HERALD: Hybrid Ensemble Approach for Robust Anomaly Detection in Encrypted DNS Traffic

Umar Sa'ad[a], Demeke Shumeye Lakew[b], Nhu-Ngoc Dao[c,*], Sungrae Cho[a,*]

[a]School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea
[b]Department of Computer Science, Kombolcha Institute of Technology, Wollo University, Dessie 1145, Ethiopia
[c]Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea

Abstract

The proliferation of encrypted Domain Name System (DNS) traffic through protocols like DNS over Hypertext Transfer Protocol Secure presents significant privacy advantages but creates new challenges for anomaly detection. Traditional security mechanisms that rely on payload inspection become ineffective, necessitating advanced strategies capable of detecting threats in encrypted traffic. This study introduces the Hybrid Ensemble Approach for Robust Anomaly Detection (HERALD), a novel framework designed to detect anomalies in encrypted DNS traffic. HERALD combines unsupervised base detectors, including Isolation Forest (IF), One-Class Support Vector Machine (OCSVM), and Local Outlier Factor (LOF), with a supervised Random Forest meta-model, leveraging the strengths of both paradigms. Our comprehensive evaluation demonstrates HERALD's exceptional performance, achieving 99.99 percent accuracy, precision, recall, and F1-score on the CIRA-CIC-DoHBrw-2020 dataset, while maintaining competitive computational efficiency with 110s training time and 2.2ms inference time. HERALD also demonstrates superior generalization capabilities on cross-dataset evaluations, exhibiting minimal performance degradation of only 2-4 percent when tested on previously unseen attack patterns, outperforming purely supervised models, which showed 5-8 percent degradation. The interpretability analysis, incorporating feature importance, accumulated local effects, and local interpretable model-agnostic explanations, provides insights into the relative contributions of each base detector, with OCSVM emerging as the most influential component, followed by IF and LOF. This study advances the field of network security by offering a robust, interpretable, and adaptable solution for detecting anomalies in encrypted DNS traffic that balances a high detection rate with a low false-positive rate.

Keywords: Anomaly detection, DNS over HTTPS, encrypted DNS traffic, hybrid ensemble models.

1. Introduction

1.1. Background

Concerns regarding pervasive monitoring, and the increasing frequency and sophistication of cyberattacks, have accelerated efforts to improve the security and privacy properties of major internet protocols and services. According to a FortiGuard Labs report [1], approximately 85% of the current web traffic is encrypted, a 30% increase from 2017. These figures are expected to rise due to increased efforts to secure the last significant unencrypted traffic on the Internet: the Domain Name System (DNS).

The DNS is a critical component of the Internet ecosystem. It is a distributed, hierarchical database used to translate network queries from human-friendly domain names to machine-readable Internet Protocol (IP) addresses. This traffic is sent in plaintext by default and can be intercepted by rogue entities. This information may include the identity of the nodes querying the database and the specific data requested. Furthermore, multiple points in the DNS resolution path, including the stub resolver, its communication links, the recursive resolver, and the authoritative nameservers, are susceptible to information leakage [2].

Given the privacy implications, several industry and academic efforts have been launched to reduce information leakage in the DNS resolution pipeline. This study explores the efforts to encrypt DNS transactions, and the most popular technology developed in this regard include DNS over Transport Layer Security (TLS) called DoT [3] and DNS over Hypertext Transfer Protocol Secure (HTTPS) called DoH [4]. Moreover, DoT improves the privacy properties of the DNS resolution pipeline by ensuring communication is secured over a TLS connection on TCP port 853. In contrast, DoH sends and receives DNS queries over an HTTPS connection on TCP port 443. Security concerns abound because of the possibility of malicious actors abusing these mechanisms.

Furthermore, existing security tools rely on the ability to analyze unencrypted DNS traffic to detect and prevent its malicious use. Data exfiltration via DNS tunneling and concealed malware command-and-control communication are two stan-

*Corresponding authors
Email addresses: umar@uclab.re.kr (Umar Sa'ad), demeke@uclab.re.kr (Demeke Shumeye Lakew), nndao@sejong.ac.kr (Nhu-Ngoc Dao), srcho@cau.ac.kr (Sungrae Cho)

dard methods of DNS abuse. Thus, the addition of encryption to DNS exacerbates the problem by allowing rogue entities to conceal their malicious activities. Moreover, unlike DoT traffic, which is distinguishable because it uses its own standard TCP port number, distinguishing DoH traffic from encrypted web traffic requires considerably more effort. Consequently, DoH may be more vulnerable to abuse than DoT. Fig. 1 presents a schematic representation of malicious data exfiltration using DoH tunneling.

1.2. Motivation

In recent years, DoH has grown in popularity due to its ability to improve user privacy and security while browsing the Internet. Moreover, DoH support has been built into most web browsers and operating systems [5]. Furthermore, several extensive DNS services, such as Google, Cloudflare, and Quad9, have built DoH support into their public DNS resolvers [6]. A study by [7] suggests that many working DoH servers are not publicly known or published, implying that considerable number of DoH servers are used for new service offerings. Approximately 73% of these unpublished resolvers lack a reverse DNS host name, which could be considered suspicious and an indication of malicious or fraudulent activity.

Thus, a critical task is to develop detection mechanisms that can navigate the intricacies of encrypted DNS traffic without infringing on user privacy. Such mechanisms must be sophisticated enough to differentiate between benign and malicious use, ensuring network security without undermining the foundational principles of Internet privacy. This backdrop of heightened security risk amid increasing privacy measures motivates this research. It underscores the imperative for innovative approaches that are adept at identifying anomalies within encrypted DNS traffic and are transparent and adaptable to the ever-evolving tactics of cyber adversaries.

1.3. Contributions

This study introduces the hybrid ensemble approach for robust anomaly detection (HERALD), an innovative approach specifically tailored to address the complexities of encrypted DNS traffic. The HERALD framework combines the strengths of unsupervised anomaly detection methods with the precision of a supervised learning meta-model. The following contributions are central to this research:

- We propose a sophisticated ensemble model that strategically integrates multiple unsupervised algorithms with a random forest (RF) meta-model. This hybridization allows for a nuanced interpretation of the encrypted DNS traffic, capitalizing on the diverse perspectives offered by each detection method.
- Our approach is underscored by a specialized training regimen that optimizes the performance of both base detectors and meta-model. By partitioning the dataset and leveraging the distinctive characteristics of benign and malicious traffic, HERALD achieves a fine-tuned balance between sensitivity and specificity, maintaining remarkably low false positive rates across diverse test scenarios.
- We implement a carefully administered feature extraction process to derive anomaly scores from the base detectors into a format that enriches the meta-model learning phase. This extraction encapsulates the insights of the detectors, furnishing the meta-model with sophisticated descriptive and predictive features that enable effective discrimination between normal and anomalous encrypted traffic.
- Through cross-dataset evaluation with three diverse datasets, we demonstrate HERALD's superior generalization capabilities, exhibiting only 2-4 percent performance degradation on previously unseen attack patterns compared to 5-8 percent for purely supervised models. This highlights HERALD's ability to adapt to novel threats in encrypted DNS traffic, a critical advantage in evolving cybersecurity landscapes.
- Beyond its robust detection capabilities, HERALD is designed with an emphasis on interpretability. Our detailed interpretability analysis, including feature importance, accumulated local effects, and local interpretable model-agnostic explanation plots, reveals the relative contributions of each base detector, with OCSVM emerging as the most influential component (42 percent), followed by IF (38 percent) and LOF (20 percent). This transparency fosters trust and provides security administrators with actionable insights on model predictions.

The remainder of this paper is organized as follows. Section 2 explores the background literature concerning the privacy enhancements and security challenges introduced by DoH, examines machine learning approaches for encrypted DNS traffic analysis, investigates deep learning and image-based detection methods, and discusses the innovative adaptation of hybrid learning models for anomaly detection. Section 3 describes HERALD's design rationale, detailing the selection methodology for unsupervised base detectors and explaining the role of the supervised meta-model. Next, Section 4 covers the dataset description, exploratory feature analysis, and outlines the preprocessing steps including handling missing values, feature elimination, standardization, and dataset resampling. Then, Section 5 details the model training process and evaluation methodology, including cross-dataset validation, and presents a comparative analysis of HERALD against purely unsupervised, supervised, and deep learning models. Section 6 offers an in-depth interpretability analysis, incorporating feature importance to quantify detector contributions, accumulated local effects to visualize feature-response relationships, statistical analysis of detector influences, and LIME explanations for individual predictions. Finally, Section 7 concludes with a discussion on the potential and limitations of HERALD and suggests directions for future work to enhance and generalize this approach.

2. Background Literature

This section explores the multifaceted landscape of anomaly detection in DoH traffic, a critical research area in the network

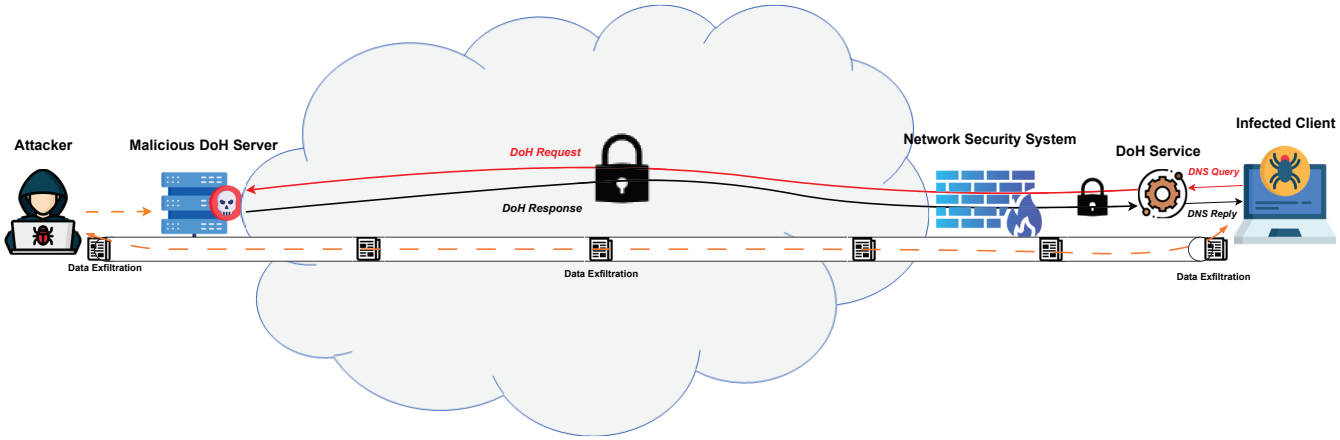


Figure 1: Schematic representation of malicious data exfiltration using DNS over HTTPS (DoH) tunneling attack.

security and privacy domain. The literature review is structured into four thematic areas: 1) an examination of the privacy enhancements and challenges introduced by adopting DoH; 2) an exploration of the role of machine learning in analyzing encrypted DNS traffic; 3) innovative approaches in deep learning and image-based detection for DoH abuse; and 4) insight into hybrid learning models that combine supervised and unsupervised techniques. Each theme collectively builds an understanding of the current state of anomaly detection within encrypted DNS traffic, providing a foundation on which the proposed HERALD framework is developed. This review highlights the complexities and advancements in this field and sets the stage for the ensuing discussion on the novel approach of HERALD in addressing these challenges.

2.1. Enhancing Privacy with DoH: Challenges and Opportunities

The evolution from conventional DNS to encrypted DNS represents a significant leap in securing user privacy. As outlined by researchers e.g., [8], DoH effectively addresses longstanding privacy concerns associated with traditional, unencrypted DNS communications. This advancement encrypts DNS queries, protecting them from external surveillance and interference. Furthermore, one study [9] made a significant contribution by implementing and deploying the Oblivious DoH, which maintains performance comparable to DoH and DoT while significantly enhancing client privacy. This innovation indicates the potential of the Oblivious DoH and similar protocols to serve as practical, privacy-enhancing alternatives to conventional DNS usage, representing a leap forward in the ongoing quest to secure user data.

