SafeSurf Darknet 2025: A Novel Dataset for Darknet Traffic Detection and Analysis

Qasem Abu Al-Haija *, Mohammad J. Obaidat , Ibrahim A. Al-Syouf , Yahea F. Awawdeh , Anas E. Masa'deh

Posted Date: 24 July 2025

doi: 10.20944/preprints202507.1926.v1

Keywords: darknet traffic detection; cyber threat intelligence; network security; anomaly detection; machine learning for cybersecurity

 doi:10.20944/preprints202507.1926.v1

Article

SafeSurf Darknet 2025: A Novel Dataset for Darknet Traffic Detection and Analysis

Qasem Abu Al-Haija *, Mohammad J. Obaidat, Ibrahim A. Al-Syouf, Yahea F. Awawdeh and Anas E. Masa'deh

Cybersecurity Department, Jordan University of Science and Technology, Irbid 22110, Jordan; qsabuhaija@just.edu.jo

Abstract

The growing threat of darknet-related activities, ranging from illegal marketplaces to command-and-control infrastructures, has made the accurate identification of darknet traffic a critical concern for cybersecurity professionals. In response to the lack of high-quality, well-labeled datasets in this domain, we present a newly created darknet traffic dataset to support research and analysis efforts in network security. The dataset was developed to address data availability, consistency, and challenges with labeling accuracy. It comprises around 92 megabytes of traffic data on the first layer and 35 megabytes of traffic data on the second and third layer, including nearly 253K individual flows and 79 distinct features (source/destination IPs, ports, protocols, timestamps, etc.), Each entry is labeled according to its nature as darknet or non-darknet traffic in the first layer, and further labeled by darknet type and behavior in the second and third layers, respectively. Potential applications include threat intelligence research, network traffic analysis, and testing security tools and policies. The dataset has a comprehensive three-layered label, indicating its relevance and practical utility for understanding darknet traffic behavior in various applications.

Keywords: darknet traffic detection; cyber threat intelligence; network security; anomaly detection; machine learning for cybersecurity

1. Introduction

The dark web, commonly accessed via specialized networks like Tor and I2P, presents significant challenges for cybersecurity due to its anonymous and unregulated nature. While advanced security tools increasingly protect the surface web, the darknet remains a haven for illicit activities, ranging from illegal marketplaces to ransomware communications. As cyber threats become more sophisticated, the need for accurate and granular classification of darknet traffic grows urgently. Despite a growing body of research in this field, progress is limited by the lack of comprehensive, labeled datasets tailored specifically for darknet traffic behaviors and types.

Existing datasets, such as CIC-Darknet2020 and CTU-13, often have limitations, including a lack of class diversity, inadequate labeling, and restricted availability of raw traffic data. Moreover, many focus primarily on detection—merely distinguishing between normal and malicious traffic—without addressing the nuanced classification of darknet traffic into distinct behavioral categories. This limitation hampers the development of robust machine learning models capable of detecting darknet activity and understanding its underlying patterns.

To address these gaps, we present a novel darknet dataset specifically designed for traffic classification. The dataset was constructed by capturing real-world darknet traffic across multiple machine environments, including physical and virtual systems operating on conventional networks. Each type of darknet behavior was intentionally invoked and recorded to ensure the traffic's authenticity, and labeling was performed accordingly. This process yielded a diverse and well-structured dataset for traffic detection, behavior-based classification, and pattern analysis.

 doi:10.20944/preprints202507.1926.v1

In addition to its comprehensiveness, the dataset has been evaluated using several machine learning models, demonstrating high classification accuracy (exact values to be inserted). The dataset is also intended for public release, offering researchers a valuable resource for advancing darknet traffic analysis. This work contributes a foundational asset for future research and developing more precise cybersecurity tools by bridging the gap between detection and detailed classification.

Darknet traffic detection and classification have received increasing attention in recent years, particularly with the rise of encrypted and decentralized communication platforms. Various studies have explored protocol-specific behaviors, developed classification models, and constructed datasets to identify darknet-related activities. However, existing efforts regarding behavioral diversity, platform scope, or real-world applicability are often limited.

1.1. Datasets and Traffic Classification Models

Several studies have utilized existing darknet datasets or created new ones to train machine learning models for traffic detection and classification. In [1], the authors focused on classifying **Tor traffic** using time-related features like flow duration and inter-arrival time. By employing a Random Forest (RF) model on the **UNB-CIC Tor network traffic dataset**, they achieved a classification accuracy of 98% between Tor and non-Tor flows. In [2], the authors proposed a machine learning-based intrusion detection system for **IoT networks** using the **CIC-Darknet2020** dataset. They evaluated six supervised algorithms and found that **BAG-DT** achieved the highest accuracy of 99.5%, indicating the effectiveness of ensemble models for darknet detection. A more comprehensive multi-platform approach was taken. [3], where the authors created a **custom dataset** that includes traffic from **Tor, I2P, ZeroNet, and Freenet**. They applied both flat and hierarchical classification techniques using CICFlowMeter-extracted features. Among various models tested—such as XGBDT, LightGBM, MLP, and LSTM—XGBDT achieved the highest performance with an accuracy of 99.42% in darknet behavior classification. This work demonstrated strong model performance but used a limited feature set (26 features), which may restrict fine-grained behavior analysis. In [4], a novel deep learning system called **Tor-VPN Detector** was proposed to classify darknet traffic into four categories: Tor, non-Tor, VPN, and non-VPN. Using the **DIDarknet dataset**, which contains over 141,000 samples and 79 flow-level features, the authors trained a six-layer deep neural network that achieved a 96% accuracy and 96.06% F1-score, without requiring preprocessing or balancing techniques.

1.2. Protocol-Specific Behavioral Analysis

Beyond detection accuracy, some studies have explored specific darknet technologies in greater technical and behavioral detail. In [5], the authors analyzed **Tor artifacts** on Windows 11 systems, proposing a virtual lab setup to simulate cyberattacks and generate traffic for forensic and machine learning research. Although methodologically strong, this work emphasized operating system traces more than network flow analysis.

In [6], the researchers studied the architecture and mechanisms of the **I2P protocol**, including NTCP communication, AES-256 encryption, and tunnel routing. They used **Tranalyzer2** to extract flow features and tested multiple supervised classifiers, finding **Random Forest** the most effective. However, their focus remained largely on technical protocol behavior rather than creating a reusable dataset.

Study [7] Focused on **ZeroNet** architecture and its robustness challenges, highlighting its dependency on centralized trackers and proposing peer-exchange improvements. While useful for understanding platform limitations, the work did not include traffic classification or dataset contributions.

