Scalable Anomaly Detection in Cybersecurity: A Data Mining Approach Using the CTU-13 Dataset

Halime Sıla ÖZYURT
Faculty of Engineering and Natural Sciences, Biruni University
Computer Engineering
İstanbul, Türkiye
210408006@st.biruni.edu.tr

Damla TEZAL
Faculty of Engineering and Natural Sciences, Biruni University
Computer Engineering
İstanbul, Türkiye
220408031@st.biruni.edu.tr

Buse Emine SÖNMEZ
Faculty of Engineering and Natural Sciences, Biruni University
Computer Engineering
İstanbul, Türkiye
210408026@st.biruni.edu.tr

***Abstract*—The advent of rapid digital transformation, coupled with an increasing dependence on interconnected devices, has underscored the necessity of robust cybersecurity solutions to address the ever-evolving threats posed. This research project utilizes the CTU-13 dataset, a benchmark collection designed to reflect real-world scenarios involving botnets, to investigate sophisticated methodologies for botnet and anomaly detection. By combining approaches from data mining and machine learning, this study employs both supervised and unsupervised learning algorithms to accurately identify botnet traffic. The integration of graph mining techniques for the clustering of communication patterns, along with deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), significantly augments the detection of nefarious activities within intricate network settings. Experimental evaluations have validated the efficacy of this hybrid approach, achieving an F1 score of 0.985 and a detection accuracy of 99.99%. Through this study, the pivotal importance of the CTU-13 dataset in the development of adaptive and scalable mechanisms for detecting attacks is highlighted. Future endeavors will concentrate on enhancing computational efficiency and broadening the applicability of the framework to a variety of real-world datasets.**

Keywords*—CTU-13 dataset, botnet detection, anomaly detection, data mining, deep learning, graph mining, cybersecurity, machine learning.

INTRODUCTION

The advent of digital transformation and interconnected devices has transformed the way we interact with technology. However, this technological advancement has led to an equally important challenge: cybersecurity. As critical infrastructures, businesses and personal devices increasingly rely on interconnected networks, the risk of cyberattacks, data breaches and system vulnerabilities has increased exponentially. Cybersecurity researchers and professionals are constantly looking for innovative approaches to combat these challenges, with data mining and machine learning emerging as undeniable tools in this effort. The development of effective cybersecurity solutions is highly dependent on datasets that accurately represent real-world attack scenarios and network behavior. These datasets are essential for training and testing intrusion detection systems (IDS), anomaly detection algorithms and other security mechanisms.

By providing the necessary context and variety to model complex attack vectors and normal traffic patterns, they enable researchers to detect, explain and mitigate evolving cyber threats.

The CTU-13 dataset, which was collected at the CTU University in the Czech Republic in 2011, contains 13 scenarios of botnet traffic mixed with other types of traffic and background activity. Of many existing cybersecurity datasets, the CTU-13 dataset holds a significant position due to its focus on botnet activity in realistic network environments [4].

In particular, it provides a benchmark for evaluating the performance of attack detection systems in identifying botnet traffic [1].

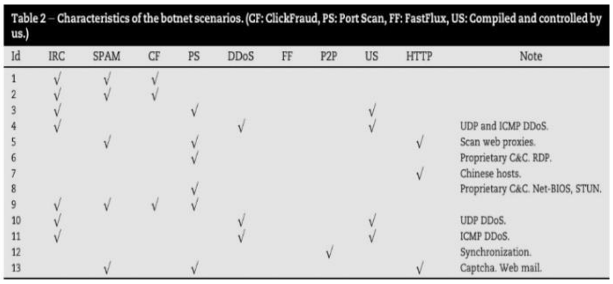


Table 2 – Characteristics of the botnet scenarios. (CF: ClickFraud, PS: Port Scan, FF: FastFlux, US: Compiled and controlled by us.)

Id	IRC	SPAM	CF	PS	DDoS	FF	P2P	US	HTTP	Note
1	√	√	√							
2	√	√	√							
3	√			√				√		
4	√				√			√		UDP and ICMP DDoS.
5		√		√					√	Scan web proxies.
6				√						Proprietary C&C. RDP.
7									√	Chinese hosts.
8				√						Proprietary C&C. Net-BIOS, STUN.
9	√	√	√	√						
10	√				√			√		UDP DDoS.
11	√				√			√		ICMP DDoS.
12							√			Synchronization.
13		√		√					√	Captcha. Web mail.