Comprehensive surveys [10, 11] highlight the dual-use nature of DNS encryption technologies, including DoH and DoT. While these protocols enhance privacy, they simultaneously complicate threat detection efforts. These surveys outline common attack vectors such as data exfiltration and command-and-control (C2) communication, emphasizing the necessary shift from traditional payload-based detection methods to more sophisticated behavioral analysis techniques that examine query patterns, timing characteristics, and entropy measures. Another study [12] further contributes to this understanding by specifically focusing on DoH abuse, categorizing various malicious use cases and advocating for machine learning-driven detection approaches that leverage metadata such as packet sizes and session duration to compensate for encrypted payloads.

The encryption that DoH provides, although safeguarding privacy, introduces new challenges in network monitoring and security because encryption conceals DNS query content, making traditional monitoring and threat detection methodologies less effective. Researchers [5] and [13] have discussed the difficulties in recognizing and controlling malicious activities, such as DNS tunneling, in the face of encrypted DoH traffic. These studies underline the complexities involved in identifying and classifying encrypted traffic, necessitating innovative techniques capable of identifying and mitigating potential threats hidden within encrypted traffic. These techniques must strike a delicate balance, maintaining robust security measures without infringing on the privacy enhancements that DoH provides.

2.2. Machine Learning for DoH Traffic Abuse Detection

Recent advancements in machine learning approaches for detecting DoH traffic abuse have underscored the significant potential of these methods to balance privacy enhancement with effective network security measures. The field has witnessed a rise in innovative machine learning-based approaches aimed at accurately detecting and classifying malicious DoH traffic. Notably, [14] introduced a systematic two-layer approach employing six machine learning algorithms to differentiate between benign and malicious DoH traffic. This study highlights the increasing reliance on sophisticated, multifaceted machine learning strategies for effective DoH traffic analysis.

Building upon foundational network traffic analysis techniques, researchers have demonstrated the effectiveness of machine learning approaches in broader network security contexts. A study [15] developed a dual-grained classification system that employs supervised machine learning models to analyze network behavioral patterns of enterprise assets. Their approach utilizes transport- and network-layer behavioral analysis to clas-

sify assets into both fine-grained specific types and coarse-grained generic categories, achieving classification accuracy of nearly 99 percent. This dual-grained methodology demonstrates the potential for sophisticated behavioral analysis that could be adapted for encrypted traffic scenarios where payload inspection is not possible. Similarly, researchers [16] contributed to the understanding of DNS traffic behavioral analysis through comprehensive passive traffic analysis techniques. Their work on enterprise DNS asset mapping and cyber-health tracking provides valuable insights into DNS behavioral profiling that can inform DoH anomaly detection approaches.

Moreover, researchers [17, 18] have conducted systematic comparisons of various machine learning models for DoH detection, revealing that tree-based models such as XGBoost often outperform neural networks in scenarios with limited training data. Their research identifies critical features for effective detection, including temporal metrics like query frequency and domain name characteristics such as entropy and subdomain counts. In another study, researchers [19] prioritize real-time detection capabilities using traditional machine learning models including Random Forest and Support Vector Machines (SVM) with carefully engineered features such as query length and domain entropy. Their system emphasizes low latency, making it particularly suitable for deployment in enterprise network environments.

Furthermore, the work by [20] provides a comparative analysis of feature selection techniques in the realm of encrypted HTTPS traffic, offering valuable insight that extends to the detection of abusive DoH traffic. This comparative perspective is crucial in understanding and enhancing the performance of machine learning models tailored for DoH abuse detection. Additionally, researchers [21, 22] have contributed to the evolving landscape of DoH traffic analysis by exploring deep learning and other learning approaches for detecting DoH traffic tunnels and identifying malicious activities. Their research provides a glimpse into the potential of these advanced methodologies to offer deeper insight and more accurate classification, thereby enhancing privacy and security.

The authors in [23] used an immensely popular public DoH dataset to develop an accurate and explainable artificial intelligence (AI)-based intrusion detection system. This approach provides a practical solution for detecting and classifying DoH attacks and emphasizes the importance of transparency and understandability in AI applications. The use of explainable AI in this context ensures that the rationale behind decisions made by machine learning models is accessible and interpretable, which is crucial for trust and validation in security applications. Similarly, another study [24] addresses interpretability challenges with a deep learning framework that provides feature importance scores, aiding administrators in understanding detection decisions and improving the practical utility of these sophisticated models.

Collectively, the literature underscores a burgeoning interest in leveraging machine learning to combat DoH traffic abuse. The development of innovative models and techniques reflects the dynamic nature of the field, continually evolving to address the intricate challenges associated with DoH abuse.

2.3. Deep Learning and Image-Based Detection Approaches

As the complexity of DoH abuse techniques increases, researchers have explored novel deep learning and image-based approaches to enhance detection capabilities. Studies by [25] and [26] have applied sophisticated deep learning techniques, including Long Short-Term Memory networks (LSTMs) and Convolutional Neural Networks (CNNs), to classify DoH traffic by leveraging raw packet sequences for automated feature learning. While these models excel in accuracy, they face significant challenges in terms of explainability, a critical factor for practical implementation in security systems.

A particularly innovative approach has emerged in the form of image-based detection methods. Researchers [27] propose converting DNS traffic into grayscale images (DNS-images) and employing CNNs for classification. This method effectively captures spatial patterns in DNS queries, achieving high accuracy rates despite requiring substantial preprocessing. Building upon this concept, researchers [28] have developed FECC, a hybrid approach that combines CNNs for feature extraction with clustering techniques (unsupervised learning) to detect novel tunneling variants. This integration of supervised and unsupervised learning methodologies significantly improves the system's adaptability to evolving threats, a crucial advantage in the rapidly changing landscape of DoH abuse.

These image-based and deep learning approaches represent a significant departure from traditional detection methods, offering new perspectives on how to analyze and identify malicious patterns in encrypted DNS traffic. By transforming the detection problem into an image recognition challenge, researchers have opened new avenues for leveraging advancements in computer vision to enhance network security, particularly in the context of encrypted protocols like DoH.

2.4. Innovative Adaptation of Hybrid Learning for DoH Anomaly Detection

While hybrid learning models have demonstrated their efficacy in various domains [29, 30, 31, 32, 33], their application to DoH abuse detection is a pioneering venture. These models synergize supervised and unsupervised learning methods, an approach previously unexplored in the context of DoH anomaly detection. The promise of this methodology, as demonstrated in other areas, offers a compelling premise for its potential effectiveness in identifying and mitigating DoH abuse. This research is among the first to apply such an innovative approach in the DoH domain, setting a new benchmark in anomaly detection strategies.

Recent work by [28] demonstrates the growing interest in hybrid approaches specifically for DoH traffic analysis. By combining CNN-based feature extraction with clustering algorithms, their research shows how hybrid models can effectively detect novel variants of tunneling attacks in encrypted DNS traffic. This approach aligns with our research direction and confirms the validity of pursuing hybrid methodologies for enhanced detection capabilities.

The unique advantage of hybrid approaches is the ability to combine the predictive accuracy of supervised learning with the

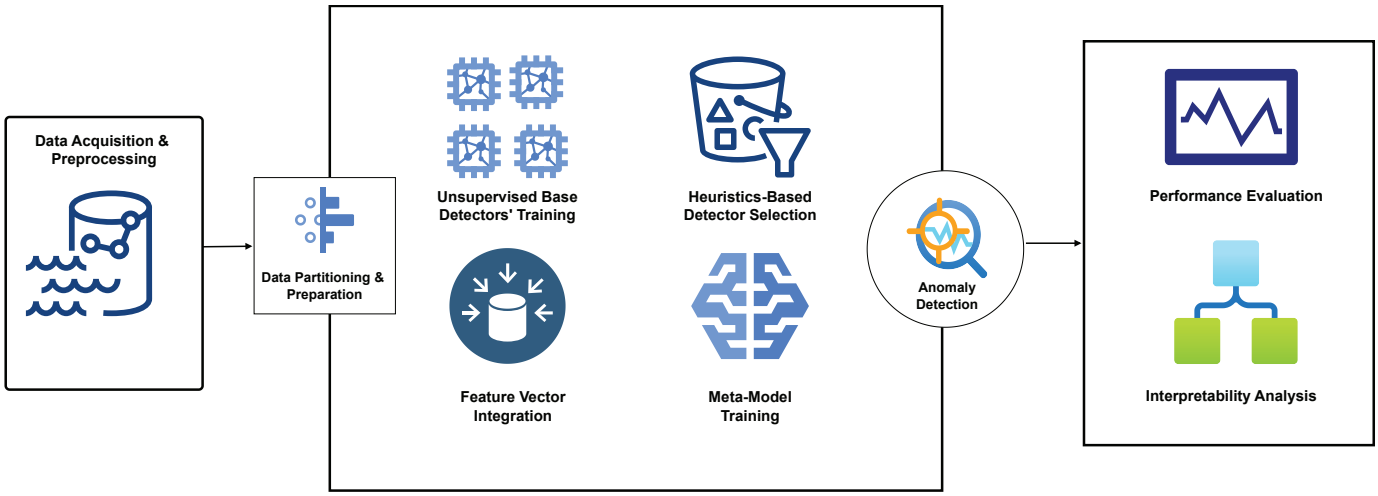


Figure 2: The HERALD Anomaly Detection Pipeline.

pattern recognition capabilities of unsupervised learning. This synergy enables a more nuanced understanding and detection of anomalies within encrypted DNS traffic. Integrating these methods permits the detection of complex patterns and subtle irregularities, which might be challenging to identify using traditional single-method approaches. This enhanced detection capability is crucial in the DoH context, where encryption adds a layer of complexity to traffic analysis.

Thus, applying hybrid learning models to DoH abuse detection represents a significant advancement in network security. By introducing this innovative approach, this research lays the foundation for more sophisticated and effective methods of detecting anomalies in encrypted DNS traffic. The promising results in this study validate the potential of hybrid models in the DoH context and encourage further exploration and development of these technologies which could transform security in the face of challenges of encrypted traffic.

3. HERALD: Hybrid Ensemble Approach

This section presents an in-depth description of HERALD, a novel approach for robust anomaly detection in encrypted DNS traffic. We begin by explaining the rationale and underlying principles that inform this approach, followed by describing the selected components, including the unsupervised base anomaly detectors and supervised meta-model.

3.1. Design Rationale

The hybrid ensemble approach, integrating unsupervised and supervised detection methodologies, is designed to address the intricate challenges of anomaly detection in encrypted DNS traffic. This approach aims to achieve a harmonious balance between a high detection rate and a low false-alarm rate, two critical metrics in network security. Fig. 2 presents a schematic overview of the HERALD framework.

3.1.1. High Detection Rate

The deployment of unsupervised detectors, which are adept at identifying diverse anomaly types, is central to the high detection rate. These detectors are particularly valuable owing to their ability to recognize aberrant patterns that may not be present in the training dataset. This capability is crucial due to the dynamic nature of network traffic, where new and unforeseen types of anomalies continually emerge.

3.1.2. Low False-Alarm Rate

Complementing the unsupervised detectors, the supervised meta-model of the ensemble refines the initial detection results. By scrutinizing the output from unsupervised models, the supervised detector plays a critical role in reducing false alarms. This aspect is essential in encrypted DNS traffic analysis, where the objective is to detect anomalies and minimize the misclassification of legitimate traffic.