1.3. Foundational and Contextual Studies

In [8], a broader perspective on the **dark web's dual-use nature** was provided, emphasizing its legitimate role in preserving privacy and anonymity and its risks in enabling criminal activity. The

 doi:10.20944/preprints202507.1926.v1

study reinforces the need for advanced detection tools and datasets to support privacy-preserving and threat-identifying research.

1.4. Gaps in Existing Work

Despite these efforts, current studies often fail to capture the **behavioral diversity** and **multi-protocol coverage** required for robust darknet traffic analysis. Most datasets are centered on a single platform (e.g., Tor) and focus on binary detection tasks (darknet vs. non-darknet), leaving a resource gap that enables behavior-level classification across multiple darknet ecosystems.

1.5. Contribution of This Work

The **SafeSurf Darknet 2025** dataset addresses these limitations by introducing a behaviorally labeled, flow-based dataset encompassing five key technologies: **Tor, I2P, Freenet, ZeroNet, and VPN**. It includes **79 flow-level features** and labels corresponding to user behaviors (e.g., email, chat, file transfer, video streaming), making it suitable for multi-class classification, behavior profiling, and anomaly detection. In contrast to prior datasets, it emphasizes **diversity** and **structure**, offering a practical resource for researchers and practitioners aiming to understand or secure darknet environments [8].

1.6. Paper Structure

The remainder of this paper is organized as follows: Section 2 outlines the methodology and data collection procedures used in building the dataset. Section 3 details the dataset's structure and labeling approach. Section 4 discusses the evaluation results using machine learning models. Finally, Section 5 concludes the paper and outlines potential future directions.

2. Dataset Description

2.1. Dataset Collection Methodology

We run darknet activity on a Raspberry Pi device and multiple virtual machine environments to develop a comprehensive and realistic darknet traffic dataset. Traffic was captured using **Wireshark** and then processed using **CICFlowMeter** to convert raw Pcap files into structured flow-based CSV data with 79 extracted features per flow. This methodology ensured depth and structure, facilitating machine learning applications such as classification and anomaly detection.

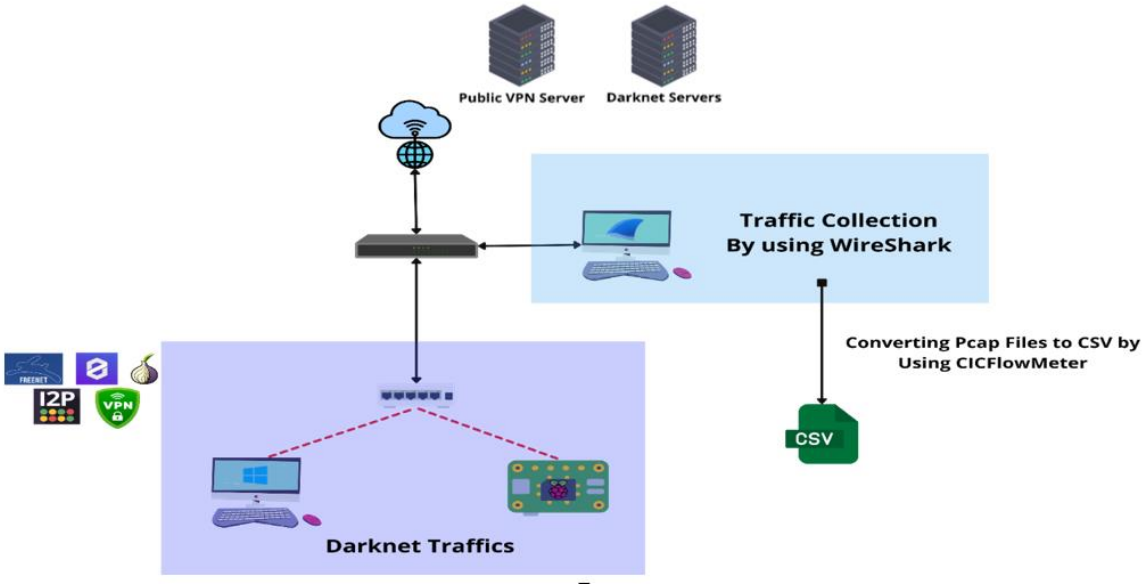


Figure 1. Dataset Capture Methodology.

 doi:10.20944/preprints202507.1926.v1

Multiple platforms and protocols representing diverse darknet behaviors were used in data collection. These included:

- **Tor (The Onion Router):** We configured Tor on a Raspberry Pi 5 (Kali Linux), Windows machines, and Azure-based VPNs. Behaviors captured included web browsing, email via Thunderbird (through proxychains), chatting and VoIP via Telegram, file transfers, audio, and video streaming (including YouTube).
- **Freenet:** We used the Freenet platform to simulate browsing, file sharing, chatting (via FMS), video streaming (via FreeTube), and decentralized email (via Web of Trust).
- **ZeroNet:** Deployed on a Windows 10 environment, this protocol allowed the collection of traffic related to ZeroMail, ZeroChat, ZeroTube, decentralized file transfers (IFS), and general browsing behaviors through peer-to-peer hosting.
- **VPN:** An OpenVPN server was deployed on Microsoft Azure to simulate encrypted communication and traffic redirection. Services accessed included Spotify (audio), YouTube (video), Discord, WhatsApp, and Telegram for chat, VoIP, and file transfer sessions.
- **I2P (Invisible Internet Project):** We installed I2P using Docker, simulating access to hidden services like ramble.i2p and i2pforum.i2p. Behaviors recorded included browsing, email via i2pmail, chatting via i2pchat, video streaming (via invidious front-end), and torrent-based file transfers.

Traffic was generated manually to reflect realistic usage patterns, capturing user-initiated behaviors in various darknet ecosystems. Some automation (e.g., repeated uploads/downloads) was used in the VPN file transfer scenario to ensure sufficient data volume. Approximately **127 MB of traffic** was collected, resulting in layer One **~253,000 flows** and **~91,000 flows** in layer 2 and 3, as shown in **Error! Reference source not found.**, each annotated with **79 distinct features**. This setup offers a wide behavioral representation while maintaining a manageable dataset size for experimentation and analysis.