Table 1.1: Characteristics of botnet scenarios [1]

The importance of CTU-13 lies not only in its comprehensive coverage of botnet behavior, but also in filling the gap between theoretical research and practical application. By providing a detailed representation of network traffic, it allows researchers to explore the application of machine learning and data mining techniques to detect advanced cyberattacks. These techniques, ranging from supervised classification to unsupervised anomaly detection, have shown great promise in improving the accuracy and efficiency of attack detection systems. [2,3].

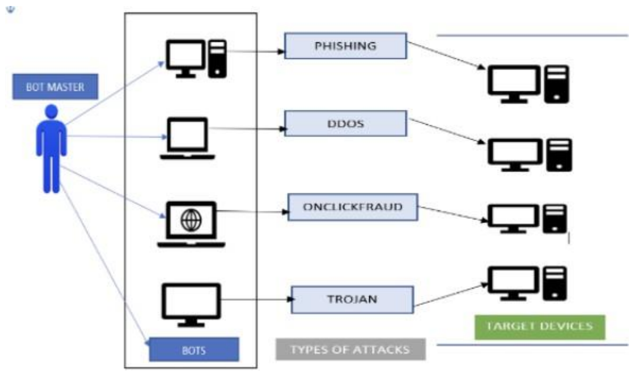


Table 1.2: Botnet Architecture [4]

This report focuses on familiarizing the CTU-13 dataset with its structure and key features and exploring its potential applications in cybersecurity. It describes how the dataset contributes to developing highly adaptive attack detection mechanisms, emphasizing its interest in data mining and machine learning. By exploring the nuances of the dataset, this study aims to communicate its importance in advancing cybersecurity research and reducing the growing cyberattack challenge.

LITERATURE REVIEW

The increase in sophisticated cyber threats, such as botnets and network anomalies, has made traditional detection methods increasingly deficit. Traditional methods like rule-based systems, signature-based detection, and machine learning models, including Random Forest and Support Vector Machines (SVM), struggle to approach the complexities of modern cyberattacks, specifically zero-day threats and localized botnets. These methods, while foundational, deal with obstacles in scalability, adaptability, and false-positive rates, submitting them less effective in dynamic and large-scale network environments. As a consequence, researchers have turned to advanced methodologies, such as graph mining and deep learning, to defeat these limitations. These technologies influence modern computational techniques to enhance the detection of complex and evolving threats, offering a more powerful and efficient alternative for cybersecurity systems.

On the flip side, deep learning techniques have revolutionized the way we detect anomalies in cybersecurity. These approaches, which include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), excel at dealing with unstructured information like system logs (syslogs) and network traffic (dataflows). The LogHub dataset has played a crucial role in this area, especially for analyzing system log anomalies. Within this dataset, CNNs have achieved an F1 score of 0.999 in cases where they had labeled data, and semi-supervised approaches scored 0.938. Moreover, models based on RNNs, such as DeepLog and LogAnomaly, have proven their effectiveness in analyzing sequential data for anomaly detection by taking advantage of the order of events [6].

In terms of handling data movement across networks, advanced machine learning approaches like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have consistently outperformed conventional methods. By utilizing datasets like NIDS and CTU-13, these algorithms have demonstrated exceptional performance indicators, including an impressive 0.985 F1 score for identifying anomalies in data flow on the NIDS dataset. Hybrid models, such as Convolutional Long Short-Term Memory networks (C-LSTM), integrate both convolutional and recurrent layers to improve the detection of anomalies in sequences. These models are particularly adept at recognizing both the spatial and temporal aspects, rendering them ideal for navigating intricate network scenarios [6] [7].

Research into techniques like Log2Vec and IP2Vec has also been conducted to improve the initial processing of syslog and dataflow data. Log2Vec, drawing inspiration from word2vec, transforms log templates into numerical vectors by considering semantic connections, thereby enhancing the comprehension of context for the purpose of identifying anomalies. Nonetheless, its significant computational expense and variable results underscore the necessity for refinement. In a similar vein, IP2Vec categorizes IPs according to their behavioral likenesses, which boosts anomaly detection but elevates the need for more computational resources. [3].