3.1.3. Additional Advantages

Beyond its core capabilities, the hybrid ensemble approach offers several additional advantages:

- Adaptability to evolving traffic patterns: The dynamic nature of network traffic, notably encrypted DNS traffic, necessitates an adaptable detection system. The hybrid ensemble approach meets this need by allowing the integration of new unsupervised base detectors as emerging threats and traffic patterns evolve. Concurrently, the supervised meta-model is designed to assimilate and learn from the evolving data, ensuring the system remains relevant and practical.

- Potential for generalization: The success of this approach in encrypted DNS traffic opens avenues for its application to other types of encrypted traffic including encrypted web traffic, encrypted email communications, and potentially other domains where encryption is prevalent. The

ability to generalize this approach to various contexts significantly enhances its utility and applicability in the broader field of network security.

3.2. Unsupervised Base Anomaly Detectors

In this section, we outline the rationale and methodology for selecting and implementing unsupervised base detectors within the HERALD framework. Our approach begins with a systematic selection process based on heuristics that aims to optimize both individual performance and methodological diversity. Formally, let $D = \{d_1, d_2, \ldots, d_n\}$ denote the set of candidate detectors, where each detector d_i represents a distinct anomaly detection algorithm. We evaluated each candidate's performance using a score $P(d_i)$, calculated on validation data through the precision between actual labels y and predictions $\hat{y}_i$. A detector is included in our ensemble if its performance exceeds $\tau \times \mathcal{P}^*$, where $\mathcal{P}^*$ is the highest performance score observed among all candidates and τ, an empirically determined threshold, is set to 0.9. This methodology offers several advantages, including linear computational complexity ($O(n)$), transparent and reproducible selection criteria, and a reduced risk of overfitting.

Consequently, three detectors were selected as candidates, each embodying a unique detection paradigm. The first detector, Isolation Forest (IF) [34], leverages random feature subspace isolation to efficiently identify global anomalies. Its anomaly score is mathematically defined as:

$$score(x_i, n) = 2^{-E(h(x_i))/c(n)}$$

where $E(h(xi))$ denotes the average path length in the constructed isolation trees.

Second, the One-Class Support Vector Machine (OCSVM) [35], implements a boundary-based approach by optimizing a hyperplane in the feature space to segregate normal data from anomalies. Its formulation involves minimizing the objective function:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - b$$

subject to the constraints:

$$w^T\phi(x_i) \geq b - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, ..., n$$

where $\phi(\cdot)$ indicates a mapping function that maps the input data to a higher-dimensional feature space, ξ_i denotes a slack variable that allows for some misclassification of training examples, and ν represents a hyperparameter that controls the fraction of outliers in the data.

The third detector, the Local Outlier Factor (LOF) [36], identifies anomalies by comparing the local density of a data point to that of its neighbors. This method computes the LOF value for a point X_i by evaluating the ratio between the average local reachability density of its neighbors to its own, thereby offering context-aware anomaly identification. Mathematically, the LOF value is computed as:

$$LOF(X_i) = \frac{\sum_{X_j \in N_k(X_i)} lrd(X_j)}{\| N_K(X_i) \|} \times \frac{1}{lrd(X_i)}$$

where $lrd(X_i)$ is the local reachability distance of X_i and $N_K(X_i)$ the set of $k-$nearest neighbors of X_i.

The integration of these detectors is a strategic decision underscored by several key advantages:

- *Diversity of detection methods*: The enhancement of ensemble models in anomaly detection, particularly in encrypted DNS traffic, is significantly bolstered by the diversity of their detection methodologies, underscored by extensive research in ensemble learning [37, 38]. Renowned for its efficiency in large data sets, IF utilizes an isolation-based mechanism to identify anomalies, which is effective for detecting rare and atypical patterns prevalent in encrypted traffic. Employing a boundary-based approach, OCSVM excels in establishing a demarcation around normative data patterns, which is essential for identifying deviations within environments where anomalies lack clear definition. With a local density-based approach, LOF adds a granular layer of detection by identifying anomalies in the context of their local data neighborhood, making it particularly adept at uncovering subtle aberrations.

- *Complementarity and robustness*: The distinct methods that these algorithms employ complement each other in how they detect anomalies. The isolation approach of IF, the delineation of the normative data boundaries of OCSVM, and the assessment of local density variances of LOF integrate synergistically to foster a robust detection framework, capable of accurately identifying a broad spectrum of anomalies and enhancing overall reliability of the system.

However, we recognize that the reliance on individual performance metrics may overlook potential interactions between detectors, and the fixed-threshold methodology presents trade-offs between computational efficiency and the exhaustive exploration of detector combinations. Despite these challenges, the selected ensemble represents a balanced compromise between simplicity and domain-specific optimization, thus forming a robust foundation for subsequent anomaly detection within the HERALD framework.

3.3. Supervised Meta-Model: Random Forest (RF)

The Random Forest (RF) [39] meta-model within the HERALD framework plays a pivotal role in aggregating and interpreting the outputs from unsupervised base detectors. This supervised meta-model operates on feature vectors defined as $F_i = [f_1, f_2, f_3]$, where f_1 is the anomaly score of the IF, f_2 represents the decision function of the OCSVM, and f_3 is the outlier factor computed by the LOF. The ensemble is configured with 100 trees, an unrestricted maximum tree depth, a minimum of 2 samples required to split a node, and employs feature sampling based on the square root of the total number of features. The architecture of the RF metamodel is shown in Fig. 3

For a given input sample X the meta-model produces an output defined as:

$$RF(X) = \text{mode}\{t_1(X), t_2(X), ..., t_T(X)\}$$

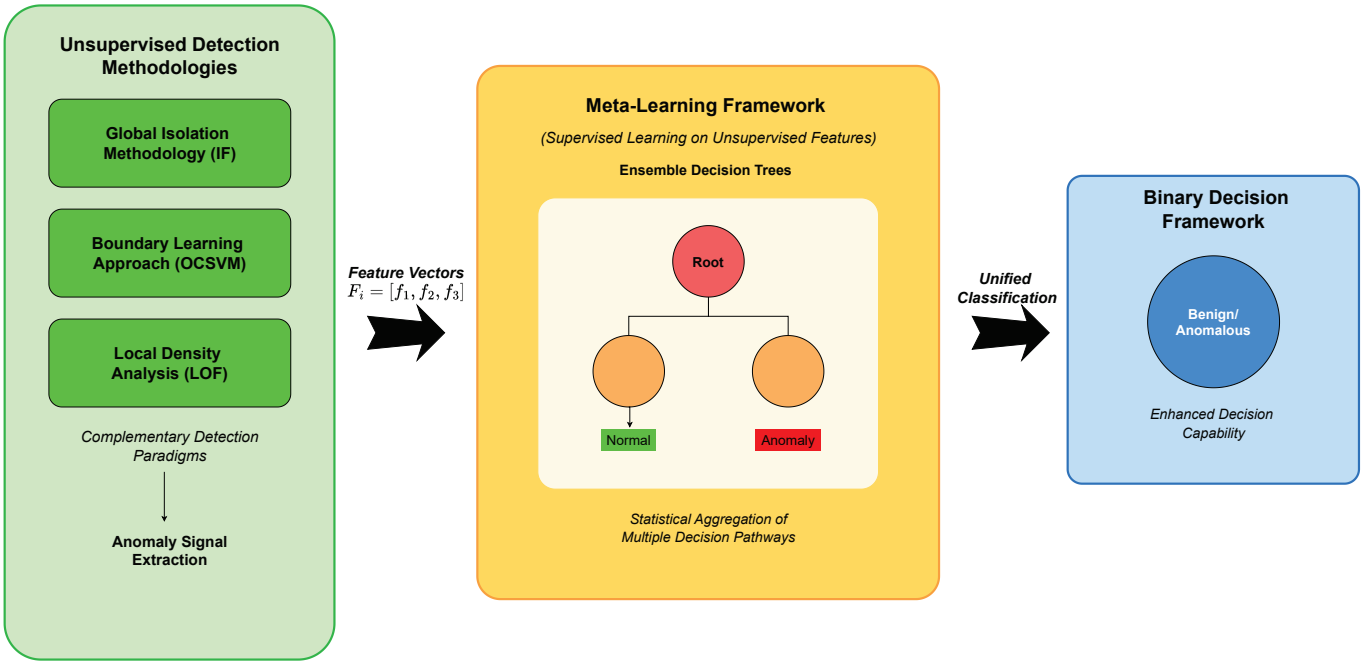


Figure 3: RF Meta-model Architecture

where each $t_i(X)$ corresponds to the prediction from an individual tree. The selection of the RF meta-model is driven by its architectural advantages, such as the ability to automatically detect interactions among features and robustly handle nonlinear relationships, while inherently leveraging ensemble characteristics. Additionally, the model offers significant implementation benefits. It requires minimal hyperparameter tuning, supports natively heterogeneous input, and provides built-in metrics to assess the importance of characteristics.

Performance-wise, the RF meta-model exhibits a training complexity of $O(T \times n \log n)$ and a prediction complexity of $O(T \times \log n)$, with a memory footprint of $O(T \times n)$, where T represents the number of trees and n denotes the sample size. This configuration ensures that the meta-model not only enhances detection accuracy through ensemble learning but also maintains computational efficiency, making it an effective component of the overall anomaly detection strategy in HERALD.

4. Dataset Description, Analysis, and Preprocessing

This section outlines the dataset employed to train and evaluate HERALD by detailing its composition and the preprocessing measures. We provide an analysis of the dataset features, highlighting the distinctions between benign and malicious traffic patterns. The preprocessing steps, which are critical to model accuracy including feature standardization, handling missing values, and class balance are concisely described, setting a solid foundation for the model's performance assessment.

4.1. Dataset Description

In this research, we used the widely acknowledged CIRA-CIC-DoHBrw-2020 dataset [40], provided by the Canadian Institute of Cybersecurity (CIC). This dataset encompasses non-DoH and DoH traffic. The former represents HTTPS traffic from visits to Alexa-ranked domains. In contrast, the latter is classified into benign traffic from browser-based DoH clients such as Mozilla Firefox and Google Chrome, and malicious traffic produced using DNS tunneling tools, including dns2tcp, DNSCat2, and Iodine [40]. This study focused exclusively on DoH traffic (benign and malicious) to build an ensemble model for detecting DoH abuse.

The dataset consists of 34 features organized into various categories such as network endpoint information, flow statistics, and packet statistics. The network endpoint information encompasses the IP addresses and port numbers of the source and destination nodes. Flow statistics are subdivided into two categories: flow timing and duration, containing attributes like *"TimeStamp"* and *"Duration"*, and flow volume and rate, including such attributes as *"FlowBytesSent"* and *"FlowSentRate"*. Packet statistics are further segmented into packet length, packet timing, and response time statistics. Table 1 provides a comprehensive breakdown of these categories and their corresponding features.

4.2. Exploratory Feature Analysis

The feature profiles were thoroughly examined to elucidate the unique attributes distinguishing benign and malicious DoH traffic. Consequently, kernel density estimation (KDE) plots were employed to visualize the classwise density distribution for *"Duration"*, *"FlowBytesSent"*, *"FlowBytesReceived"*, and *"PacketLengthMean"*. This is as shown in Fig. 4. The KDE plots provide a smooth, continuous visualization of the probability density functions of various features which is advantageous in discerning subtle differences in the distribution pat-

Table 1: Categorization of the Dataset's Features

Category	Features
Network endpoint information	SourceIP, DestinationIP, SourcePort, DestinationPort
Flow timing and duration	TimeStamp, Duration
Flow volume and rate	FlowBytesSent, FlowSentRate, FlowBytesReceived, FlowReceivedRate
Packet length statistics	PacketLengthVariance, PacketLengthStandardDeviation, PacketLengthMean, PacketLengthMedian, PacketLengthMode, PacketLengthSkewFromMedian, PacketLengthSkewFromMode, PacketLengthCoefficientofVariation
Packet timing statistics	PacketTimeVariance, PacketTimeStandardDeviation, PacketTimeMean, PacketTimeMedian, PacketTimeMode, PacketTimeSkewFromMedian, PacketTimeSkewFromMode, PacketTimeCoefficientofVariation
Response time statistics	ResponseTimeTimeVariance, ResponseTimeTimeStandardDeviation, ResponseTimeTimeMean, ResponseTimeTimeMedian, ResponseTimeTimeMode, ResponseTimeTimeSkewFromMedian, ResponseTimeTimeSkewFromMode, ResponseTimeTimeCoefficientofVariation

terns between benign and malicious traffic. Unlike histograms or bar charts, KDE plots offer a nuanced and detailed representation, allowing for a clear comparison of overlapping densities and identification of distinct distributional characteristics.