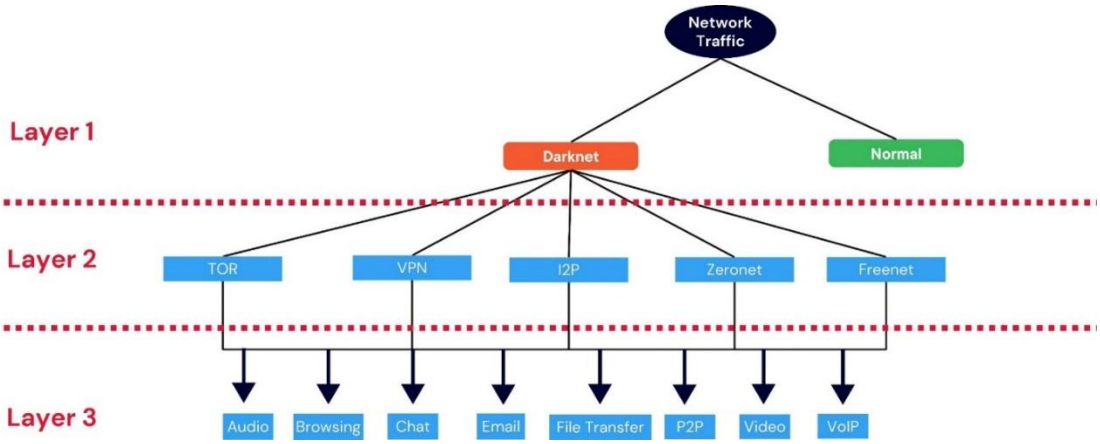


Figure 2. Hierarchical Classification Method.

2.2. Dataset Characteristics

Labeling was achieved by **manually assigning a behavior class to each traffic capture session**, based on the known generation context. For instance, if a capture session involved streaming a video over Tor, it was labeled accordingly. No automated labeling tools were used; the behavior was isolated by setup and traffic timing, ensuring ground-truth integrity. The dataset is organized around **behavioral classes** rather than protocol or port distinctions. Across the Tor, Freenet, I2P, Zeronet, and VPN environments, we defined and captured traffic for nine behaviors: browsing, Email, chatting, VOIP, file transfer, audio streaming, video streaming, and P2P sharing. Each class has

 doi:10.20944/preprints202507.1926.v1

multiple examples derived from different darknet technologies to allow for intra-class variation and robust classifier training. For example, video streaming traffic includes captures from YouTube over Tor, FreeTube on Freenet, and ZeroTube on ZeroNet, each exhibiting underlying traffic signatures.

The dataset is structured in CSV format, with each row representing a single network flow and labeled according to the behavior class. The features include timestamp, flow duration, packet statistics, byte counts, inter-arrival times, and header information. This structure is intended to support tasks such as classifying darknet behaviors, traffic profiling, and anomaly detection in academic and operational cybersecurity research.

2.3. Data Annotation and Labeling

Due to availability issues, we did not cover all nine behaviors in all darknet types in the table. The captured behaviors for each darknet type are shown below in Table 1.

Table 1. Captured Behaviors.

Darknet type	Behaviors captured
TOR	Browsing, Email, Chatting, VOIP, File Transfer, Streaming
Freenet	Browsing, Chatting, Email, File Sharing, Video
Zeronet	Email, Chatting, File Sharing, Video
VPN	Streaming, File Transfer, Chatting, VOIP
I2P	Browsing, Email, Chatting, File Sharing, Video

The following table, Table 2, shows the label counts of the layer one dataset:

Table 2. Layer 1 Label Counts.

Label	Count
Normal	360,358
Darknet	91,404

The following table, Table 3, shows the label counts of the layer two dataset:

Table 3. Layer 2 Label Count.

Label	Count
Freenet	26,284
Zeronet	25,499
I2P	22,958
Tor	12,546
VPN	4,117

The following table, **Error! Reference source not found.**, shows the label counts of the layer three dataset:

Table 4. Layer 3 Label Count.

Label	Count
Browsing	33,586
FTP	20,214
Video	9,559
P2P	9,392
Email	7,873
Audio	5,953

 doi:10.20944/preprints202507.1926.v1

Chat	3,489
VOIP	1,338

The following pie charts in **Error! Reference source not found.** Illustrate each label's distribution in the three layers of the captured data.

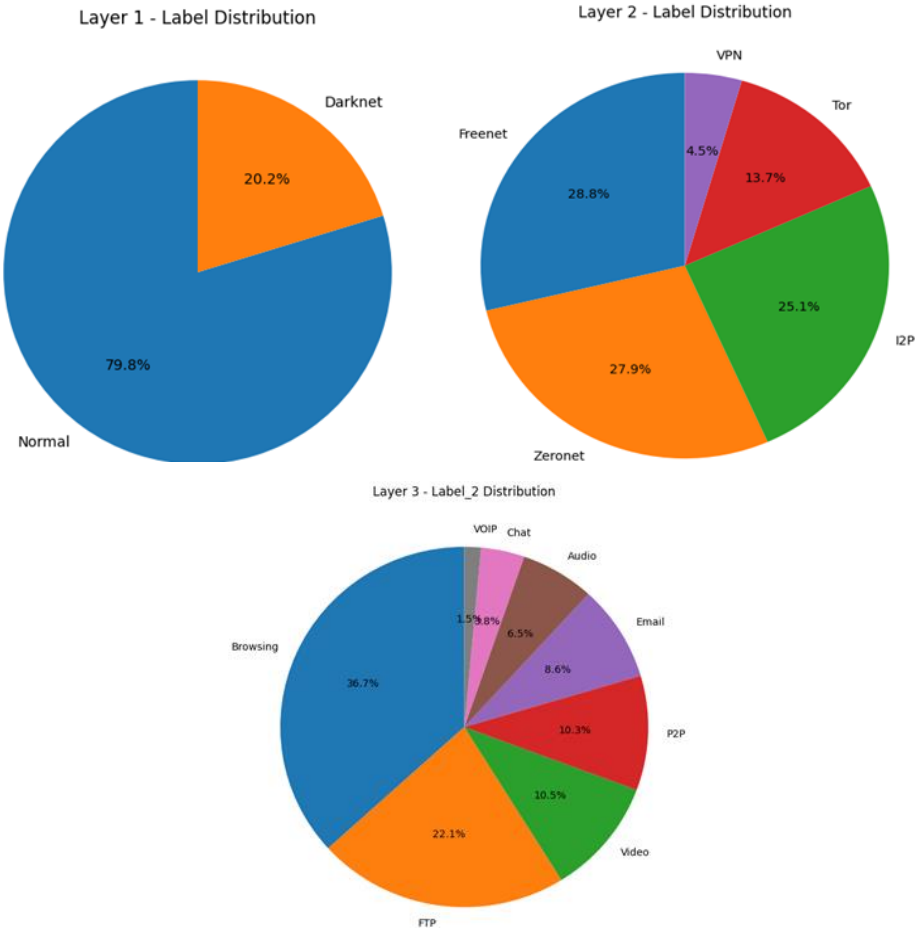


Figure 3. Traffic Distribution Charts.

