CTU-13 DATASET

MTH-410

Beyza Akgün
Ahmet Yiğit Özkoca
Yusuf Eskiocak

CTU 13 Dataset -42

Origin:

The CTU-13 dataset was created by researchers at the Czech Technical University (CTU) in Prague, Czech Republic.

It is part of the CTU Prague Capture Traffic Archive, which contains network traffic data captured in various network environments and scenarios.

Intended Purpose:

The primary purpose of the CTU-13 dataset is to serve as a benchmark dataset for evaluating and comparing cybersecurity algorithms and systems.

It provides a diverse set of network traffic data collected from real-world network environments, including both benign and malicious activities.

The dataset includes a variety of network traffic features such as source/destination IP addresses, ports, protocols, packet sizes, timestamps, etc.

Researchers can use the CTU-13 dataset to develop, test, and validate cybersecurity solutions, including intrusion detection systems (IDS), malware detection algorithms, network anomaly detection techniques, and more.

Significance in the Realm of Cybersecurity:

The CTU-13 dataset plays a significant role in advancing research and development efforts in the field of cybersecurity.

It provides a standardized and realistic testbed for evaluating the effectiveness of cybersecurity tools and techniques under various network conditions and attack scenarios.

By using the CTU-13 dataset, researchers can benchmark their algorithms against a common set of data, facilitating fair comparisons and reproducibility of results.

The dataset helps address the growing need for robust cybersecurity solutions in the face of evolving cyber threats and attacks.

Insights gained from analyzing the CTU-13 dataset contribute to improving the understanding of cyber threats and vulnerabilities, leading to better defense strategies and countermeasures.

In summary, the CTU-13 dataset is a valuable resource for the cybersecurity community, providing researchers with access to real-world network traffic data for evaluating and advancing cybersecurity algorithms and systems. Its standardized format and diverse content make it an essential tool for developing effective cybersecurity solutions to protect against emerging cyber threats.

Enumeration and Description of Features

Enumeration of Features:

The CTU 13 dataset, encapsulating network traffic data, serves as a vital resource for analyzing cybersecurity threats and anomalies. Enumerating its features lays the groundwork for understanding the dataset's structure, facilitating subsequent analysis and modeling tasks.

The Python code snippet below demonstrates how to enumerate the features of the CTU 13 dataset using the Pandas library:

```
import pandas as pd
# File path to the CTU 13 dataset
binetflow_file = 'C:\\Users\\Asus\\Desktop\\capture20110810.binetflow.xz'
# Read the dataset into a Pandas DataFrame
df = pd.read_csv(binetflow_file)
# Enumerate features
for feature in df.columns:
    print("Feature:", feature)
    print("Data Type:", df[feature].dtype)
    print("Sample Values:", df[feature].unique())
```

In this code snippet:

The pd.read_csv() function is used to read the dataset into a Pandas DataFrame from the specified file path.

A loop iterates over each column (feature) in the DataFrame.

For each feature, its name, data type, and unique sample values are printed.

This enumeration provides a comprehensive overview of the dataset's structure, including the types of features available and the range of values they encompass.

And as a output we have these:

1. **StartTime**
    - Feature: StartTime
    - Data Type: object
    - Sample Values: ['2011/08/10 09:46:59.607825' '2011/08/10 09:47:00.634364' ...]
2. **Dur**
    - Feature: Dur
    - Data Type: float64
    - Sample Values: [1.026539 1.009595 ...]
3. **Proto**
    - Feature: Proto
    - Data Type: object
    - Sample Values: ['tcp' 'udp' ...]