The KDE plots in Fig. 4 reveal distinctive differences between benign and malicious network activities. Malicious activities are characterized by concentrated spikes in short durations and specific average packet lengths, suggesting a pattern of behavior. In contrast, benign activities display a wider range of durations and a broader, more varied distribution in packet lengths, including occasional high-volume data transmissions that are atypical of malicious traffic. While malicious activities might follow a more predictable pattern, benign activities encompass a more diverse spectrum of network behavior. These findings can be instrumental in developing cybersecurity measures, as the defined characteristics of malicious traffic could be employed to enhance anomaly detection algorithms and improve network security protocols.

4.3. Data Preprocessing

A systematic approach was employed to prepare the dataset for analysis and modeling, addressing common data problems such as missing values, discrepancies in feature scale, and class imbalance.

4.3.1. Handling Missing Values

The dataset contains 688 missing data points across two features, which was addressed by imputing missing data points with the median of the relevant feature distributions. This method was selected owing to its superior statistical properties, notably its robustness against the distortion effects of outliers and skewed data. This approach ensures that the central tendency of the dataset remains intact, conserving the intrinsic distributional properties of the data and maintaining the structural integrity of the dataset. This method facilitates an unbiased analytical foundation for modeling exercises.

4.3.2. Feature Elimination

Feature elimination was performed to enhance the generalization of the model by eliminating attributes that could introduce bias or redundancy. Specifically, categorical network endpoint information such as "*SourceIP*", "*DestinationIP*", "*SourcePort*", and "*DestinationPort*" was excluded from the data set. These features inherently encode node-specific attributes that may inadvertently lead to model overfitting by associating specific endpoints with malicious or benign activity rather than learning intrinsic traffic patterns. Additionally, the "*TimeStamp*" feature was removed to prevent potential temporal biases that could skew the anomaly detection process.

4.3.3. Feature Standardization

Feature standardization was implemented to ensure uniformity in feature scales using the Robust Scaler, which scales the features based on the median and interquartile range (IQR). The transformation is expressed as follows.

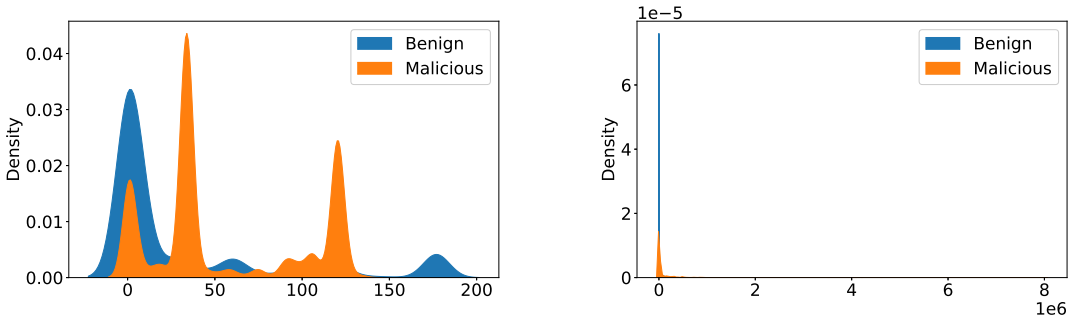
$$X_{scaled} = \frac{X - \text{median}(X)}{IQR(X)}$$

Unlike standardization techniques that assume a Gaussian distribution, the Robust Scaler was specifically chosen because the KDE plots in Fig. 4 revealed that our dataset exhibits non-Gaussian characteristics with significant skewness and outliers. By centering the data around the median and scaling it using the interquartile range, RobustScaler effectively mitigates the influence of extreme values, preserving the intrinsic structure of the dataset while ensuring that feature magnitudes remain comparable. This choice is particularly advantageous for anomaly detection tasks, where preserving the relative spacing between data points is crucial to identifying deviations from normal patterns.

4.3.4. Dataset Resampling

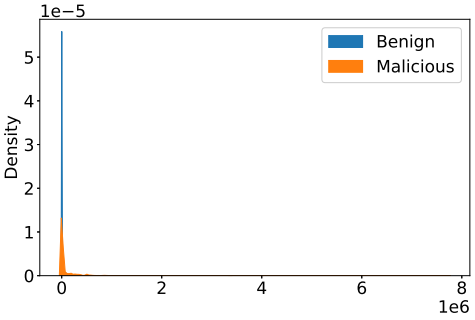
Initially, the dataset was characterized by a significant class imbalance, with an overrepresentation of malicious instances compared to benign ones. Several resampling techniques were considered, including synthetic minority over-sampling technique (SMOTE) [41], adaptive synthetic sampling (ADASYN) [42], and packet conditional generative adversarial networks (PacketCGAN) [43]. However, SMOTE was ultimately selected due to its computational efficiency, stability, and demonstrated effectiveness in preserving decision boundaries in high-dimensional feature spaces.

SMOTE was applied exclusively to the training set to prevent data leakage and ensure that the validation and test sets

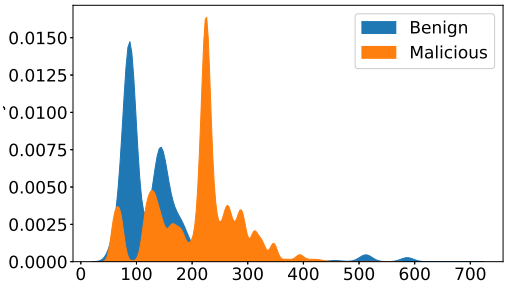


(a) Duration

(b) FlowBytesSent



(c) FlowBytesReceived



(d) PacketLengthMean

Figure 4: Comparative kernel density estimation (KDE) plots of network traffic features.

.

remained representative of real-world distributions. This technique synthesizes new instances in the minority class by interpolating between existing, similar instances. It selects a minority class instance a and determines its k-nearest minority class neighbors. A synthetic instance is created by choosing one of these k neighbors, b, and generating a new instance at a random point along the line segment joining a and b. Formally, the synthetic instance s is given by:

$$s = a + (b - a) \cdot \lambda$$

where λ is a random number between 0 and 1.

The preference for SMOTE was due to three key contextual factors. First, its deterministic interpolation mechanism facilitated reproducible and computationally efficient sample generation, aligning with the need for resource-aware training cycles. Second, SMOTE's localized sampling strategy mitigated the risk of over-adaptation, a challenge associated with ADASYN's boundary-focused approach, which could reduce generalizability in imbalanced threat detection scenarios [44]. Lastly, by restricting SMOTE's application to the training set, the validation and test sets preserved their original real-world distributions, ensuring an unbiased evaluation of model performance. Consequently, SMOTE provided an optimal balance between synthetic diversity, computational efficiency, and feature space integrity, making it well-suited for the HERALD framework's operational requirements.

5. Model Training and Evaluation

This section delineates the structured training process of the individual base detectors and the RF meta-model constituting the HERALD approach. This section details the dataset partitioning, specialized training, feature extraction, and their integration within the ensemble. Evaluation metrics, performance comparisons, and a comprehensive assessment of the predictive capabilities of the ensemble are also presented.

5.1. Model Training

The HERALD framework implements a sophisticated training methodology that systematically integrates unsupervised base detectors with a supervised meta-learning approach to achieve optimal anomaly detection performance. This section details the structured training protocol across the pipeline's components.

5.1.1. Dataset Partitioning and Preparation

Following the data processing phase, the dataset underwent a three-way split: 70 percent for training, 15 percent for validation, and 15 percent for testing, as illustrated in Fig. 2. To address class imbalance challenges inherent in cybersecurity datasets, SMOTE balancing was exclusively applied to the training partition while validation and test sets maintained their original distributions to ensure realistic performance evaluation. The

training set was further segregated based on traffic classification (benign/malicious) to accommodate the specific learning requirements of each base detector.

5.1.2. Unsupervised Base Detector Training

The HERALD pipeline incorporates three complementary unsupervised detection algorithms, each trained according to its optimal learning paradigm:

- IF was trained on the complete spectrum of training data (benign and malicious) to effectively model outliers within the feature space.
- OCSVM was trained exclusively on benign traffic samples to establish a comprehensive boundary representing normal network behavior.
- LOF similarly focused on benign traffic patterns to develop localized density estimations for anomaly identification.

Each detector generates a corresponding anomaly score (IF Score, OCSVM Score, and LOF Score) that quantifies deviation from expected behavior patterns.

5.1.3. Heuristic-based Detector Selection

Prior to meta-learning, HERALD employs a critical detector selection phase as detailed in Section III. This selection mechanism evaluates the performance of individual base detectors using a comparative threshold approach. Specifically, a detector is included in the ensemble if and only if its performance exceeds $\tau \times \mathcal{P}^*$, where $\mathcal{P}^*$ represents the highest performance score observed among all candidate detectors, and τ is an empirically determined threshold set to 0.9. This selection criterion ensures that only detectors performing within 10 percent of the best-performing detector are propagated to the meta-learning phase. The heuristic-based detector selection mechanism illustrated in Fig. 2 implements this threshold-based filtering, which optimizes the ensemble's discrimination capabilities by eliminating underperforming detectors while maintaining sufficient diversity in detection approaches. This intermediate selection step enhances both the computational efficiency and overall effectiveness of the final ensemble model.

5.1.4. Feature Vector Integration

Following detector selection, the anomaly scores produced by the selected base detectors were extracted and consolidated into an integrated feature vector. This transformation process converts the raw outputs from the unsupervised models into structured numerical representations that encapsulate diverse perspectives on potential anomalies. These feature vectors serve as the foundation for the subsequent meta-learning phase, bridging the unsupervised and supervised components of the HERALD pipeline.

5.1.5. Meta-Model Training

The RF meta-model represents the supervised learning component of HERALD, utilizing the integrated feature vectors from the selected base detectors. This meta-model was trained to synthesize the insights from individual detectors into a cohesive classification framework. During training, the meta-model learns optimal detector weighting patterns based on their demonstrated effectiveness during the selection phase. This approach enables HERALD to leverage complementary strengths across detection methodologies while minimizing their individual limitations. The final ensemble model resulting from this training process integrates both unsupervised anomaly detection capabilities and supervised classification precision, culminating in a robust binary classification system for DoH traffic.

5.2. Model Evaluation

Following the systematic training of HERALD including the unsupervised base models and supervised meta-model, we evaluated how the model performs on the testing dataset. The testing dataset comprises 49968 malicious samples and 3961 benign samples. This section describes the structured assessment methodology, including the model evaluation metrics, confusion matrix, and performance comparison of the base models and hybrid ensemble model.