3. Experimental Setup and Benchmarking

3.1. Data Preprocessing

The preprocessing phase involved several key steps to prepare the dataset from all three layers for machine learning:

- **Feature Inspection**: All dataset features across the three layers were initially examined.
- **Handling Missing Values**: Null values were identified and removed to ensure data completeness.
- **Removing Duplicates**: Duplicate records were checked and addressed.
- **Zero-Value Features**: Features with only zero values were detected and excluded from the dataset.
- **Filtering Misclassified and Non-Relevant Flows**: (a) Traffic involving well-known DNS servers (e.g., Google, Cloudflare) was removed. (b) Communications between private IPs were also excluded.
- **Eliminating String-Based Features**: Features containing non-numeric (string) values unsuitable for machine learning were removed.

 doi:10.20944/preprints202507.1926.v1

- **Feature Selection via ANOVA**: (a) Analysis of Variance (ANOVA) was applied to select the most relevant numerical features. (b) Features with a p-value ≤ 0.05 were retained as they showed statistically significant differences between label groups.

3.2. Machine Learning Training

In this section, cleaned and pre-processed data are trained through various machine learning models, including the following:

- **Gaussian Naive Bayes (GNB) is a type of Naive Bayes method based on continuous attributes and the data features that follow a** Gaussian distribution throughout the dataset. This "naive" assumption simplifies calculations and makes the model fast and efficient. Gaussian Naive Bayes is widely used because it performs well even with small datasets and is easy to implement and interpret
- **Decision Tree Classifier (DT):** is a flowchart-like tree structure where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome.
- **Random Forest (RF)** is a method that combines the predictions of multiple decision trees to produce a more accurate and stable result. It can be used for both classification and regression tasks.
- The random forest classifier has two main parameters that can be adjusted to maximize prediction accuracy: the number of estimators and the maximum leaf nodes. The following table, **Error! Reference source not found.**, represents the best parameters of each layer, which we determined through trial and comparison.

Table 5. Best RF Parameters.

Layer	N Estimators	Maximum Leaf Nodes
Layer 1	300	50
Layer 2	300	50
Layer 3	100	50

- **Logistic Regression** is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm that analyzes the relationship between two data factors.
- **XGBoost** is a high-performance machine learning algorithm that utilizes gradient boosting and ensemble learning techniques to optimize predictive accuracy and computational speed by efficiently combining multiple decision trees for improved model performance.
- **Multilayer Perceptron (MLP):** is an artificial neural network consisting of multiple layers of interconnected nodes, each performing specific computations to process input data and make predictions.
- **A Support Vector Machine (SVM)** is a supervised machine learning algorithm for classification and regression tasks. It finds the optimal hyperplane (or line in 2D space) that best separates different classes in a dataset, maximizing the margin between them.
- **Bagging Decision Trees (Bag-DT)** is an ensembled modelling technique, used to combine multiple Decision Trees in parallel into memory, to build a stronger model

 doi:10.20944/preprints202507.1926.v1

- **Ensemble Learning (DT, RF, XGBoost):** is a method where we use many small models instead of just one. Each model may not be strong, but when we combine their results, we get a better and more accurate answer. It's like asking a group of people for advice instead of just one person. Each one might be a little wrong, but together, they usually give a better answer

Testing and evaluation. The table summarizes the classification accuracy of various machine learning models across three hierarchical layers. **Decision Tree** achieved the highest accuracy in Layer 1 (**99.46%**) and Layer 3 (**84.93%**), while **XGBoost** slightly outperformed others in Layer 2 with **96.21%** accuracy. Ensemble methods and Random Forest also demonstrated strong performance, especially in Layers 1 and 2. In contrast, models like **GaussianNB** and **Logistic Regression** underperformed across all layers, indicating their limited suitability for this task. Overall, tree-based models and ensemble techniques proved to be the most effective for this multi-layer classification problem as provided in Table 6.

Table 6. Machine Learning Models' Accuracies.

Model	Layer 1 Accuracy	Layer 2 Accuracy	Layer 3 Accuracy
Decision Trees	**0.994626903**	0.953850251	**0.849317619**
Ensemble (XGBoost, DT, RF)	0.992257785	0.957730939	0.80936156
XGBoost	0.991808736	**0.962068178**	0.804557895
Random Forests	0.976448181	0.870453508	0.535347893
MLP	0.972066088	0.906787399	0.598763708
Bagging-DT	0.953809944	0.863491097	0.570835428
SVM - Minmax Scaler	0.882535111	0.731281388	0.476884693
GaussianNB	0.667456373	0.385063156	0.178573423
Logistic Regression	0.64958734	0.365659717	0.188013182

Error! Reference source not found. Evaluates the computational efficiency of each model across the three classification layers. We measured the Training time in seconds. Simpler models like **Decision Tree** and **SVM (with MinMax Scaler)** demonstrated the lowest time complexity, completing execution in under 0.1 seconds across all layers. **GaussianNB** also maintained low execution times, though slightly higher in Layer 3 due to increased complexity. In contrast, ensemble methods such as **Bagging-DT** and **Ensemble (XGBoost, DT, RF)** required significantly more processing time, particularly in Layers 1 and 3, highlighting the trade-off between accuracy and computational cost. For instance, Bagging-DT took over 9 seconds in Layer 1.

Table 7. Machine Learning Models Performance.

Model	Layer 1 Time (s)	Layer 2 Time (s)	Layer 3 Time (s)
GaussianNB	0.121964693	0.143498898	0.418830156
Decision Trees	**0.045069218**	**0.024540663**	**0.057661057**
Random Forests	0.279803038	0.203600407	0.223201990
Bagging-DT	9.853935242	2.273411274	6.706970453
MLP	0.655432940	0.320156097	0.702657700
SVM - Minmax Scaler	0.017041683	0.021026850	0.062469721
XGBoost	0.053077459	0.092761278	0.219833136
Logistic Regression	0.047414303	0.038285732	0.107613802
Ensemble (XGBoost, DT, RF)	1.206273079	0.638847589	1.271958113

MLP and **Random Forests** offered a balanced performance in terms of time and accuracy, maintaining moderate training times below 1 second for most layers. This comparison helps choose models based on accuracy and resource efficiency, which are crucial in real-time or resource-constrained environments. The following tables, Tables 8–10, show the summary of the study, including the accuracy performance of various machine learning models across three classification

 doi:10.20944/preprints202507.1926.v1

layers and their corresponding training time (in seconds), highlighting the trade-offs between predictive performance and computational efficiency. Based on the results, the Decision Tree model was selected as the final choice due to its consistently high accuracy across all layers and low computational cost.