4. **SrcAddr**
    - Feature: SrcAddr
    - Data Type: object
    - Sample Values: ['94.44.127.113' '147.32.86.89' ...]
5. **Sport**
    - Feature: Sport
    - Data Type: object
    - Sample Values: ['1577' '4768' ...]
6. **Dir**
    - Feature: Dir
    - Data Type: object
    - Sample Values: [' ->' ' ?>' ...]
7. **DstAddr**
    - Feature: DstAddr
    - Data Type: object
    - Sample Values: ['147.32.84.59' '77.75.73.33' ...]
8. **Dport**
    - Feature: Dport
    - Data Type: object
    - Sample Values: ['6881' '80' ...]
9. **State**
    - Feature: State
    - Data Type: object
    - Sample Values: ['S_RA' 'SR_A' ...]
10. **sTos**
    - Feature: sTos
    - Data Type: float64
    - Sample Values: [0. nan ...]
11. **dTos**
    - Feature: dTos
    - Data Type: float64
    - Sample Values: [0. nan ...]
12. **TotPkts**
    - Feature: TotPkts
    - Data Type: int64
    - Sample Values: [4 3 ...]
13. **TotBytes**
    - Feature: TotBytes
    - Data Type: int64
    - Sample Values: [276 182 ...]
14. **SrcBytes**
    - Feature: SrcBytes
    - Data Type: int64
    - Sample Values: [156 122 ...]
15. **Label**
    - Feature: Label
    - Data Type: object

- Sample Values: ['flow=Background-Established-cmpgw-CVUT' ...]

Description of Features:

The CTU 13 dataset encompasses a wide array of features, each providing valuable insights into network traffic behavior and aiding in the detection of cybersecurity threats. Below is a detailed description of the key features present in the dataset:

StartTime:

Represents the timestamp indicating when the network activity associated with a particular record began.

It plays a crucial role in analyzing network traffic patterns over time, identifying trends, and correlating events.

Dur:

Stands for duration and represents the duration of the network activity in seconds.

Provides information about how long a particular network interaction lasted, which is valuable for understanding the nature and extent of the activity.

Proto:

Stands for protocol and indicates the network protocol used for the communication, such as TCP, UDP, ICMP, etc.

Understanding the protocol used is essential for analyzing network behavior, as different protocols have different characteristics and behaviors.

SrcAddr:

Represents the source IP address of the network communication.

Identifies the origin of the network activity and is crucial for tracing the source of potential security threats or anomalies.

Sport:

Stands for source port and represents the port number used by the source device for the communication.

Port numbers help differentiate between different network services or applications running on a device.

Dir:

Stands for direction and indicates the direction of the network traffic flow, such as inbound or outbound.

Understanding the direction of traffic flow is essential for analyzing network security incidents and identifying potential threats.

DstAddr:

Represents the destination IP address of the network communication.

Identifies the target of the network activity and is crucial for understanding the scope and impact of the activity.

Dport:

Stands for destination port and represents the port number used by the destination device for the communication.

Similar to the source port, destination port numbers help differentiate between different network services or applications.

State:

Indicates the state of the network connection, such as established, closed, or other relevant states.

Provides information about the current status of the network connection, which is useful for detecting abnormal behavior or unauthorized access.

sTos and dTos:

Represent the Type of Service (TOS) values for the source and destination, respectively.

TOS values are used to prioritize network traffic based on characteristics such as delay, throughput, reliability, etc.

TotPkts

Stands for total packets and represents the total number of packets exchanged during the network activity.

Packet count is an essential metric for analyzing network traffic volume and intensity.

TotBytes:

Represents the total number of bytes exchanged during the network activity.

Provides information about the data transfer volume associated with the communication.

SrcBytes:

Stands for source bytes and represents the number of bytes sent by the source device during the network activity.

Analyzing the source bytes can help identify potential data exfiltration or unauthorized data transfer activities.

Label:

Indicates the class or category assigned to each network activity, such as normal, malicious, suspicious, etc.

Serves as the target variable for supervised machine learning algorithms, allowing the detection of network anomalies or security threats based on historical patterns.