5.2.1. Model Evaluation Metrics

The hybrid ensemble model was rigorously evaluated using a range of metrics crucial to understanding its performance in categorizing encrypted DNS traffic. The critical metrics used for this evaluation include the following:

- Accuracy: This metric measures the proportion of true results (both true positives and true negatives) among the total number of examined cases defined as follows:
$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{1}$$
where T_P denotes true positives, T_N represents true negatives, F_P indicates false positives, and F_N denotes false negatives.
- Precision: Precision measures the accuracy of the positive predictions defined as follows:
$$Precision = \frac{T_P}{T_P + F_P} \tag{2}$$
indicating the proportion of positive detections that are correct.
- Recall (sensitivity): Recall measures the ability of the model to identify all relevant instances, representing the proportion of correctly identified positives, defined as follows:
$$Recall = \frac{T_P}{T_P + F_N} \tag{3}$$
- F1-Score: The F1-score is the harmonic mean of the precision and recall, offering a single metric that balances them, calculated as follows:
$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

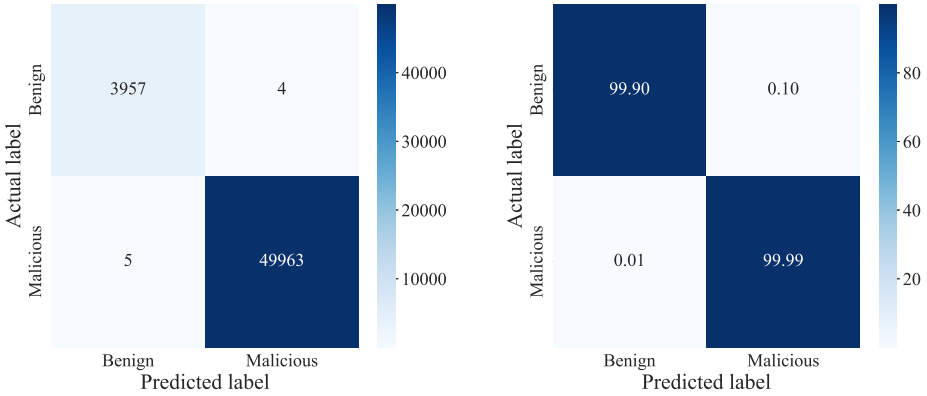


Figure 5: Confusion matrices for HERALD model performance evaluation showing both absolute counts (left) and percentages (right).

Table 2: Performance Comparison of All Models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Training Time (s)	Inference Time (ms)
IF	0.58	0.57	0.58	0.57	0.58	40	0.5
OCSVM	0.56	0.55	0.56	0.56	0.55	45	13.0
LOF	0.60	0.59	0.60	0.59	0.58	50	2.5
RF	0.9998	0.9999	0.9998	0.9999	1.0000	100	1.2
GB	0.9999	0.9999	0.9999	0.9999	1.0000	120	0.2
SVM	0.93	0.89	1.00	0.96	0.96	290	32.0
CNN	0.98	0.97	0.98	0.98	0.99	140	2.3
RNN	0.98	0.98	0.98	0.98	0.99	160	2.9
HERALD	**0.9999**	**0.9999**	**0.9999**	**0.9999**	**1.0000**	**110**	**2.2**

5.2.2. Confusion Matrices

The confusion matrices for the hybrid ensemble model, as depicted in Fig. 5, provide a nuanced assessment of its detection capabilities. The model demonstrated exceptional performance in identifying true negatives, correctly classifying 49,963 malicious samples out of 49,968, with a precision rate of 99.99 percent. Similarly, the model exhibited strong accuracy in recognizing benign traffic, correctly identifying 3,957 out of 3,961 benign samples, achieving a 99.90 percent true positive rate. The model demonstrated minimal error in both directions, with only 4 benign samples (0.10 percent) misclassified as malicious and merely 5 malicious samples (0.01 percent) incorrectly labeled as benign. This balanced performance is particularly noteworthy given the significant class imbalance in the dataset, where malicious samples constitute approximately 92.7 percent of the total. The model's ability to maintain high detection rates while minimizing both false positives and false negatives underscores its effectiveness in security applications where both accurate threat detection and reduction of false alarms are critical requirements.

5.3. Performance Comparison

The proposed approach was compared with alternative modeling approaches to contextualize the performance of the hybrid ensemble model: purely unsupervised models, purely supervised models, and deep learning models.

5.3.1. Comparison with Purely Unsupervised Models

For purely unsupervised models, we compared HERALD against the individual unsupervised base detectors, revealing a clear advantage of the hybrid ensemble approach. As shown in Fig. 6 and Table 2, the base unsupervised models demonstrated limited detection capabilities. The IF model achieved accuracy, precision, and recall values of approximately 0.58, indicating a balanced but significantly constrained capability in identifying malicious activities. Similarly, OCSVM displayed comparable limitations with performance metrics around 0.56, suggesting difficulties in distinguishing between classes. The LOF model performed marginally better with metrics near 0.60, but still fell substantially short of acceptable standards for operational deployment. In contrast, HERALD significantly outperformed each of these base detectors, achieving near-perfect scores across all metrics, demonstrating the efficacy of integrating these models within a hybrid ensemble framework.

5.3.2. Comparison with Purely Supervised Models

In the evaluation against leading supervised models including RF, Gradient Boosting (GB), and SVM, HERALD exhibited exceptional performance, establishing new benchmarks in encrypted DNS traffic anomaly detection. As illustrated in Figure 6 and detailed in Table 2, traditional supervised models like RF and GB demonstrated strong performance with metrics con-

sistently above 0.99. The SVM model showed slightly lower performance with an accuracy of 0.93 and recall of 1.0, but with a reduced F1-score of 0.96 due to precision limitations. Furthermore, SVM demonstrated prohibitively high computational requirements with a training time of 290 seconds—nearly three times higher than HERALD. HERALD achieved comparable or marginally superior performance to the best supervised models with an accuracy of 0.9999, F1-score of 0.9999, precision of 0.9999, and recall of 0.9999. Importantly, as shown in Figure 7, HERALD achieved this performance with moderate computational requirements, representing an optimal balance between detection capability and efficiency compared to purely supervised approaches.

5.3.3. Comparison with Deep Learning Models

To provide a comprehensive evaluation, we compared HERALD against state-of-the-art deep learning approaches, specifically Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), which have gained prominence in network traffic analysis. As shown in Table 2, both CNN and RNN models demonstrated strong performance with consistent accuracy, precision, recall, and F1-scores of 0.98, and AUC-ROC values of 0.99. While these results highlight the capability of deep learning methods to effectively model complex patterns in encrypted DNS traffic, they still fall short of the near-perfect performance achieved by HERALD (0.9999 across all metrics with an AUC-ROC of 1.0000). These findings suggest that while deep learning models offer sophisticated pattern recognition capabilities, the structured integration of multiple complementary models in HERALD's hybrid framework provides a more effective and balanced solution for DoH traffic anomaly detection.

5.3.4. Computational Efficiency Analysis

Beyond training efficiency, operational deployment considerations necessitate evaluating model inference time, i.e., the computational cost of generating predictions in real-world settings. As shown in Table 2 and Figure 7, inference time varied significantly across models. The GB model demonstrated exceptional inference speed (0.2 ms), followed by IF (0.5 ms) and RF (1.2 ms). However, this computational efficiency must be contextualized alongside detection performance; while GB and IF offer rapid inference, only GB provides comparable detection capabilities to HERALD. HERALD achieved competitive inference performance (2.2 ms), comparable to CNN (2.3 ms) and marginally faster than RNN (2.9 ms). The most computationally intensive models were SVM (32 ms inference, 290 seconds training) and OCSVM (13 ms inference), exhibiting latencies significantly higher than other approaches. SVM, in particular, represents the most computationally demanding model in both training and inference phases. These findings reveal that training time and inference time are not necessarily correlated, with some models demonstrating inefficient performance in both dimensions (SVM), while others exhibited efficient inference despite varying training requirements (GB, IF).

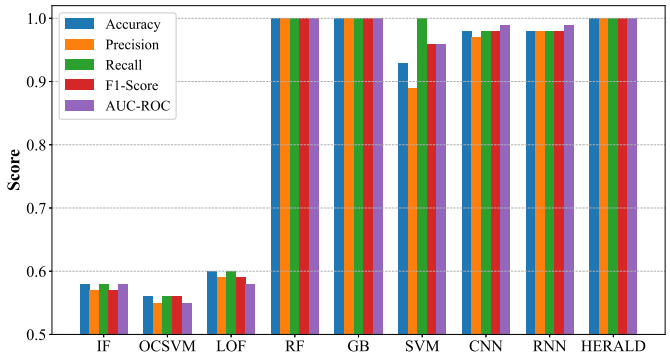


Figure 6: Performance metrics comparison across different models. The chart compares accuracy, precision, recall, F1-score, and AUC-ROC values for unsupervised learning methods (IF, OCSVM, LOF), supervised learning methods (RF, GB, SVM), and deep learning approaches (CNN, RNN) alongside our proposed HERALD approach.

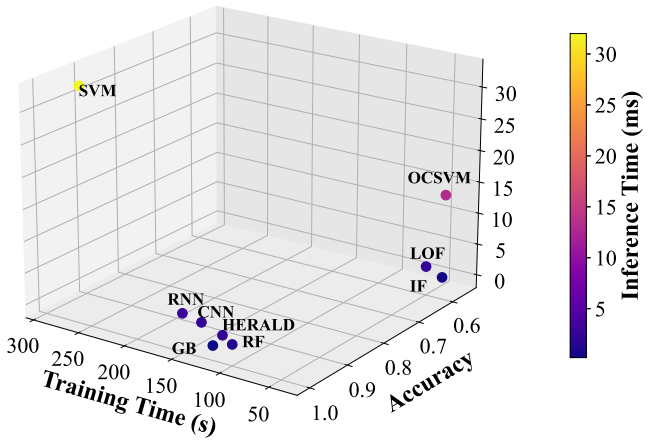


Figure 7: Performance-efficiency tradeoff visualization showing the relationship between accuracy, training time, and inference time across all evaluated models.

5.4. Cross-Dataset Performance Analysis

To comprehensively evaluate HERALD's generalizability and robustness in detecting DoH anomalies across different traffic contexts, we extended our evaluation beyond the CIRA-CIC-DoHBrw-2020 dataset to include two specialized datasets: DoH-DGA-Malware-Traffic-HKD [45] and DoH-Tunnel-Traffic-HKD [46]. The datasets are described in Table 3.

The DoH-DGA-Malware-Traffic-HKD dataset contains 4,212 flows representing four malware families: Tinba (42.9%), Padcrypt (19.9%), Zloader (19.5%), and Sisron (17.7%), with flow durations ranging from 0 to 170.7 seconds and an average of 72.4 seconds. The significantly larger DoH-Tunnel-Traffic-HKD dataset comprises 98,080 flows from three DNS tunneling tools: dns2tcp (47.0%), dnstt (30.6%), and iodine (29.6%), with durations spanning 0 to 135 seconds and averaging 68.2 seconds. Both datasets utilize the same 34 statistical features as the CIRA-CIC-DoHBrw-2020 dataset, ensuring feature compatibility for cross-dataset evaluation. While the malware dataset presents

challenges through high behavioral diversity across families, the tunneling dataset introduces complexity via obfuscation techniques and legitimate tool mimicry. These datasets enable comprehensive evaluation of HERALD's generalization capabilities across malware family identification and covert tunneling detection scenarios.

Furthermore, unlike the CIRA-CIC-DoHBrw-2020 dataset, which contains both benign and malicious traffic samples, the DoH-DGA-Malware-Traffic-HKD and DoH-Tunnel-Traffic-HKD datasets predominantly contain malicious traffic. This presents a methodological challenge for binary classification models that require both positive and negative samples for training and evaluation. To address this challenge, we employed a specialized evaluation framework as detailed below.