Even with these progressions, obstacles persist. Datasets such as CTU-13 and LogHub, though commonly utilized, do not completely mirror the intricacies found in actual-world situations. Moreover, the computational demands of deep learning models, especially their reliance on extensive, high-quality labeled datasets, restrict their ability to scale in environments with limited resources. A potential solution lies in hybrid strategies that combine graph mining for preliminary grouping with deep learning for precise categorization. By simplifying the data through clustering, these techniques enable deep learning models to concentrate on areas of higher risk, enhancing both efficiency and the use of resources [5] [6] [7] [8].

"Real-Time Identification of Cyber Threats: An Overview of Supervised Learning Approaches with the CTU-13 dataset" explores the use of supervised learning algorithms for the immediate identification of cyber threats through the CTU-13 dataset. This dataset includes a wide variety of network traffic events, offering a strong basis for creating intrusion detection models that can differentiate between harmless and harmful activities. The study utilizes established supervised learning techniques to improve the accuracy and dependability of intrusion detection systems, with the goal of correctly recognizing and classifying examples of both harmless and harmful activities within network data. [3].

Interconnection of Graph mining and deep learning is changed framework of botnet and anomaly detection, addressing the limitations of conventional methods. Graph mining is great for preprocessing and scalability, while deep learning provides better accuracy and flexibility for data analysis. Future research works on combining these methods, for improving efficiency and testing them to ensure they are scalable and reliable. Altogether, this creates a strong system for dealing with cyber threats. By using these improvements, cybersecurity can better handle the growing challenges of advanced attacks, protecting important systems and networks more accurately and reliably.

DATA PREPROCESSING

Data pre-processing is an important, necessary and critical step to make raw data usable for analysis and modeling processes. Data sets, especially those used in cyber security, contain missing values, conflicting information and complex structures. Such problems directly affect model performance and the analysis results to be obtained. For this reason, the data preprocessing phase stands out as one of the fundamental building blocks of the project.

The datasets used are extracted from the cybersecurity dataset CTU-13 and named as processed_pcap_data1.csv, processed_pcap_data2.csv, ..., processed_pcap_data13.csv. These datasets contain different features for analyzing network traffic. These features include IP addresses, timestamps, protocol information and data length. Analyses of the data showed that operations such as filling in missing values, transforming and scaling numeric and categorical features were necessary.

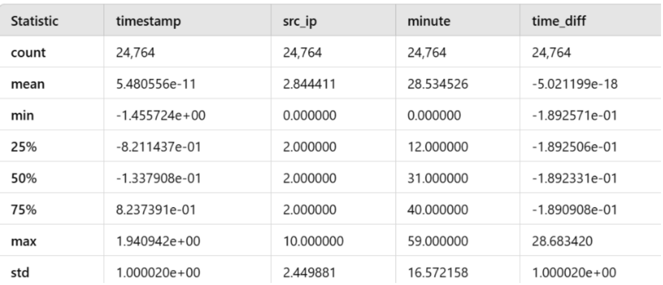
In the process of filling in the missing values, the mean value was used for numerical columns and the most frequently repeated value (mode) was used for categorical columns. This method allowed the missing values in the datasets to be filled and the data to have a more complete structure. In addition, new features such as hours and minutes derived from timestamps were added to give meaning to the datasets, and the time_diff column was created by calculating the differences between consecutive timestamps. These features facilitate the detection of time-based anomalies in network traffic. In addition, conflicting or inaccurate data in the dataset was also cleaned and implausible values were removed. In this process, duplicate data were removed and anomalous appearances were checked.

The fact that the numeric columns in the datasets have different scales was considered as a factor that could negatively affect the model performance. To resolve this issue, all numeric columns were converted to a standard scale using StandardScaler. In addition, categorical features such as protocols were converted into numerical data using Label Encoding. This process enabled the model to handle all features.