Exploration of First Five Rows of Each Column:

To gain a more detailed understanding of the data within each column of the CTU 13 dataset, we examined the first five rows of every column. This exploration provides insight into the range of values, format, and characteristics present in each feature. Below is the code used to extract and display the first five rows of each column:

```
# Extract and display the first five rows of each column
for column in df.columns:
    first_five_rows = df[column].head()
    print("First Five Rows of Column {}: \n{}".format(column,
first_five_rows))
```

First Five Rows of Each Column:

Start Time (StartTime):

```
First Five Rows of Column StartTime:
0    2011/08/10 09:46:59.607825
1    2011/08/10 09:47:00.634364
2    2011/08/10 09:47:48.185538
3    2011/08/10 09:47:48.230897
4    2011/08/10 09:47:48.963351
Name: StartTime, dtype: object
```

Duration (Dur):

```
First Five Rows of Column Dur:
0    1.026539
1    1.009595
2    3.056586
3    3.111769
4    3.083411
Name: Dur, dtype: float6
```

Protocol (Proto):

```
First Five Rows of Column Proto:
0    tcp
1    tcp
2    tcp
3    tcp
4    tcp
Name: Proto, dtype: object
```

Source Address (SrcAddr):

```
First Five Rows of Column SrcAddr:
0    94.44.127.113
1    94.44.127.113
2     147.32.86.89
3     147.32.86.89
4     147.32.86.89
Name: SrcAddr, dtype: object
```

Source Port (Sport):

```
First Five Rows of Column Sport:
0    1577
1    1577
2    4768
3    4788
4    4850
Name: Sport, dtype: object
```

Direction (Dir):

```
First Five Rows of Column Dir:
0      ->
1       ->
2       ->
3       ->
4       ->
Name: Dir, dtype: object
```

Destination Address (DstAddr):

```
First Five Rows of Column DstrAddr:

0    147.32.84.59
1    147.32.84.59
2     77.75.73.33
3     77.75.73.33
4     77.75.73.33
Name: DstAddr, dtype: object
```

Destination Port (Dport):

```
First Five Rows of Column Dport:
0    6881
1    6881
2      80
3      80
4      80
Name: Dport, dtype: object
```

State (State):

```
First Five Rows of Column State:
0    S_RA
1    S_RA
2    SR_A
3    SR_A
4    SR_A
Name: State, dtype: object
```

Source Type of Service (sTos):

```
First Five Rows of Column sTos:
0    0.0
1    0.0
2    0.0
3    0.0
4    0.0
Name: sTos, dtype: float64
```

Destination Type of Service (dTos):

```
First Five Rows of Column dTos:
0    0.0
1    0.0
2    0.0
3    0.0
4    0.0
Name: dTos, dtype: float64
```

Total Packets (TotPkts):

```
First Five Rows of Column TotPkts:
0    4
1    4
2    3
3    3
4    3
Name: TotPkts, dtype: int64
```

Total Bytes (TotBytes):

```
First Five Rows of Column TotBytes:

0    276
1    276
2    182
3    182
4    182
Name: TotBytes, dtype: int64
```

Source Bytes (SrcBytes):

```
First Five Rows of Column SrcBytes:
0    156
1    156
2    122
3    122
4    122
Name: SrcBytes, dtype: int64
```

Label:

```
First Five Rows of Column Label:
0   flow=Background-Established-cmpgw-CVUT
1    flow=Background-Established-cmpgw-CVUT
2               flow=Background-TCP-Attempt
3               flow=Background-TCP-Attempt
4               flow=Background-TCP-Attempt
Name: Label, dtype: object
```

This presentation organizes the first five rows of each column in the dataset, providing a structured overview of the data's content and format.

Exploring CTU-13 Dataset

In this section, we delve into the CTU-13 dataset to gain insights into network traffic patterns and characteristics. The CTU-13 dataset comprises network traffic data captured during a malware analysis project conducted by the Czech Technical University (CTU). Through visualizations and analyses, we aim to better understand the behavior of network traffic, distinguish between normal and malicious activities, and identify potential anomalies. The following code snippets and accompanying visualizations demonstrate our exploration of the dataset, shedding light on various aspects of network traffic behavior and aiding in the detection of potential security threats.