5.4.1. Methodology and Performance Metrics

The methodology for cross-dataset evaluation was designed to rigorously assess HERALD's detection capabilities across diverse datasets and attack scenarios. The process began with the construction of hybrid evaluation datasets, which were created by combining malicious traffic from the DoH-DGA-Malware-Traffic-HKD and DoH-Tunnel-Traffic-HKD datasets with benign traffic samples from the CIRA-CIC-DoHBrw-2020 dataset. This approach preserved the binary classification paradigm while introducing novel malicious patterns for evaluation.

To evaluate HERALD's ability to generalize to previously unseen attack vectors, a transfer learning paradigm was employed. Models were trained on the CIRA-CIC-DoHBrw-2020 dataset and tested on the constructed hybrid datasets. This zero-shot detection approach assessed the model's capability to identify malicious patterns without prior exposure to specific attack types, highlighting its adaptability to new threats.

Further granularity was achieved through attack-type-specific analysis, where malicious traffic was categorized to evaluate HERALD's performance across different threat vectors. This included assessing detection capabilities for Domain Generation Algorithm (DGA)-based malware traffic, analyzing performance on various DNS tunneling techniques, and identifying novel attack patterns that posed detection challenges.

To ensure a comprehensive and rigorous evaluation, a range of performance metrics was employed. Standard classification metrics such as accuracy, precision, recall, F1-score, and AUC-ROC were used alongside security-focused metrics, including detection rate (DR), false positive rate (FPR), and false negative rate (FNR). Additionally, stability metrics such as performance degradation rate (PDR) and Matthew's Correlation Coefficient (MCC) were utilized to evaluate performance consistency and robustness, particularly in imbalanced dataset scenarios. This multi-faceted approach provided a thorough assessment of HERALD's effectiveness and generalizability in detecting diverse and evolving threats.

5.4.2. Performance Comparison

Table 4 presents a comprehensive performance comparison, revealing key insights into model behavior. All models exhibited some degree of performance degradation when evaluated on additional datasets compared to the original CIRA-CIC-DoHBrw-2020 dataset, indicating that the new datasets introduced novel attack patterns with distinctive characteristics. HERALD consistently outperformed all baseline models, demonstrating significant advantages on the additional datasets. While purely supervised models like RF and GB showed considerable performance drops when faced with new attack patterns, HERALD maintained exceptional detection capabilities with minimal degradation, underscoring the effectiveness of its hybrid architecture. The DoH-Tunnel-Traffic-HKD dataset proved more challenging than the DoH-DGA-Malware-HKD dataset across all models, suggesting that tunneling techniques employ more sophisticated obfuscation methods that closely mimic legitimate traffic patterns. HERALD exhibited only 2-4 percent performance degradation on new datasets, compared to 5-8 percent for purely supervised models, highlighting its superior generalization ability by leveraging the complementary strengths of supervised and unsupervised approaches. Deep learning models like CNN and RNN showed significant performance degradation, indicating potential overfitting to specific patterns in the training data, a limitation effectively mitigated by HERALD's hybrid ensemble approach.

Further analysis by attack type provided targeted insights for security improvements. HERALD achieved a 97.32 percent detection rate for DGA-based malicious traffic and 95.73 percent for DoH tunneling traffic, outperforming baseline models such as GB (95.83 and 93.11 percent respectively) and RF (93.56 and 91.02 percent respectively). Additionally, HERALD maintained remarkably low false positive rates across datasets (2.12 percent for DGA-Malware and 3.67 percent for Tunneling), significantly outperforming baseline models like GB (3.12 percent and 5.42 percent) and RF (4.78 percent and 7.69 percent). This balance between high detection capability and minimal false alarms makes HERALD particularly suitable for operational deployment in production environments, where false positives can significantly impact workflow efficiency.

6. Interpretability Analysis

This section presents a comprehensive interpretability analysis of HERALD. We employed feature importance plots to extract the relative influence of the integrated detection methodologies, ALE plots to elucidate the relationships between critical features and the predictive outcomes, and LIME plots to provide instance-based insight into the decision-making process for HERALD. This analysis aims to clarify the operational mechanics of the ensemble, explaining the contributions of individual algorithms and the influence of specific feature variations on the model predictions.

6.1. Feature Importance Analysis

Understanding the relative contributions of individual anomaly detection components within HERALD is essential for optimizing its predictive performance in encrypted DNS traffic. To this end, we conducted a built-in feature importance analysis, quantifying the contributions of IF, OCSVM, and LOF to the ensemble model.

Table 3: Comprehensive Characteristics of Supplementary Evaluation Datasets

Aspect	DoH-DGA-Malware-Traffic-HKD	DoH-Tunnel-Traffic-HKD
Scale	4,212 flows, 4 malware families	98,080 flows, 3 tunneling tools
Families/Tools	Tinba (42.9%), Padcrypt (19.9%), Zloader (19.5%), Sisron (17.7%)	dns2tcp (47.0%), dnstt (30.6%), iodine (29.6%)
Features	34 features (same as CIRA- CIC-DoHBrw-2020)	34 features (same as CIRA- CIC-DoHBrw-2020)
Duration Range	0–170.7 seconds (avg: 72.4s)	0–135 seconds (avg: 68.28s)
Key Challenge	High behavioral diversity across families	Obfuscation and legitimate tool mimicry
Primary Application	Malware family identification	Covert tunneling detection

Table 4: Cross-Dataset Performance Comparison of All Models

Model	Dataset	Acc.	Prec.	Rec.	F1	AUC	DR	FPR	PDR	MCC
IF	CIRA-CIC-DoHBrw-2020	0.58	0.57	0.58	0.57	0.58	0.58	0.43	-	0.15
	DoH-DGA-Malware-HKD	0.55	0.54	0.56	0.55	0.56	0.56	0.46	0.035	0.10
	DoH-Tunnel-Traffic-HKD	0.53	0.52	0.54	0.53	0.54	0.54	0.48	0.070	0.06
OCSVM	CIRA-CIC-DoHBrw-2020	0.56	0.55	0.56	0.56	0.55	0.56	0.45	-	0.11
	DoH-DGA-Malware-HKD	0.54	0.53	0.55	0.54	0.54	0.55	0.47	0.036	0.08
	DoH-Tunnel-Traffic-HKD	0.52	0.51	0.53	0.52	0.53	0.53	0.49	0.071	0.04
LOF	CIRA-CIC-DoHBrw-2020	0.60	0.59	0.60	0.59	0.58	0.60	0.41	-	0.19
	DoH-DGA-Malware-HKD	0.58	0.57	0.59	0.58	0.57	0.59	0.43	0.017	0.16
	DoH-Tunnel-Traffic-HKD	0.56	0.55	0.57	0.56	0.56	0.57	0.44	0.051	0.13
RF	CIRA-CIC-DoHBrw-2020	0.9998	0.9999	0.9998	0.9999	1.0000	0.9998	0.0001	-	0.9997
	DoH-DGA-Malware-HKD	0.9423	0.9534	0.9356	0.9444	0.9612	0.9356	0.0478	0.0555	0.8889
	DoH-Tunnel-Traffic-HKD	0.9156	0.9245	0.9102	0.9173	0.9394	0.9102	0.0769	0.0826	0.8333
GB	CIRA-CIC-DoHBrw-2020	0.9999	0.9999	0.9999	0.9999	1.0000	0.9999	0.0001	-	0.9998
	DoH-DGA-Malware-HKD	0.9612	0.9699	0.9583	0.9640	0.9785	0.9583	0.0312	0.0359	0.9280
	DoH-Tunnel-Traffic-HKD	0.9347	0.9462	0.9311	0.9386	0.9568	0.9311	0.0542	0.0613	0.8774
SVM	CIRA-CIC-DoHBrw-2020	0.93	0.89	1.00	0.96	0.96	1.00	0.12	-	0.88
	DoH-DGA-Malware-HKD	0.88	0.83	0.96	0.89	0.91	0.96	0.20	0.0729	0.76
	DoH-Tunnel-Traffic-HKD	0.85	0.79	0.93	0.85	0.87	0.93	0.25	0.1146	0.69
CNN	CIRA-CIC-DoHBrw-2020	0.98	0.97	0.98	0.98	0.99	0.98	0.03	-	0.95
	DoH-DGA-Malware-HKD	0.93	0.92	0.94	0.93	0.94	0.94	0.08	0.0510	0.86
	DoH-Tunnel-Traffic-HKD	0.90	0.89	0.91	0.90	0.92	0.91	0.12	0.0816	0.79
RNN	CIRA-CIC-DoHBrw-2020	0.98	0.98	0.98	0.98	0.99	0.98	0.02	-	0.96
	DoH-DGA-Malware-HKD	0.94	0.93	0.94	0.93	0.95	0.94	0.07	0.0510	0.87
	DoH-Tunnel-Traffic-HKD	0.91	0.90	0.91	0.90	0.93	0.91	0.10	0.0816	0.81
HERALD	CIRA-CIC-DoHBrw-2020	0.9999	0.9999	0.9999	0.9999	1.0000	0.9999	0.0001	-	0.9998
	DoH-DGA-Malware-HKD	0.9764	0.9795	0.9732	0.9763	0.9912	0.9732	0.0212	0.0236	0.9527
	DoH-Tunnel-Traffic-HKD	0.9587	0.9642	0.9573	0.9607	0.9832	0.9573	0.0367	0.0392	0.9215

The feature importance analysis reveals a distinct hierarchy in the influence of each algorithm. OCSVM emerges as the most influential component, contributing the highest proportion (48 percent) to the model's predictive capability. This suggests that the decision boundary learned by OCSVM effectively differentiates normal traffic from anomalies in the encrypted DNS setting. The high importance score of OCSVM may stem from its ability to generalize across diverse anomaly types, particularly in scenarios where normal traffic distributions are complex and non-stationary.

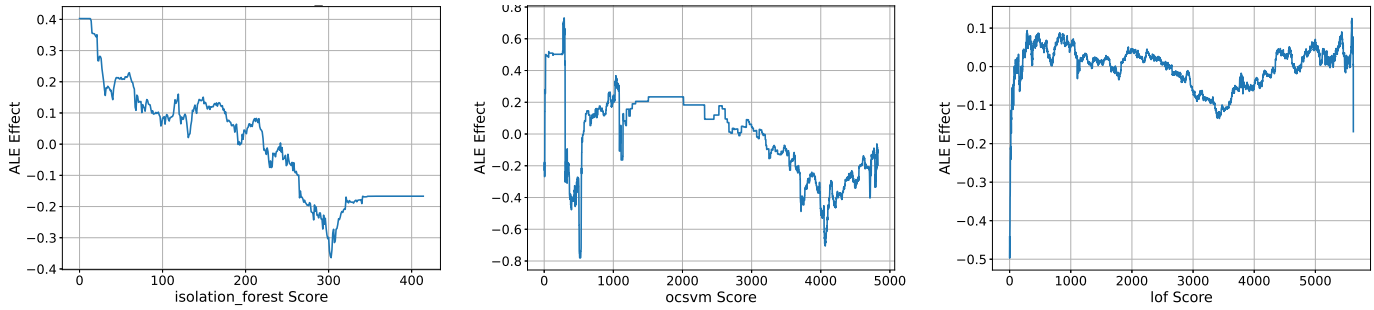
Following closely, IF demonstrates substantial influence (40 percent), reinforcing its effectiveness in detecting anomalies by isolating instances that deviate significantly from the majority of the data. IF's contribution highlights its robustness in handling encrypted DNS traffic, where anomalous patterns are often scattered across different regions of the feature space. The relatively high importance score of IF suggests that its partitioning-based anomaly detection mechanism is well-suited for this problem domain.

In contrast, LOF exhibits the lowest feature importance (12 percent), indicating that its local density-based anomaly detection approach contributes less significantly to the overall ensemble. While LOF has traditionally been effective in identifying anomalies based on local neighborhood deviations, its lower influence in this setting suggests that anomalies in encrypted DNS traffic may not always be well-captured by local density variations. This finding warrants further investigation into the parameter tuning of LOF or the potential integration of additional anomaly detection techniques that can better complement the strengths of IF and OCSVM.