Table 8. Layer 1 - Per Class Metrics.

Model	Class	Precision	Recall	F1-score	Support
GaussianNB	Darknet	0.97	0.35	0.51	**32147**
	Normal	0.6	0.99	0.75	**32434**
Decision Trees	Darknet	1.0	0.99	0.99	**32147**
	Normal	0.99	1.0	0.99	**32434**
Random Forests	Darknet	0.97	0.98	0.98	**32147**
	Normal	0.98	0.97	0.98	**32434**
Bagging-DT	Normal	0.96	0.95	0.95	**32147**
	Darknet	0.95	0.96	0.95	**32434**
MLP	Normal	0.98	0.97	0.97	**32147**
	Darknet	0.97	0.98	0.97	**32434**
SVM	Darknet	0.97	0.8	0.87	**32147**
	Normal	0.83	0.97	0.89	**32434**
XGBoost	Normal	0.99	0.99	0.99	**32434**
	Darknet	0.99	0.99	0.99	**32147**
Ensemble	Normal	0.99	0.99	0.99	**32434**
	Darknet	0.99	0.99	0.99	**32147**
Logistic Regression	Darknet	0.63	0.74	0.68	**32147**
	Normal	**0.69**	**0.57**	**0.62**	**32434**

Table 9. Layer 2 - Per Class Metrics.

Model	Class	Precision	Recall	F1-score	Support
GaussianNB	Freenet	0.57	0.47	0.52	**5290**
	I2P	0.33	0.09	0.14	**5315**
	Tor	0.76	0.15	0.25	**5249**
	VPN	0.31	0.95	0.47	**5157**
	Zeronet	0.38	0.27	0.32	**5273**
Decision Trees	Freenet	0.96	0.96	0.96	**5290**
	I2P	0.93	0.93	0.93	**5315**
	Tor	0.92	0.92	0.92	**5249**
	VPN	0.99	0.99	0.99	**5157**
	Zeronet	0.96	0.97	0.96	**5273**
Random Forests	Freenet	0.94	0.91	0.93	**5290**
	I2P	0.86	0.74	0.8	**5315**
	Tor	0.81	0.83	0.82	**5249**
	VPN	0.91	0.98	0.94	**5157**
	Zeronet	0.83	0.89	0.86	**5273**
Bagging-DT	Tor	0.9	0.9	0.9	**5290**
	VPN	0.86	0.75	0.8	**5315**
	Freenet	0.78	0.86	0.82	**5249**
	I2P	0.88	0.93	0.91	**5157**
	Zeronet	0.88	0.85	0.87	**5273**
MLP	Tor	0.96	0.93	0.94	**5290**
	VPN	0.91	0.8	0.85	**5315**
	Freenet	0.83	0.91	0.87	**5249**
	I2P	0.93	0.99	0.96	**5157**
	Zeronet	0.89	0.89	0.89	**5273**
SVM	Freenet	0.71	0.87	0.78	**5290**
	I2P	0.82	0.41	0.55	**5315**
	Tor	0.74	0.67	0.7	**5249**
	VPN	0.74	0.91	0.82	**5157**
	Zeronet	0.7	0.79	0.74	**5273**
XGBoost	Tor	0.91	0.95	0.93	**5249**
	VPN	0.99	1.0	0.99	**5157**
	Freenet	0.98	0.96	0.97	**5290**
	I2P	0.95	0.94	0.94	**5315**
	Zeronet	0.98	0.97	0.97	**5273**
Logistic Regression	Freenet	0.37	0.44	0.4	**5290**
	I2P	0.26	0.47	0.34	**5315**

 doi:10.20944/preprints202507.1926.v1

	Tor	0.49	0.34	0.41	**5249**
	VPN	0.79	0.22	0.34	**5157**
	Zeronet	0.34	0.36	0.35	**5273**
Ensemble	Tor	0.89	0.96	0.92	**5249**
	VPN	0.98	1.0	0.99	**5157**
	Freenet	0.98	0.95	0.97	**5290**
	I2P	0.96	0.92	0.94	**5315**
	Zeronet	**0.98**	**0.96**	**0.97**	**5273**

Table 10. Layer 3 - Per Class Metrics.

Model	Class	Precision	Recall	F1-score	Support
GaussianNB	Audio	0.47	0.07	0.12	**6674**
	Browsing	0.2	0.03	0.06	**6593**
	Chat	0.27	0.04	0.07	**6809**
	Email	0.18	0.24	0.21	**6786**
	FTP	0.24	0.0	0.01	**6681**
	P2P	0.35	0.03	0.05	**6745**
	VOIP	0.16	0.92	0.27	**6798**
	Video	0.34	0.08	0.13	**6623**
Decision Trees	Audio	0.92	0.93	0.93	**6674**
	Browsing	0.8	0.8	0.8	**6593**
	Chat	0.86	0.85	0.86	**6809**
	Email	0.8	0.81	0.8	**6786**
	FTP	0.75	0.75	0.75	**6681**
	P2P	0.87	0.85	0.86	**6745**
	VOIP	0.97	0.97	0.97	**6798**
	Video	0.82	0.83	0.83	**6623**
Random Forests	Audio	0.69	0.75	0.72	**6674**
	Browsing	0.53	0.41	0.46	**6593**
	Chat	0.49	0.37	0.42	**6809**
	Email	0.44	0.36	0.4	**6786**
	FTP	0.44	0.26	0.32	**6681**
	P2P	0.45	0.6	0.52	**6745**
	VOIP	0.73	0.94	0.83	**6798**
	Video	0.43	0.59	0.49	**6623**
Bagging-DT	Audio	0.73	0.76	0.74	**6674**
	Chat	0.53	0.43	0.47	**6593**
	FTP	0.62	0.37	0.46	**6809**
	Browsing	0.46	0.37	0.41	**6786**
	Email	0.51	0.35	0.41	**6681**
	Video	0.49	0.69	0.57	**6745**
	P2P	0.7	0.95	0.81	**6798**
	VOIP	0.47	0.6	0.53	**6623**
MLP	Audio	0.76	0.75	0.76	**6674**
	Chat	0.5	0.57	0.54	**6593**
	FTP	0.55	0.53	0.54	**6809**
	Browsing	0.55	0.46	0.5	**6786**
	Email	0.47	0.37	0.41	**6681**
	Video	0.58	0.58	0.58	**6745**
	P2P	0.8	0.96	0.87	**6798**
	VOIP	0.54	0.59	0.57	**6623**
SVM	Audio	0.64	0.71	0.67	**6674**
	Browsing	**0.36**	**0.32**	**0.34**	6593
	Chat	**0.43**	**0.26**	**0.33**	**6809**
	Email	0.42	0.16	0.23	**6786**
	FTP	0.37	0.23	0.28	**6681**
	P2P	0.4	0.62	0.49	**6745**
	VOIP	0.66	0.97	0.78	**6798**
	Video	0.38	0.54	0.45	**6623**
XGBoost	Audio	0.91	0.89	0.9	**6674**
	Chat	0.8	0.81	0.8	**6809**
	FTP	0.75	0.66	0.7	**6681**
	Browsing	0.82	0.73	0.77	**6593**
	Email	0.78	0.7	0.74	**6786**
	Video	0.72	0.81	0.76	**6623**
	P2P	0.76	0.84	0.79	**6745**
	VOIP	0.92	0.99	0.95	**6798**
Logistic Regression	Audio	0.21	0.18	0.2	**6674**
	Browsing	0.3	0.0	0.01	**6593**
	Chat	0.21	0.1	0.14	**6809**
	Email	0.15	0.28	0.19	**6786**
	FTP	0.23	0.33	0.27	**6681**
	P2P	0.14	0.34	0.2	**6745**
	VOIP	0.35	0.0	0.0	**6798**