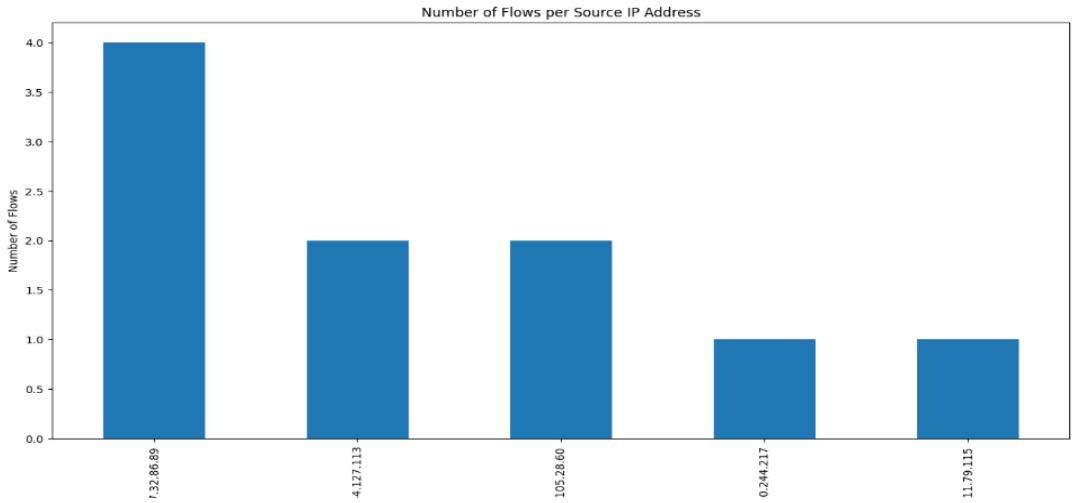

Figure 1

Figure 1: Number of Flows per Source IP Address

This bar graph illustrates the distribution of network traffic flows originating from different source IP addresses within a subset of the CTU-13 dataset. The X-axis lists the source IP addresses, while the Y-axis indicates the number of flows generated by each IP address. The bars represent the count of flows, revealing the activity level of each source IP within the captured network data. It is a visual representation of network activity from the first 10 rows of the dataset.

```
# Plotting network traffic patterns
df = pd.read_csv(binetflow_file, nrows=10)
plt.figure(figsize=(10, 6))
df['SrcAddr'].value_counts().plot(kind='bar', figsize=(10, 6))
plt.xlabel('Source IP Address')
plt.ylabel('Number of Flows')
plt.title('Nüme of Flows per Source IP Address')
plt.show()
```