As a result of all these processes, the data sets were made suitable for modeling. Missing values were filled in, new and meaningful features were derived, categorical data were transformed and scaling processes were performed. In their latest form, the data sets have gained usability in the analysis process and have been equipped with an infrastructure that can detect potential threats and anomalies. These pre-processing steps enabled more accurate detection of anomalies in the dataset and identification of potential cyber threats. This comprehensive preprocessing process has provided an important basis for obtaining healthy results in the modeling process and creating a reliable analysis infrastructure

EXPLORATORY DATA ANALYSIS (EDA)

The purpose of this section is to perform an exploratory data analysis (EDA) on two datasets, **Dataset 6** and **Dataset 9**, **Dataset 8** and **Dataset 12** to identify patterns, detect anomalies, and understand feature relationships. The analysis employs the Isolation Forest algorithm to highlight anomalies within network traffic data. The features analyzed include protocol, timestamp, source IP, destination IP, and packet length.



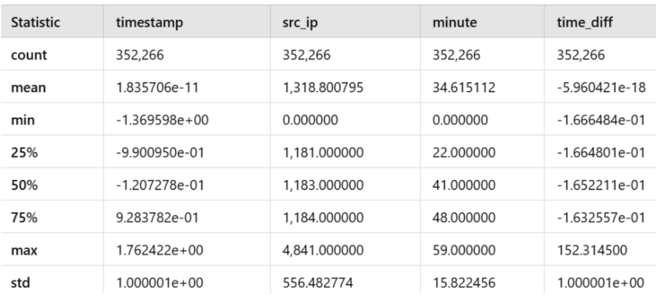
Statistic	timestamp	src_ip	minute	time_diff
count	24,764	24,764	24,764	24,764
mean	5.480556e-11	2.844411	28.534526	-5.021199e-18
min	-1.455724e+00	0.000000	0.000000	-1.892571e-01
25%	-8.211437e-01	2.000000	12.000000	-1.892506e-01
50%	-1.337908e-01	2.000000	31.000000	-1.892331e-01
75%	8.237391e-01	2.000000	40.000000	-1.890908e-01
max	1.940942e+00	10.000000	59.000000	28.683420
std	1.000020e+00	2.449881	16.572158	1.000020e+00

Table 3: Dataset 6 Summary Statistics

Dataset 6 consists of 24,764 data points, offering key insights into network traffic behavior. The src_ip feature demonstrates a notable concentration on two primary IP addresses, with a maximum of 10 unique IPs observed in the dataset. This indicates that a majority of the network activity originates from a limited number of sources. The minute feature exhibits a mean value of 28.53, suggesting that most traffic occurs within 30-minute intervals, which may reflect consistent network activity patterns.

The time_diff feature presents both negative and positive values, showcasing significant variability in time differences between events. This wide range may suggest the presence of outliers, which could signal irregular or anomalous behavior in the dataset. Additionally, the standard deviation values for these features indicate some degree of dispersion, further emphasizing the dynamic nature of the observed data.

Overall, the statistical distribution of features in Dataset 6 provides a comprehensive foundation for deeper exploration of network anomalies, particularly in relation to temporal behaviors and source activity patterns. These findings can guide further analyses aimed at identifying irregular traffic or security threats within the network.



Statistic	timestamp	src_ip	minute	time_diff
count	352,266	352,266	352,266	352,266
mean	1.835706e-11	1,318.800795	34.615112	-5.960421e-18
min	-1.369598e+00	0.000000	0.000000	-1.666484e-01
25%	-9.900950e-01	1,181.000000	22.000000	-1.664801e-01
50%	-1.207278e-01	1,183.000000	41.000000	-1.652211e-01
75%	9.283782e-01	1,184.000000	48.000000	-1.632557e-01
max	1.762422e+00	4,841.000000	59.000000	152.314500
std	1.000001e+00	556.482774	15.822456	1.000001e+00

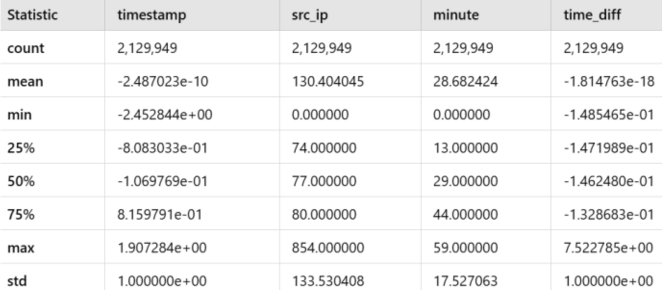
Table 4: Dataset 12 - Summary Statistics

Dataset 12 consists of 352,266 data points, providing valuable insights into the characteristics of network traffic. The src_ip feature shows a significant variation in source IP addresses, with an average of 1,318.8 IPs and a maximum of 4,841 unique IPs. This diversity reflects the extensive origins of network activity captured in the dataset.