 doi:10.20944/preprints202507.1926.v1

	Video	0.29	0.26	0.27	**6623**
Ensemble	Audio	0.85	0.91	0.88	**6674**
	Chat	0.79	0.83	0.81	**6809**
	FTP	0.72	0.7	0.71	**6681**
	Browsing	0.83	0.72	0.77	**6593**
	Email	0.83	0.69	0.75	**6786**
	Video	0.75	0.8	0.77	**6623**
	P2P	0.78	0.84	0.81	**6745**
	VOIP	**0.92**	**0.99**	**0.95**	**6798**

After we conducted this comprehensive evaluation of various supervised learning algorithms, we assessed each model based on standard metrics such as accuracy, precision, recall, F1-score, and performance, and the Decision Trees consistently demonstrated inferior results. Consequently, we selected the Decision Tree algorithm as the best algorithm. More details about the results of the Decision Trees algorithm are provided below. The following **Error! Reference source not found.** shows the Confusion Matrix for Layer 2 (Darknet Protocol Classification)

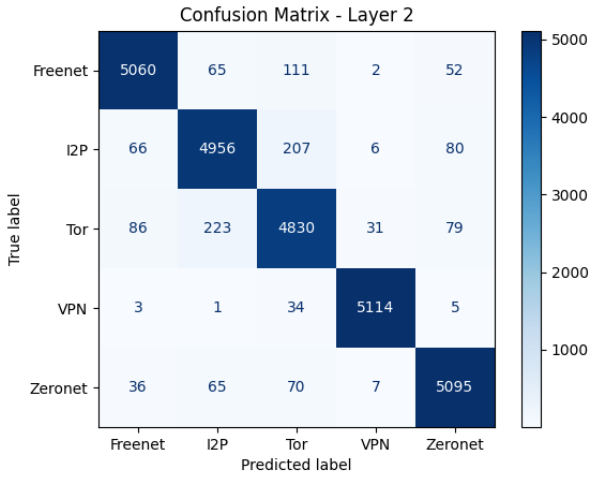


Figure 4. Layer 2 Confusion Matrix.

The following **Error! Reference source not found.** shows the Confusion Matrix for Layer 3 (Darknet Behavior Classification)

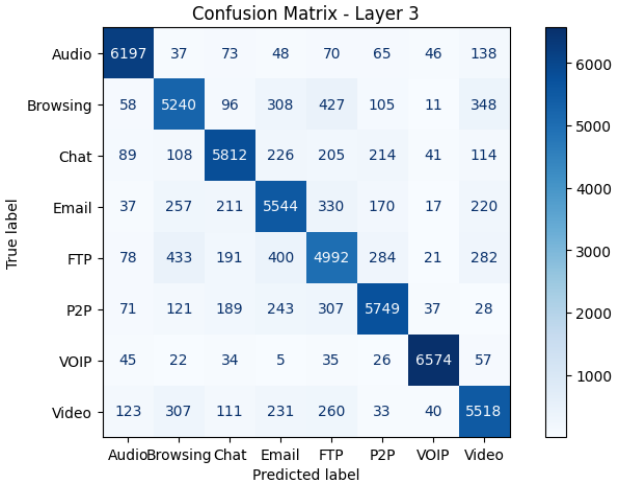


Figure 5. Layer 3 Confusion Matrix.

The following **Error! Reference source not found.** shows the ROC Curve for Layer 1 (Normal vs Darknet)

 doi:10.20944/preprints202507.1926.v1

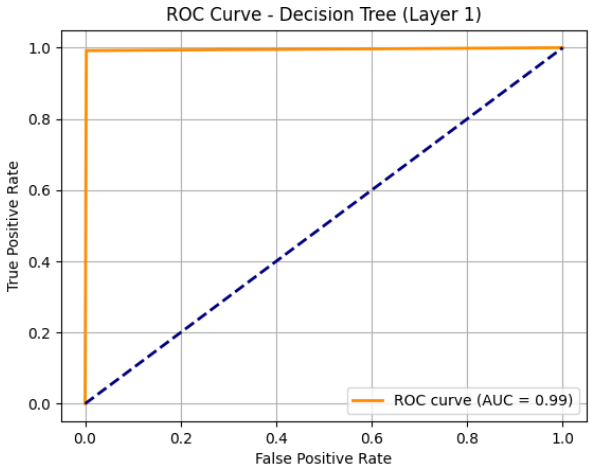


Figure 6. Layer 1 ROC Curve.