This feature importance analysis underscores the pivotal role of OCSVM and IF in HERALD's anomaly detection framework, with LOF playing a more limited role. These insights provide a foundation for refining the ensemble by adjusting algorithmic weights or exploring alternative detection mechanisms to enhance performance.

6.2. Accumulated Local Effects Analysis

To complement the feature importance analysis, ALE plots (Fig. 8) were employed to examine the influence of individual features on the model's predictive behavior. These plots provide an interpretable visualization of feature effects across their value range while accounting for interactions with other features.



(a) Isolation Forest (b) One-Class SVM (c) Local Outlier Factor

Figure 8: Accumulated Local Effects (ALE) plots for base detectors used in HERALD. The plots show the effect of detector score variations on the final meta-model output probability, providing insights into how each detector contributes to the ensemble decision.

The ALE plot for the IF Score shows a monotonically decreasing trend, indicating that as the IF Score increases, its effect on the model prediction decreases. This behavior aligns with the IF methodology, where lower anomaly scores typically indicate more anomalous instances. The effect size is moderate, suggesting that while the IF Score influences the model's predictions, its impact is not dominant.

The ALE plot for the OCSVM Score reveals a highly nonlinear effect on model predictions. There are sharp fluctuations in the lower score range, suggesting that the OCSVM score strongly impacts anomaly detection for extreme values. However, as the score increases, the effect stabilizes, indicating that OCSVM's influence is more significant in detecting pronounced anomalies but diminishes for scores closer to the normal range.

For the LOF Score, the ALE plot demonstrates a mixed influence on model predictions. Initially, the effect is negative for low values, suggesting a contribution to anomaly detection. However, as the LOF score increases, the effect becomes more variable, indicating that LOF's influence on predictions is more complex and dataset-dependent. This behavior highlights the adaptive nature of LOF, which considers local density variations when identifying anomalies.

Overall, the ALE analysis underscores that OCSVM exhibits the most dynamic influence, particularly in detecting extreme anomalies. IF contributes more consistently but with a decreasing effect as scores rise. LOF shows a more variable and dataset-dependent effect, reflecting its sensitivity to local density structures. These insights enhance model interpretability and provide avenues for refining anomaly detection strategies.

6.3. Statistical Analysis of Detector Contributions

To quantify the influence of each anomaly detection model, statistical contributions were analyzed from the ALE results as shown in Table 5. The mean effect values suggest that IF (0.0020) has a slight positive contribution, while OCSVM (-0.0001) and LOF (-0.0000) have negligible average effects. However, their standard deviations (std_effect) indicate that OCSVM (0.2641) exhibits the highest variability in its effect, implying a more dynamic influence on the predictions.

The maximum and minimum effect values further highlight this variability, with OCSVM (max: 0.7316, min: -0.7810) showing the largest spread, suggesting that it strongly influences model decisions in both positive and negative directions. In contrast, LOF (max: 0.1251, min: -0.4971) has a more constrained effect, meaning that it contributes less significantly to model decisions.

Additionally, the stability metric suggests that IF (0.0094) is the most stable detector, while LOF (0.0023) is the least stable, indicating that LOF's impact on predictions fluctuates more depending on feature values.

To statistically assess differences between detectors, Kolmogorov-Smirnov (KS) tests were performed. The results reveal significant differences between all three detectors (p-value ≈ 0.0000 in all comparisons). The highest KS statistic (0.4659) between OCSVM and LOF suggests that these two detectors behave most differently in their feature contributions.

6.4. Local Interpretable Model-agnostic Explanations

In advancing the interpretability of HERALD, we employed LIME to analyze model predictions at the instance level. LIME provides localized explanations, shedding light on the specific contributions of individual features in each decision, thereby clarifying how HERALD differentiates between normal and anomalous encrypted DNS traffic.

LIME visualizations were generated for multiple instances to ensure a comprehensive analysis. The findings from three representative instances, corresponding to IF, OCSVM, and LOF, reveal distinct patterns in anomaly detection:

- The LIME plot in Fig. 9a shows that IF made a strong positive contribution (approximately +0.15) when its score was above -0.35, meaning it strongly signaled this case as anomalous. The LOF detector also contributed positively (+0.08) when its score exceeded 1.18, reinforcing the anomaly classification. OCSVM had minimal influence in this case, with only a slight contribution in its operating range. Overall, both IF and LOF aligned in their assessment, driving this instance toward being classified as an anomaly.

Table 5: Statistical Contributions and Significance Tests of Anomaly Detectors

Metric	IF	OCSVM	LOF
Mean Effect	0.0020	-0.0001	-0.0000
Std Effect	0.1764	0.2641	0.0530
Max Effect	0.4026	0.7316	0.1251
Min Effect	-0.3640	-0.7810	-0.4971
Effect Range	0.7667	1.5126	0.6222
Abs Mean Effect	0.1528	0.2203	0.0407
Stability	0.0094	0.0040	0.0023

Comparison	IF vs OCSVM	IF vs LOF	OCSVM vs LOF
KS Statistic	0.2028	0.395	0.4659
P-Value	0.0000	0.0000	0.0000

- In contrast, the LIME plot in Fig. 9b reveals a different dynamic between the detectors. OCSVM was the dominant influence, but with a significant negative contribution (approximately -0.08), pushing against classifying this as an anomaly. This occurred when OCSVM's score was quite low (3099.54), outside its typical range for anomalies. Meanwhile, LOF provided a moderate positive contribution (+0.02) within a specific score range (1.06-1.18), and IF showed a small positive effect (+0.06) when its score was below -0.43. This instance demonstrates how HERALD resolves competing signals, with OCSVM's strong negative influence partially offset by the positive contributions from the other detectors.

- The instance in Fig. 9c highlights a complex interaction between all three detectors. LOF made a substantial positive contribution (+0.06), strongly suggesting an anomaly. However, both OCSVM and IF countered this assessment with negative contributions (-0.03 and -0.05 respectively). OCSVM's negative influence occurred when its score was between 17505.44 and 17505.61, while IF's negative contribution appeared when its score was below 0.43. This case illustrates HERALD's ability to weigh conflicting signals from different detection methods, with LOF's anomaly indication partially counterbalanced by opposing signals from both OCSVM and IF.

These LIME insights complement the broader feature importance and ALE analyses by offering granular explanations for individual decisions. The results underscore HERALD's capacity to integrate diverse anomaly detection mechanisms, leveraging the RF meta-model's ability to synthesize heterogeneous signals dynamically. Furthermore, the interplay between base detectors emphasizes the importance of feature space calibration, particularly for OCSVM, which exhibited high sensitivity across different instances.

6.5. Synthesis of Interpretability Analysis

A comprehensive understanding of HERALD's decision-making process requires integrating multiple interpretability techniques: Feature Importance, ALE, and LIME. Each method illuminates distinct yet complementary aspects of the model's behavior. Feature Importance quantifies each base detector's overall contribution, ALE reveals their global effects across different score ranges, and LIME provides granular insights into specific instance classifications.

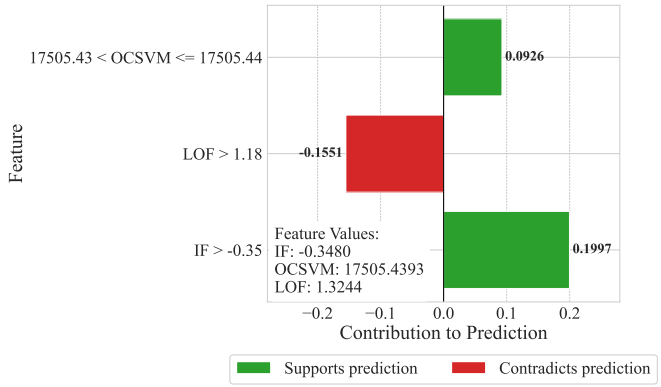
The Feature Importance analysis establishes a clear influence hierarchy among base detectors. OCSVM emerges as the most influential component (0.48 importance), followed closely by IF (0.40), with LOF contributing substantially less (0.12). This distribution indicates that while OCSVM and IF serve as primary drivers of HERALD's classification outcomes, all three detectors play meaningful roles in the ensemble.

ALE plots (Fig. 8) reveal more nuanced patterns in how each detector's influence varies across different score ranges. OCSVM exhibits the most dramatic effect fluctuations (ranging from -0.78 to 0.73), with particularly strong responses at extreme values. This suggests OCSVM excels at distinguishing clear anomalies from normal traffic. IF demonstrates a more gradual transition from positive to negative effects as scores increase, indicating a more consistent but evolving influence. LOF, despite having the smallest overall impact, shows complex oscillating patterns, confirming its contribution is highly context-dependent and particularly valuable for specific traffic patterns.

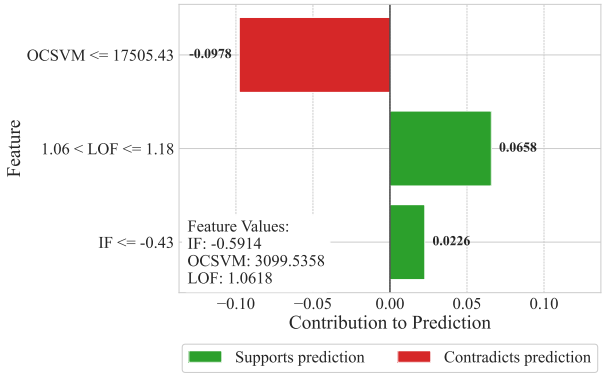
The statistical analysis reinforces these observations, with OCSVM showing the highest standard deviation (0.2641) and widest effect range (1.5126), while IF demonstrates the highest stability (0.0094). These metrics quantify OCSVM's dynamic response capabilities versus IF's more stable contribution pattern.

At the instance level, LIME plots (Fig. 9) illustrate how detector contributions vary across specific cases. In some instances, IF and LOF align to reinforce a classification decision (as in Instance 1). In others, OCSVM may provide a strong signal in opposition to the other detectors (Instance 2), or the detectors may exhibit competing influences that create a more balanced decision process (Instance 3). These varied interaction patterns demonstrate HERALD's adaptability to different anomaly manifestations.

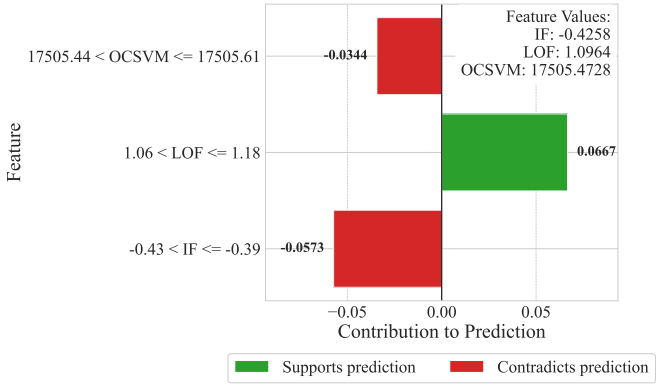
By synthesizing these interpretability findings, a key characteristic of HERALD becomes evident: its ability to balance strong global influences with localized adaptability. OCSVM provides powerful discrimination for clear anomalies, IF offers consistent and gradually evolving signals across the score spectrum, and LOF contributes targeted refinements for specific data patterns. This complementary integration ensures the model



(a) Instance 1



(b) Instance 2



(c) Instance 3

Figure 9: Local interpretable model-agnostic explanations (LIME) interpretability plots for individual predictions in HERALD.

leverages each detector's strengths while compensating for individual weaknesses.