The minute feature has a mean value of 34.61, indicating that traffic is relatively evenly distributed across an hour, with data points ranging from 0 to 59 minutes. The standard deviation of 15.82 suggests moderate variation in the temporal distribution of events.

The time_diff feature spans a wide range, from -0.166 to 152.31, with a mean value close to zero. This variability indicates the presence of irregular events or potential anomalies. The standard deviation of 1.0 highlights that most data points are concentrated near the mean, but extreme deviations suggest noteworthy outliers.

Overall, Dataset 12 reflects a dynamic and diverse network activity pattern, with significant variations in source IPs and temporal behaviors. These characteristics make it a strong candidate for advanced anomaly detection and further exploration of network dynamics.



Statistic	timestamp	src_ip	minute	time_diff
count	2,129,949	2,129,949	2,129,949	2,129,949
mean	-2.487023e-10	130.404045	28.682424	-1.814763e-18
min	-2.452844e+00	0.000000	0.000000	-1.485465e-01
25%	-8.083033e-01	74.000000	13.000000	-1.471989e-01
50%	-1.069769e-01	77.000000	29.000000	-1.462480e-01
75%	8.159791e-01	80.000000	44.000000	-1.328683e-01
max	1.907284e+00	854.000000	59.000000	7.522785e+00
std	1.000000e+00	133.530408	17.527063	1.000000e+00

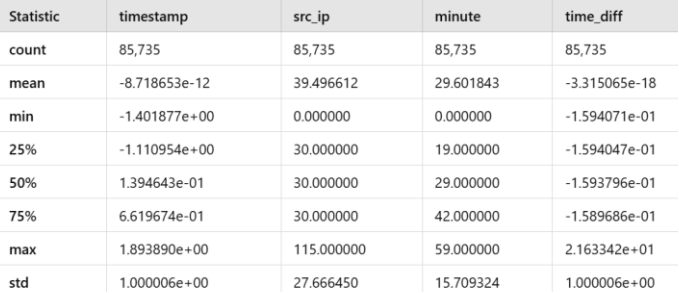
Table 5: Dataset 9 - Summary Statistics

Dataset 9 contains 2,129,949 data points, making it one of the most extensive datasets analyzed. The src_ip feature displays a diverse range of source IP addresses, with a mean value of 130.40 and a maximum of 854 unique IPs. This indicates substantial network activity originating from a wide variety of sources.

The minute feature has a mean value of 28.68, demonstrating that the majority of the network activity is evenly distributed across an hour, with data points ranging from 0 to 59 minutes. The standard deviation of 17.53 highlights moderate temporal variability, suggesting that some intervals have a higher concentration of events.

The time_diff feature ranges from -0.148 to 7.52, with a mean value near zero. This suggests a relatively balanced temporal distribution, though the maximum value indicates the presence of significant outliers. The standard deviation of 1.0 supports the observation that most values are clustered close to the mean.

Overall, Dataset 9 provides a comprehensive view of highly varied network activity, with significant diversity in IP sources and temporal behaviors. The presence of extreme values in time_diff and a wide range of source IPs makes this dataset an excellent candidate for detecting anomalies and investigating irregular network traffic patterns.



Statistic	timestamp	src_ip	minute	time_diff
count	85,735	85,735	85,735	85,735
mean	-8.718653e-12	39.496612	29.601843	-3.315065e-18
min	-1.401877e+00	0.000000	0.000000	-1.594071e-01
25%	-1.110954e+00	30.000000	19.000000	-1.594047e-01
50%	1.394643e-01	30.000000	29.000000	-1.593796e-01
75%	6.619674e-01	30.000000	42.000000	-1.589686e-01
max	1.893890e+00	115.000000	59.000000	2.163342e+01
std	1.000006e+00	27.666450	15.709324	1.000006e+00

Table 4: Dataset 8 - Summary Statistics

Dataset 8 contains 85,735 data points and showcases distinct network traffic patterns. The src_ip feature indicates a moderate diversity in source IP addresses, with an average of 39.49 and a maximum of 115 unique IPs. This range suggests that network traffic originates from a smaller number of consistent sources compared to other datasets.