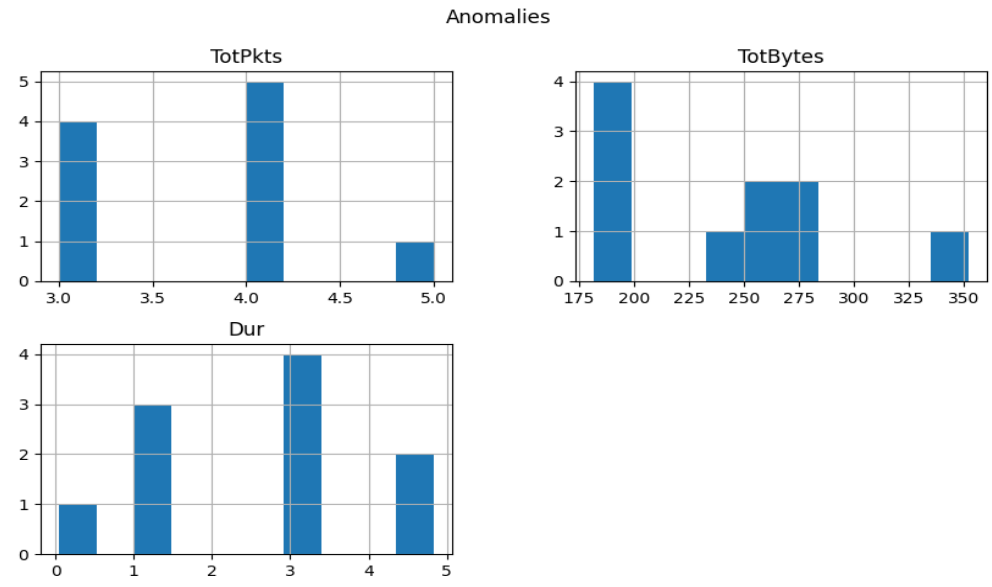

Figure 2

Figure 2: Histograms of Selected Features

These histograms illustrate the distribution of three selected features—Total Packets (TotPkts), Total Bytes (TotBytes), and Duration (Dur)—across the network traffic captured in the CTU-13 dataset. Each subplot provides insights into the frequency distribution of a particular feature, showcasing the variance and commonality in packet counts, byte sizes, and session durations in the network traffic. This aids in understanding the typical traffic patterns and identifying outliers or anomalies.

```
# Plot histograms of selected features
selected_features = ['TotPkts', 'TotBytes', 'Dur']
df[selected_features].hist(figsize=(10, 6))
plt.suptitle('Anomalies')
plt.show()
```

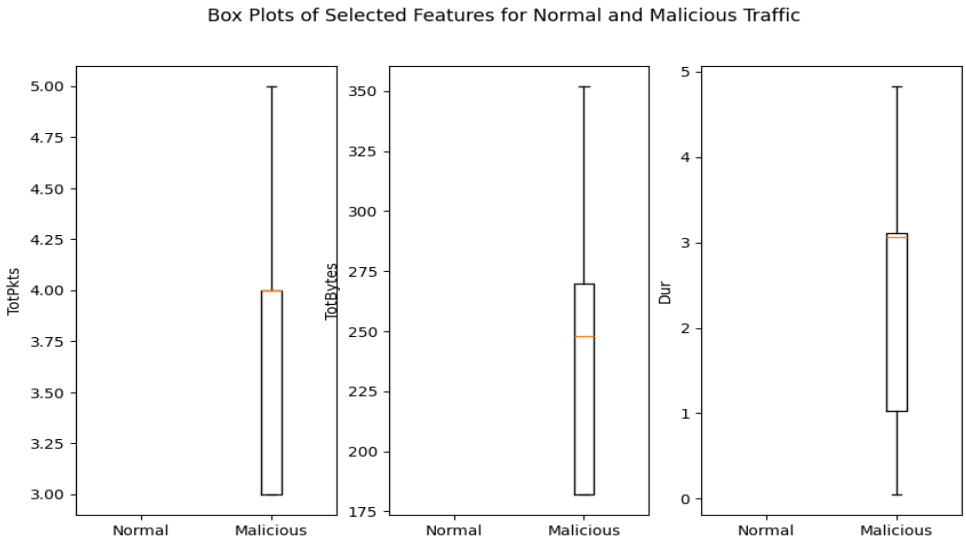

Figure 3

Figure 3: Box Plots of Selected Features for Normal and Malicious Traffic

The box plots compare the distributions of Total Packets (TotPkts), Total Bytes (TotBytes), and Duration (Dur) between normal and malicious traffic in the CTU-13 dataset. For each feature, two box plots are drawn side by side—one for normal traffic and the other for malicious traffic. These plots include the median, quartiles, and potential outliers, providing a clear visual comparison between the typical behavior of normal and malicious network flows.

```
# Filter the dataset for normal and malicious traffic
normal_traffic = df[df['Label'] == 'Normal']
malicious_traffic = df[df['Label'] != 'Normal']
# Plot box plots for selected features
selected_features = ['TotPkts', 'TotBytes', 'Dur']
plt.figure(figsize=(10, 6))
for i, feature in enumerate(selected_features):
    plt.subplot(1, len(selected_features), i+1)
    plt.boxplot([normal_traffic[feature], malicious_traffic[feature]])
    plt.xticks([1, 2], ['Normal', 'Malicious'])
    plt.ylabel(feature)
```

```
plt.suptitle('Box Plots of Selected Features for Normal and Malicious
Traffic')
plt.show()
```

