fine-grained features," *IEEE Trans on Net and Serv Man*, 2021.

[3] R. Houser, Z. Zhou, S. Kumbhare, N. Feamster, A. Terzis, L. Iannone, and R. Zhu, "A comprehensive and operational risk assessment framework for dns over https abuse," in *42nd IEEE Symposium on Security and Privacy*, 2020.

[4] C. Deccio, S. Mitchell, and E. Osterweil, "Dns privacy considerations for operator networks," in *2020 IEEE Sym on Privacy-Aware Net*, 2020.

[5] S. Montazeri, A. Torabi, and A. H. Lashkari, "A deep dive into the cira-cic-dohbrw-2020 dataset: A benchmark for encrypted dns traffic analysis," *International Journal of Network Security*, 2020.

[6] R. Rathi and J. Kumar, "Explainable artificial intelligence in cybersecurity: A review," *ACM Computing Surveys*, 2021.

[7] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Leanpub, 2022.

[8] H. Jha, I. Patel, G. Li, A. K. Cherukuri, and I. S. Thaseen, "Detection of tunneling in dns over https," in *2021 7th Int Conf on Sig Pro and Com (ICSC)*, 2021.

[9] A. K. Bozkurt, B. S. Sarıkaya, H. E. Aköz, A. Taşpınar, and Bahtiyar, "A new method to detect malicious dns over https via feature reduction," in *2024 9th Int Conf on Com Sci and Engr (UBMK)*, 2024.

[10] P. Boonyopakorn and U. Changsan, "Malicious traffic detection in dns over https (doh): Edge prediction with graph convolutional network," in *2024 Int Tech Conf on Circuits/Systems, Computers, and Communications (ITC-CSCC)*, 2024.

[11] S. Mahdavifar, N. Maleki, A. H. Lashkari, M. Broda, and A. H. Razavi, "Classifying malicious domains using dns traffic analysis," in *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing (DASC/PiCom/CBDCom/CyberSciTech)*, 2021.

[12] Z. Deng, K. Jang, R. Huang, B. Wu, and S. Zhu, "Doh detection method based on self-attention bilstm," in *2024 17th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2024.

[13] A. M. A. Badri and S. Alouneh, "Detection of malicious requests to protect web applications and dns servers against sql injection using machine learning," in *2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)*, 2023.

[14] M. F. B. Hafiz, N. A. Khan, Z. Kamal, S. Hossain, and S. Barman, "A robust malware classification approach leveraging explainable ai," in *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, 2024.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 2002.

[16] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE Symposium on Security and Privacy (SP)*, 2016.

[17] C. Chen, W. Gan, J. Ma, J. Yang, and Y. Yang, "Featuregame: An efficient shap value estimation framework for tree ensembles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[18] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Know and Inf Systems*, 2014.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144.

[20] V. Arya, F. Nargesian, and A. Meliou, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," *IEEE Data Eng. Bull.*, 2019.