The minute feature has a mean value of 29.60, highlighting that the majority of network activity is distributed evenly across an hour, with most events occurring between the 19th and 42nd minutes. The standard deviation of 15.70 suggests moderate variability in the temporal distribution of events.

The time_diff feature spans from -0.159 to 21.63, with a mean value close to zero. This implies that while the majority of values are near the mean, there are occasional extreme differences, which may indicate significant outliers or anomalies. The standard deviation of 1.0 reinforces that most data points are tightly clustered around the mean.

In summary, Dataset 8 reveals a relatively stable and consistent network traffic pattern with limited variability in temporal and source IP behavior. The presence of a few extreme values in time_diff warrants further investigation for potential anomalies or irregular traffic patterns.

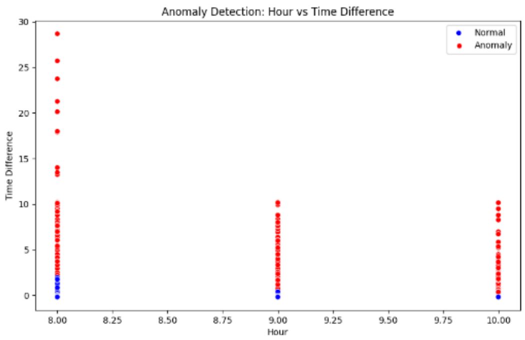


Figure 1.1 Dataset 6 -Anomaly Detection: Hour vs Time Difference.

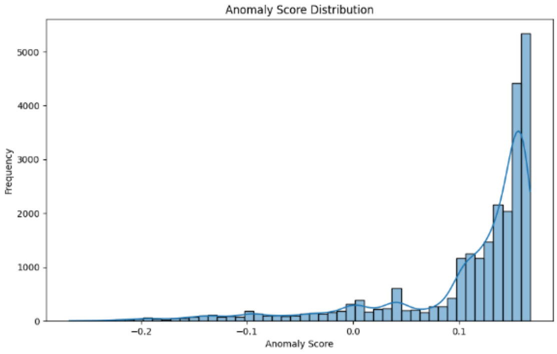


Figure 2.1 Dataset 6 - Anomaly Score Distribution

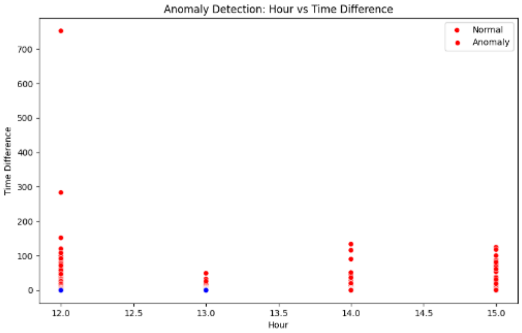


Figure 1.2 Dataset 9- Anomaly Detection: Hour vs Time Difference.

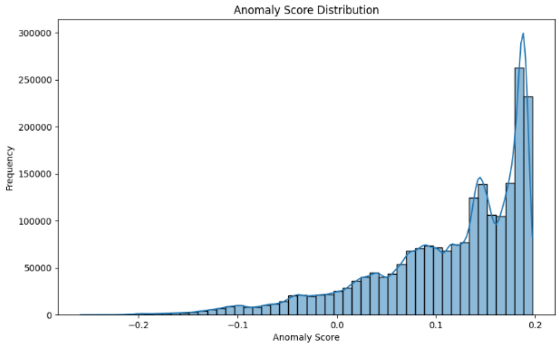


Figure 2.2 Dataset 9 - Anomaly Score Distribution

Figure 1.1 (Dataset 6): Anomalies (red points) are concentrated around hours 8, 9, and 10, suggesting unusual activity during these intervals. Most anomalies occur within a time difference range of 10 or higher, indicating distinct patterns of irregular network behavior during the morning hours.