Literature Review: Machine Learning Applications Utilizing the CTU-13 Dataset for Cybersecurity Analysis

In the realm of cybersecurity, traditional methods often struggle to detect emerging threats and novel attack patterns effectively. To address this challenge, researchers have turned to machine learning (ML) as a promising solution due to its adaptability and ability to learn from data. One prominent dataset used in ML applications for cybersecurity analysis is the CTU-13 dataset, which contains diverse network traffic captures, including various types of botnet activities.

Notable Findings:

Significance of Feature Selection Techniques:

ML models heavily rely on feature selection techniques to enhance efficiency and accuracy in cybersecurity analysis. Methods such as Pearson correlation filtering, wrapper methods like backward feature elimination, and embedded methods within classifiers like Random Forest have been utilized to reduce the dimensionality of input data and improve model performance.

Supervised Learning with Logistic Regression:

Logistic regression emerges as a notable supervised learning model for binary classification tasks in cybersecurity, particularly in distinguishing between normal and malicious network traffic. Its use of sigmoid functions to model outcome probabilities makes it suitable for identifying malicious activities. Extensive experiments have demonstrated the importance of parameter tuning in optimizing logistic regression models for achieving high precision, recall, and F1-score values.

Application of Deep Learning Models:

Deep learning models, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and hybrid CNN-LSTM architectures, have shown promising results in bot detection and malware analysis. These models exhibit high precision and recall rates, particularly in scenarios with limited malicious traffic, thus contributing to robust cybersecurity defenses.

Adversarial Robustness Training (ART):

Research emphasizes the importance of adversarial robustness training for enhancing cybersecurity defenses against sophisticated attacks. Leveraging the CTU-13 dataset, studies have explored algorithm selection, evaluation metrics, feature engineering, and training methodologies to improve model resilience against adversarial attacks, thereby bolstering cybersecurity analysis.

Exemplary Methodologies:

Ensemble Learning for Anomaly Detection:

Integration of multiple ML models using ensemble learning techniques, incorporating both supervised and unsupervised learning approaches, enhances the accuracy and robustness of anomaly detection systems. Ensemble methods capitalize on the diversity of individual models to achieve superior performance in cybersecurity analysis.

Adversarially Trained Models:

Development of adversarially trained models capable of detecting and mitigating adversarial attacks in real-time is crucial for bolstering cybersecurity defenses. Adversarial training involves exposing models to adversarial examples during training to improve their resilience against sophisticated attacks, thereby ensuring robust cybersecurity analysis.

Continuous Model Evaluation and Refinement:

Continuous evaluation and refinement of ML models are essential to adapt to evolving threats and maintain effective cybersecurity defenses. Regular updates and adjustments based on new data and emerging attack patterns enable cybersecurity systems to stay ahead of adversaries and effectively mitigate potential risks.

In conclusion, the literature review highlights the significance of machine learning applications utilizing the CTU-13 dataset for bolstering cybersecurity analysis. Notable findings underscore the importance of feature selection techniques, supervised learning with logistic regression, application of deep learning models, and the implementation of adversarial robustness training. Exemplary methodologies focus on ensemble learning for anomaly detection, adversarially trained models for adversarial resilience, and continuous model evaluation and refinement for adaptive cybersecurity defenses. These approaches collectively contribute to the development of effective cybersecurity solutions capable of combating emerging threats in network security.