The following **Error! Reference source not found.** shows the ROC Curve for Layer 2 (Darknet Protocol Classification)

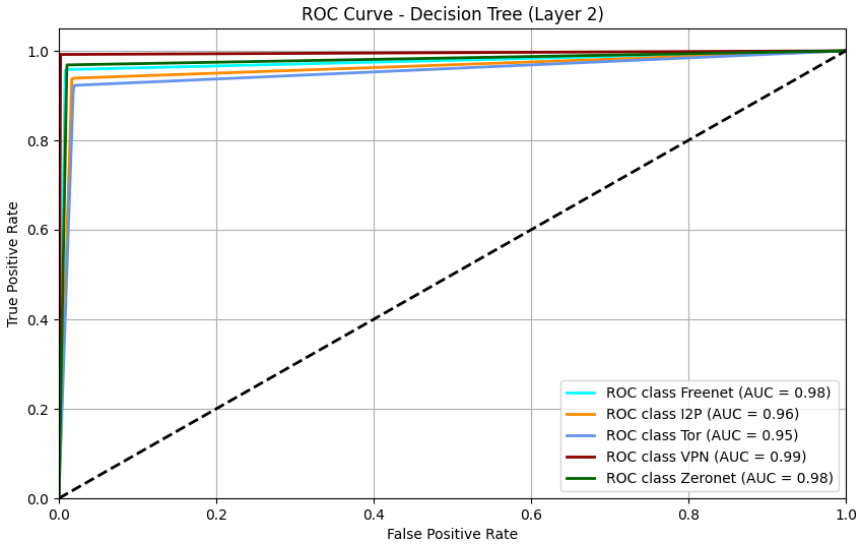
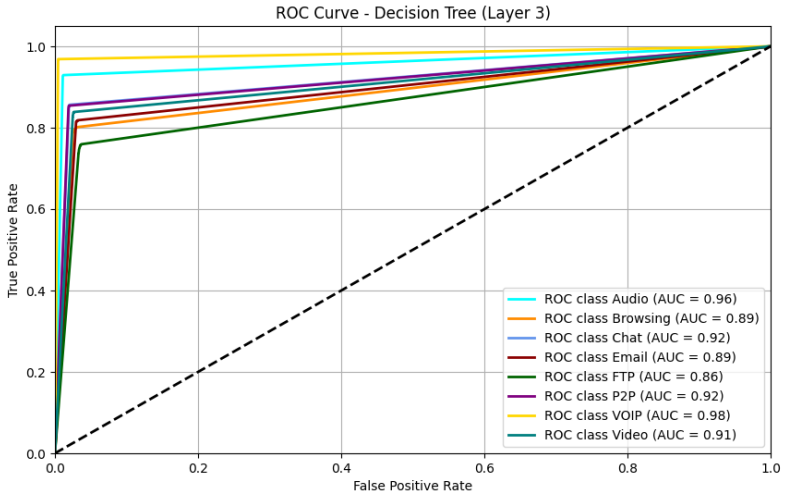


Figure 7. Layer 2 ROC Curve.

The following **Error! Reference source not found.** shows the ROC Curve for Layer 3 (Darknet Behavior Classification)



 doi:10.20944/preprints202507.1926.v1

Figure 8. Layer 3 ROC Curve.

4. Potential Applications

The SafeSurf Darknet 2025 dataset presents numerous opportunities for researchers and cybersecurity practitioners by providing rich, labeled traffic data across multiple darknet technologies and behavioral categories. Its detailed structure and behavior-focused labeling make it suitable for real-world and academic applications.

4.1. Intrusion Detection and Anomaly Detection

One of the primary applications of this dataset is the development and evaluation of **Intrusion Detection Systems (IDS)** and **Anomaly Detection models**. Due to its encrypted and obfuscated nature, traditional IDS solutions often struggle to detect or correctly interpret darknet traffic. By offering labeled samples of diverse darknet behaviors such as P2P sharing, encrypted chat, and decentralized video streaming, the dataset allows training models to detect anomalous patterns within network traffic, even when payload data is inaccessible.

4.2. Cyber Threat Intelligence

The dataset supports extracting behavioral signatures associated with various darknet technologies, enabling improved **threat profiling** and **network intelligence**. For example, by analyzing flow-level characteristics of Freenet or I2P traffic, analysts can better understand how these protocols operate in the wild. This aids in constructing threat indicators for Security Operations Centers (SOCs), law enforcement investigations, and national cybersecurity agencies aiming to identify or disrupt illicit darknet activities.

4.3. Behavioral Traffic Classification

Unlike many existing datasets that only distinguish between benign and malicious traffic, SafeSurf Darknet 2025 supports **multi-class behavioral classification**. Machine learning models can be trained to detect darknet traffic and classify the specific type of activity (e.g., video streaming vs. file transfer). This capability is crucial for building **context-aware security systems** that adapt responses based on the detected activity rather than relying solely on binary decisions.

4.4. Network Traffic Analysis Research

The dataset is also well-suited for **academic research in network traffic analysis**, particularly studies focusing on encrypted traffic, flow-based feature engineering, and darknet protocol identification. With a diverse mix of traffic types across Tor, Freenet, ZeroNet, I2P, and VPN environments, researchers can experiment with various analytical techniques, from feature selection to time-series modeling.

4.5. Curriculum and Educational Use

Finally, the dataset can be a practical resource for educational purposes in cybersecurity and data science courses. It provides students with hands-on experience in preprocessing, exploratory data analysis, and machine learning on real-world darknet traffic, preparing them for cyber defense and intelligence careers.

5. Ethical Considerations and Limitations

- Ethical concerns related to darknet data collection.
- Steps taken to ensure privacy and compliance with regulations.
- Limitations of the dataset (e.g., biases, data representativeness).

 doi:10.20944/preprints202507.1926.v1

5.1. Ethical Considerations

Creating and analyzing darknet traffic datasets involves unique ethical challenges due to the nature of the data and its potential associations with privacy-sensitive or illicit content. In the SafeSurf Darknet 2025 dataset, we took deliberate steps to ensure the ethical integrity of our data collection and handling procedures.

First, **no engagement with illegal darknet services** or marketplaces occurred during the data collection. All behaviors were simulated using publicly accessible darknet tools and services that support anonymity but are not inherently illegal, such as Tor, I2P, Freenet, and ZeroNet. The browsing and communication patterns captured were confined to generic activities (e.g., accessing blogs, downloading sample files, sending test messages), avoiding exposure to illicit content.

Second, **no real personal data** was included in the dataset. All traffic was generated in a controlled and consented environment using test accounts and machines. Any potentially sensitive identifiers (e.g., IP addresses, usernames) were anonymized or sanitized where applicable to preserve privacy.

Third, the dataset aligns with **ethical guidelines for cybersecurity research** and complies with local data protection standards. The traffic does not contain payload data; only flow-level metadata and statistical features were retained to reduce any risk of content-level privacy violations.

The dataset is intended strictly for **research and educational purposes**, and users are expected to comply with institutional review board (IRB) policies or national regulations applicable in their jurisdictions.

5.2. Limitations

While the SafeSurf Darknet 2025 dataset introduces several improvements over existing datasets, certain limitations must be acknowledged to contextualize its appropriate use and interpretation.