HERALD's ensemble architecture effectively combines these diverse detection signals through a meta-model that dynamically weighs their contributions according to the specific characteristics of each traffic instance. This adaptive approach enables superior performance across diverse network environments and threat scenarios, optimizing both predictive accuracy and operational reliability. Moreover, the interpretability mechanisms integrated into HERALD provide security analysts with transparent insights into detection decisions, enhancing trust and facilitating effective response to potential threats.

7. Concluding Remarks

This study introduces HERALD, a novel hybrid ensemble approach for detecting anomalies within encrypted DNS traffic. By integrating unsupervised base detectors (IF, OCSVM, and LOF) with a supervised RF meta-model, HERALD effectively leverages the complementary strengths of both learning paradigms. Our comprehensive evaluation demonstrates HERALD's exceptional performance, achieving 99.99 percent across accuracy, precision, recall, and F1-score metrics, while maintaining competitive computational efficiency with 110s training time and 2.2ms inference time. The cross-dataset evaluation particularly highlights HERALD's superior generalization capabilities, exhibiting only 2-4 percent performance degradation when tested on previously unseen attack patterns compared to 5-8 percent for purely supervised models. This robustness to novel threats represents a significant advancement for operational security systems that must adapt to evolving attack techniques.

Our interpretability analysis illuminates the inner workings of HERALD, revealing that OCSVM serves as the primary driver of classifications, contributing the highest influence to the model's decisions, followed by IF, with LOF playing a more supportive role. The synthesis of these interpretability findings demonstrates HERALD's ability to balance strong global influences with localized adaptability, dynamically integrating diverse detection signals based on specific traffic characteristics.

However, the transition from experimental validation to operational deployment requires careful consideration of several practical factors. HERALD's deployment necessitates sufficient computational resources, as the 2.2ms inference time al-

lows processing approximately 450 samples per second per thread, potentially requiring distributed processing or hardware acceleration for high-volume enterprise networks. Furthermore, seamless integration with existing SIEM systems demands standardized APIs and compatible data formats. Additionally, operational deployment requires continuous access to representative training data reflecting current threat landscapes, necessitating automated data collection pipelines while maintaining privacy compliance.

While HERALD demonstrates exceptional performance in controlled settings, several limitations must be acknowledged. The framework's performance is intrinsically linked to training data quality, with evaluation primarily relying on CIRA-CIC-DoHBrw-2020, which may not fully capture real-world DoH traffic diversity across different organizational contexts. The current implementation depends on 34 pre-defined statistical features that may not capture all behavioral patterns and are sensitive to network infrastructure variations. Temporal validity concerns arise from training on specific time-period datasets, as rapid evolution of attack techniques may require more frequent updates than anticipated. Moreover, despite employing SMOTE for class balance, performance under extreme class imbalances or novel attack distributions remains uncertain.

Future research should focus on the following directions. First, validating and adapting the framework across more diverse datasets would further enhance its generalizability and robustness. Second, incorporating incremental learning capabilities would address the need for continuous adaptation in the face of evolving network threats. Third, integrating additional unsupervised detectors and advanced feature selection techniques could improve HERALD's scalability and efficiency, particularly in resource-constrained environments. Finally, expanding HERALD's application to other types of encrypted traffic would broaden its utility in comprehensive network security architectures.

References

[1] N. Shah, The challenges of inspecting encrypted network traffic: Fortinet (Aug 2020).
URL https://www.fortinet.com/blog/industry-trends/keeping-up-with-performance-demands-of-encrypted-web-traffic

[2] Introduction to DNS Privacy.
URL https://www.internetsociety.org/resources/deploy360/dns-privacy/intro/

[3] Z. Hu, L. Zhu, J. Heidemann, A. Mankin, D. Wessels, P. Hoffman, Specification for dns over transport layer security (tls), Tech. rep. (2016).

[4] P. Hoffman, P. McManus, Dns queries over https (doh), Tech. rep. (2018).

[5] R. Mitsuhashi, Y. Jin, K. Iida, T. Shinagawa, Y. Takai, Malicious dns tunnel tool recognition using persistent doh traffic analysis, IEEE Transactions on Network and Service Management 20 (2023) 2086–2095. doi:10.1109/tnsm.2022.3215681.

[6] C. Deccio, J. Davis, Dns privacy in practice and preparation, in: Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies, 2019, pp. 138–143.

[7] S. García, K. Hynek, D. Vekshin, T. Čejka, A. Wasicek, Large scale measurement on the adoption of encrypted dns, arXiv preprint arXiv:2107.04436 (2021).

[8] T. H. Kim, D. S. Reeves, A survey of domain name system vulnerabilities and attacks, Journal of Surveillance, Security and Safety (2020). doi:10.20517/jsss.2020.14.

[9] S. Singanamalla, S. Chunhapanya, M. Vavruša, T. Verma, P. Wu, M. Fayed, K. Heimerl, N. Sullivan, C. Wood, Oblivious dns over https (odoh): A practical privacy enhancement to dns, arXiv preprint arXiv:2011.10121 (2020).

[10] M. Lyu, H. H. Gharakheili, V. Sivaraman, A survey on dns encryption: Current development, malware misuse, and inference techniques, ACM Computing Surveys 55 (8) (2022) 1–28.

[11] Y. Wang, A. Zhou, S. Liao, R. Zheng, R. Hu, L. Zhang, A comprehensive survey on dns tunnel detection, Computer Networks 197 (2021) 108322.

[12] K. Hynek, D. Vekshin, J. Luxemburk, T. Cejka, A. Wasicek, Summary of dns over https abuse, IEEE Access 10 (2022) 54668–54680.

[13] M. MontazeriShatoori, L. Davidson, G. Kaur, A. H. Lashkari, Detection of doh tunnels using time-series classification of encrypted traffic, 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf (2020). doi:10.1109/dasc-picom-cbdcom-cyberscitech49142.2020.00026.

[14] Y. M. Banadaki, Detecting malicious dns over https traffic in domain name system using machine learning classifiers, Journal of Computer Sciences and Applications 8 (2020) 46–55. doi:10.12691/jcsa-8-2-2.

[15] M. Lyu, H. H. Gharakheili, V. Sivaraman, Classifying and tracking enterprise assets via dual-grained network behavioral analysis, Computer Networks 218 (2022) 109387.

[16] M. Lyu, H. H. Gharakheili, C. Russell, V. Sivaraman, Enterprise dns asset mapping and cyber-health tracking via passive traffic analysis, IEEE Transactions on Network and Service Management 20 (3) (2022) 3699–3716.

[17] M. Behnke, N. Briner, D. Cullen, K. Schwerdtfeger, J. Warren, R. Basnet, T. Doleck, Feature engineering and machine learning model comparison for malicious activity detection in the dns-over-https protocol, IEEE Access 9 (2021) 129902–129916.

[18] D. Vekshin, K. Hynek, T. Cejka, Doh insight: Detecting dns over https by machine learning, in: Proceedings of the 15th International Conference on Availability, Reliability and Security, 2020, pp. 1–8.

[19] O. Abualghanam, H. Alazzam, B. Elshqeirat, M. Qatawneh, M. A. Almaiah, Real-time detection system for data exfiltration over dns tunneling using machine learning, Electronics 12 (6) (2023) 1467.

[20] H. R. Ibraheem, N. D. Zaki, M. I. A. Al-Mashhadani, Anomaly detection in encrypted https traffic using machine learning: a comparative analysis of feature selection techniques, Mesopotamian Journal of Computer Science (2022) 17–28doi:10.58496/mjcsc/2022/005.

[21] A. R. Alzighaibi, Detection of doh traffic tunnels using deep learning for encrypted traffic classification, Computers 12 (2023) 47. doi:10.3390/computers12030047.

[22] Q. A. Al-Haija, M. Alohaly, A. Odeh, A lightweight double-stage scheme to identify malicious dns over https traffic using a hybrid learning approach, Sensors 23 (2023) 3489. doi:10.3390/s23073489.

[23] T. Zebin, S. Rezvy, Y. Luo, An explainable ai-based intrusion detection system for dns over https (doh) attacks, IEEE Transactions on Information Forensics and Security 17 (2022) 2339–2349.

[24] R. Huang, A doh traffic detection system based on interpretable deep learning neural networks, in: 2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE), IEEE, 2024, pp. 802–807.

[25] L. F. G. Casanova, P.-C. Lin, Generalized classification of dns over https traffic with deep learning, in: 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2021, pp. 1903–1907.

[26] T. A. Nguyen, M. Park, Doh tunneling detection system for enterprise network using deep learning technique, Applied Sciences 12 (5) (2022) 2416.

[27] G. D'Angelo, A. Castiglione, F. Palmieri, Dns tunnels detection via dns-images, Information Processing & Management 59 (3) (2022) 102930.

[28] J. Liang, S. Wang, S. Zhao, S. Chen, Fecc: Dns tunnel detection model based on cnn and clustering, Computers & Security 128 (2023) 103132.

[29] P. Kadebu, R. T. Shoniwa, K. Zvarevashe, A. Mukwazvure, I. Mapanga, N. F. Thusabantu, T. T. Gotora, A hybrid machine learning approach for analysis of stegomalware, International Journal of Industrial Engineering and Operations Management (ahead-of-print) (2023).

[30] L. K. Lok, V. A. Hameed, M. E. Rana, Hybrid machine learning approach for anomaly detection, Indonesian Journal of Electrical Engineering and Computer Science 27 (2) (2022) 1016.

[31] P. Gogoi, B. Borah, D. K. Bhattacharyya, Anomaly detection analysis of intrusion data using supervised & unsupervised approach., J. Conver-

gence Inf. Technol. 5 (1) (2010) 95–110.
[32] V. T. Gowda, Credit card fraud detection using supervised and unsupervised learning, in: CS & IT Conference Proceedings, Vol. 11, CS & IT Conference Proceedings, 2021.
[33] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, G. Bontempi, Combining unsupervised and supervised learning in credit card fraud detection, Information sciences 557 (2021) 317–331.
[34] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 eighth ieee international conference on data mining, IEEE, 2008, pp. 413–422.
[35] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson, Estimating the support of a high-dimensional distribution, Neural computation 13 (7) (2001) 1443–1471.
[36] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 93–104.
[37] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, M. Woniak, Ensemble learning for data stream analysis: a survey, Information Fusion 37 (2017) 132–156. doi:10.1016/j.inffus.2017.02.004.
[38] A. Zimek, R. J. Campello, J. Sander, Ensembles for unsupervised outlier detection: challenges and research questions a position paper, Acm Sigkdd Explorations Newsletter 15 (1) (2014) 11–22.
[39] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[40] M. MontazeriShatoori, L. Davidson, G. Kaur, A. H. Lashkari, Detection of doh tunnels using time-series classification of encrypted traffic, in: 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), IEEE, 2020, pp. 63–70.
[41] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.
[42] H. He, Y. Bai, E. A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), Ieee, 2008, pp. 1322–1328.
[43] P. Wang, S. Li, F. Ye, Z. Wang, M. Zhang, Packetcgan: Exploratory study of class imbalance for encrypted traffic classification using cgan, in: ICC 2020-2020 IEEE International Conference on Communications (ICC), IEEE, 2020, pp. 1–7.
[44] A. Bansal, M. Saini, R. Singh, J. K. Yadav, Analysis of smote, International Journal of Information Retrieval Research 11 (2021) 15–37. doi:10.4018/ijirr.2021040102.
[45] R. Mitsuhashi, Y. Jin, K. Iida, T. Shinagawa, Y. Takai, Detection of dga-based malware communications from doh traffic using machine learning analysis, in: 2023 IEEE 20th Consumer Communications & Networking Conference (CCNC), IEEE, 2023, pp. 224–229.
[46] R. Mitsuhashi, Y. Jin, K. Iida, T. Shinagawa, Y. Takai, Malicious dns tunnel tool recognition using persistent doh traffic analysis, IEEE Transactions on Network and Service Management (2022).