Exploring Future Frontiers in Cybersecurity Research

As the cybersecurity landscape continues to evolve, so too must our approaches to defending against emerging threats. While significant progress has been made in leveraging machine learning (ML) techniques, particularly with the utilization of the CTU-13 dataset, there remain numerous untapped opportunities and challenges on the horizon. In this final section, we delve into the uncharted territories of cybersecurity research, aiming to identify potential avenues for future endeavors and untapped applications. By examining emerging trends, technological advancements, and evolving threat landscapes, we seek to illuminate pathways for innovation that promise to shape the future of cybersecurity. From enhancing the robustness of ML models to exploring novel use cases and addressing pressing cybersecurity challenges, this exploration serves as a call to action for researchers and practitioners to push the boundaries of what is possible in safeguarding our digital world.

Adversarial Robustness in ML Models:

Further research can focus on enhancing the adversarial robustness of ML models trained on the CTU-13 dataset. Investigate advanced adversarial training techniques and explore how robust models can be developed to withstand sophisticated attacks, thereby improving cybersecurity defenses.

Explainable AI for Interpretability:

Explore the integration of explainable AI techniques with ML models trained on the CTU-13 dataset to enhance interpretability. Investigate methods for providing human-understandable explanations for

model decisions, enabling cybersecurity analysts to gain insights into the underlying rationale behind detected threats.

Scalability and Deployment Challenges:

Address scalability and deployment challenges associated with deploying ML models trained on the CTU-13 dataset in real-world cybersecurity environments. Investigate strategies for optimizing model performance and resource utilization to ensure efficient deployment on large-scale networks.

Dynamic Threat Detection:

Develop dynamic threat detection mechanisms using ML models trained on the CTU-13 dataset to adapt to evolving cybersecurity threats in real-time. Explore techniques for continuous monitoring and analysis of network traffic data to identify new attack patterns and mitigate emerging threats effectively.

Privacy-Preserving Techniques:

Investigate privacy-preserving techniques for ML models trained on the CTU-13 dataset to protect sensitive information while maintaining detection accuracy. Explore methods such as federated learning, differential privacy, and homomorphic encryption to ensure data privacy and security in cybersecurity applications.

Integration with IoT Security:

Explore the application of ML models trained on the CTU-13 dataset for enhancing security in Internet of Things (IoT) environments. Investigate how ML-based anomaly detection techniques can be applied to detect and mitigate threats in IoT networks, addressing the unique challenges posed by IoT devices and communication protocols.

Threat Intelligence and Information Sharing:

Investigate the use of ML models trained on the CTU-13 dataset for threat intelligence and information sharing among cybersecurity professionals and organizations. Develop collaborative platforms and frameworks for sharing threat data and insights derived from ML-based analysis, enabling proactive defense against cyber threats.

Human-Machine Collaboration:

Explore the potential for human-machine collaboration in cybersecurity analysis using ML models trained on the CTU-13 dataset. Investigate how ML-based tools can augment the capabilities of cybersecurity analysts, facilitating faster threat detection, response, and decision-making.

By pursuing these avenues for future research endeavors and untapped applications, researchers can further leverage the CTU-13 dataset to advance the state-of-the-art in cybersecurity and develop innovative solutions for addressing evolving cyber threats.

In conclusion, our examination of the CTU-13 dataset and its integration with machine learning methodologies has shed light on the critical role of data-driven approaches in bolstering cybersecurity defenses. Through a comprehensive analysis of the dataset's features and a thorough review of existing literature, we have elucidated the significance of leveraging ML techniques for detecting and mitigating cyber threats.

The insights gleaned from our exploration underscore the importance of continuous learning, adaptation, and collaboration in the field of cybersecurity. While the CTU-13 dataset serves as a valuable foundation for current research endeavors, it also presents boundless opportunities for future innovation and discovery.

References

https://stumejournals.com/journals/confsec/2019/3/109.full.pdf

https://arxiv.org/ftp/arxiv/papers/2109/2109.07593.pdf

https://www.stratosphereips.org/datasets-ctu13