- **Synthetic Environment Bias**: The traffic was generated through real applications and intentional behaviors but was still conducted in a controlled and partially simulated environment. This may not fully reflect the complexity, noise, or unpredictability of actual darknet communications in the wild, potentially affecting the robustness of models trained exclusively on this dataset.
- **Representativeness and Overfitting Risk**: The dataset, while diverse, may not capture the full spectrum of darknet traffic variations due to differences in user behavior, time, geography, and network conditions. As a result, machine learning models trained on this dataset may risk **overfitting** to its specific traffic patterns, particularly when applied to unseen or real-world environments. Researchers are encouraged to use this dataset alongside other sources or perform cross-dataset validation when possible.
- **Incomplete Behavioral Coverage**: In some darknet technologies, we could not simulate or access certain behaviors, especially those in deeper functional layers (e.g., decentralized email, P2P file sharing, or video streaming in inactive or partially deprecated networks). This was due to service unavailability, inactive peer networks, or technical restrictions in protocols like Freenet and I2P during the data collection.
- **Protocol and Platform Scope**: The dataset includes a targeted set of darknet platforms (Tor, Freenet, ZeroNet, I2P, VPN), but omits other technologies such as GNUnet, LokiNet, or RetroShare. This may limit its comprehensiveness in representing the entire darknet ecosystem.
- **Temporal Constraints**: All traffic was collected over a specific period, which may exclude evolving usage patterns or emerging darknet platforms. Periodic updates would be needed to maintain relevance and adaptability to current threats.

 doi:10.20944/preprints202507.1926.v1

- **Encrypted Traffic Complexity**: As with many flow-based datasets, payload content is unavailable. While this enhances privacy and ethics compliance, it also restricts the depth of semantic analysis possible in applications like content-based filtering or deep behavioral profiling.

Despite these limitations, the dataset remains a valuable and much-needed resource for advancing darknet detection, classification, and threat intelligence research. Future iterations aim to address these gaps through broader behavioral coverage, longer data collection windows, and integration with additional darknet technologies.

6. Conclusions and Future Work

This paper introduced the **SafeSurf Darknet 2025** dataset — a novel and behavior-focused darknet traffic analysis and classification resource. Unlike many datasets primarily focusing on binary detection tasks, our dataset is structured around various behavioral classes across multiple darknet technologies, including **Tor, I2P, Freenet, ZeroNet, and VPN**. Traffic was captured in realistic environments using physical and virtual machines, and labeling was performed based on direct observation and controlled behavior invocation, ensuring strong ground-truth accuracy.

The dataset offers a rich foundation for various applications, such as **anomaly detection**, **traffic classification**, and **cyber threat intelligence**. Its structured format and feature-rich design support traditional analysis and modern machine learning research. It is also intended for public release, making it a valuable resource for academic and professional cybersecurity communities.

Looking ahead, several directions can be pursued to enhance the scope and utility of the dataset:

- **Expansion of Behavioral Coverage**: Additional behaviors that could not be captured due to technical or network limitations—such as decentralized email or video streaming on inactive darknet services—can be included in future iterations.
- **Inclusion of More Protocols**: Future versions may incorporate other darknet technologies such as **GNUnet**, **LokiNet**, or **Retroshare** to broaden the representational coverage of darknet traffic.
- **Long-Term Data Collection**: Extending the capture period will enable the observation of temporal changes in darknet traffic, allowing for the study of long-term trends and protocol evolution.
- **Machine Learning Benchmarking**: While initial tests showed promising results, future work may include formal benchmarking of classification models across multiple behaviors, feature sets, and darknet technologies. This could also include comparisons with other datasets to evaluate generalization.
- **Cross-Dataset Validation**: To reduce overfitting risks and improve robustness, future studies may combine this dataset with others in joint evaluations.
- **Payload-Inclusive Variant**
  (If ethically and legally viable) Consider a secure variant with limited payload samples for advanced analysis or encrypted traffic modeling.
- **Public Release and Community Support**: We aim to release the dataset with full documentation, usage guidelines, and version control to encourage reproducibility, community feedback, and collaborative improvements.

In summary, the SafeSurf Darknet 2025 dataset addresses key limitations in existing darknet traffic datasets and sets the stage for more nuanced and effective research in cybersecurity. Future developments will continue to refine its accuracy, scope, and relevance in detecting and understanding darknet activities.

 doi:10.20944/preprints202507.1926.v1

Data Availability Statement: The dataset generated in this research is published by Mendeley Data/Elsevier, a public repository, and can be accessed through: **Dataset Link**: https://data.mendeley.com/drafts/kcrnj6z4rm. **Dataset DOI:** 10.17632/kcrnj6z4rm.2.

References

1. M. Al-Fayoumi, A. Elayyan, A. Odeh and Q. A. Al-Haija, "Tor network traffic classification using machine learning based on time-related feature," 6th Smart Cities Symposium (SCS 2022), Hybrid Conference, Bahrain, 2022, pp. 92-97, doi: 10.1049/icp. 2023.0354.
2. Abu Al-Haija, Q.; Krichen, M.; Abu Elhaija, W. Machine-Learning-Based Darknet Traffic Detection System for IoT Applications. Electronics 2022, 11, 556. https://doi.org/10.3390/electronics11040556
3. Y. Hu, F. Zou, L. Li and P. Yi, "Traffic Classification of User Behaviors in Tor, I2P, ZeroNet, Freenet," 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, 2020, pp. 418-424, doi: 10.1109/TrustCom50675.2020.00064.
4. M. Alimoradi, M. Zabihimayvan, A. Daliri, R. Sledzik, and R. Sadeghi, "Deep Neural Classification of Darknet," Frontiers in Artificial Intelligence and Applications, Volume 356: Artificial Intelligence Research and Development.
5. M. Rawashdeh, Q. A. Al-Haija and M. Qasaimeh, "Analysis of TOR Artifacts and Traffic in Windows 11: A Virtual Lab Approach and Dataset Creation," 2023 14th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2023, pp. 1-6, doi: 10.1109/ICICS60529.2023.10330539.
6. H. Yin and Y. He, "I2P Anonymous Traffic Detection and Identification," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 157-162, doi: 10.1109/ICACCS.2019.8728517.
7. Wang, S., Gao, Y., Shi, J., Wang, X., Zhao, C. and Yin, Z., 2020. Look deep into the new deep network: A measurement study on the ZeroNet. In Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part I 20 (pp. 595-608). Springer International Publishing.
8. Q. A. Al-Haija and R. Ibrahim, "Introduction to Dark Web," in Perspectives on Ethical Hacking and Penetration Testing, IGI Global, 2023, p. 445.