Figure 1.2 (Dataset 9): Anomalies are primarily observed around **hours 12, 13, and 14**, reflecting a different pattern compared to Dataset 6. This dataset also exhibits a **wider range of time differences,** with anomalies occurring at higher values, suggesting more complex or diverse network activity.

These findings highlight distinct anomaly patterns in the two datasets. The differences may be attributed to variations in network traffic protocols, user behavior, or the underlying structure of the datasets.

In **Dataset 6**, the anomaly scores are mostly concentrated between **0.0 and 0.1**, indicating that the majority of data points were classified as **normal** by the Isolation Forest algorithm. Only a small fraction of data points, with lower anomaly scores, were flagged as **anomalies**.

This dataset displays relatively consistent network behavior, characterized by minimal variations and fewer anomalies, reflecting a stable traffic pattern.

In **Dataset 9**, the anomaly scores exhibit a much broader distribution compared to Dataset 6. A significant number of data points have anomaly scores above **0.1**, indicating more diverse and complex network behavior.

The broader distribution suggests that this dataset contains more variability in network activity, possibly due to different protocols, user behaviors, or an increased volume of traffic. While most data points remain classified as **normal**, the presence of a wider range of anomaly scores highlights the dataset's dynamic nature.

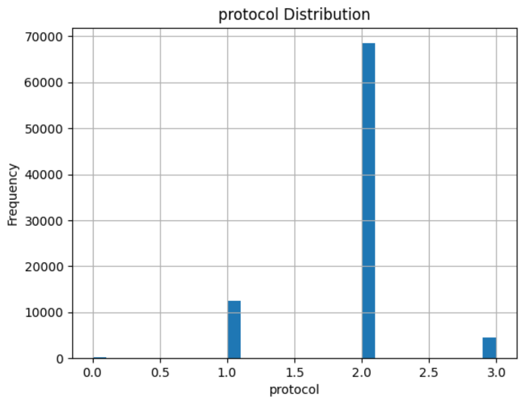


Figure 3.1 Dataset 8-Protocol Distributions

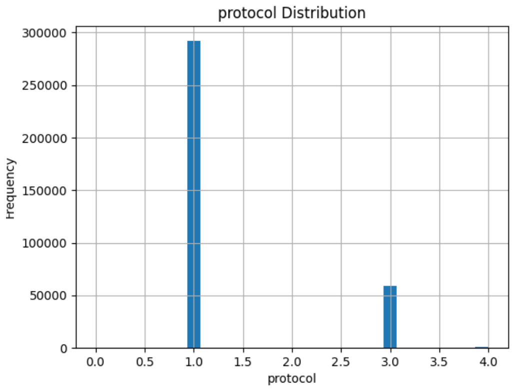


Figure 3.2 Dataset 12-Protocol Distributions

Dataset 8 (Figure 3.1)

The protocol distribution in Dataset 8 reveals a dominant presence of `Protocol 2` (e.g., TCP), with over 70,000 occurrences. This indicates a network heavily reliant on TCP traffic, typically associated with reliable data transmission. `Protocol 1` (e.g., UDP) accounts for approximately 10,000 occurrences, suggesting limited use for streaming or low-latency services. `Protocol 3` is minimally represented.

This pattern indicates a more uniform network traffic pattern in Dataset 8, likely resulting in concentrated anomalies within `Protocol 2`.

Dataset 12 (Figure 3.2)

The protocol distribution in Dataset 12 is markedly different, with `Protocol 1` dominating (over 250,000 occurrences), followed by a significant presence of `Protocol 3` (50,000+ occurrences). `Protocol 2` is nearly absent, indicating a shift from TCP to UDP-heavy traffic, potentially for real-time applications like video streaming or VoIP

The contrasting distributions between the two datasets highlight diverse network behaviors:

- Dataset 8: Uniform, TCP-dominated traffic.

-Dataset 12: Diverse traffic patterns, primarily driven by UDP and secondary protocols.

These differences may significantly impact anomaly detection patterns, with Dataset 8 anomalies likely concentrated in TCP traffic, while Dataset 12 anomalies may be distributed across UDP and other protocols.

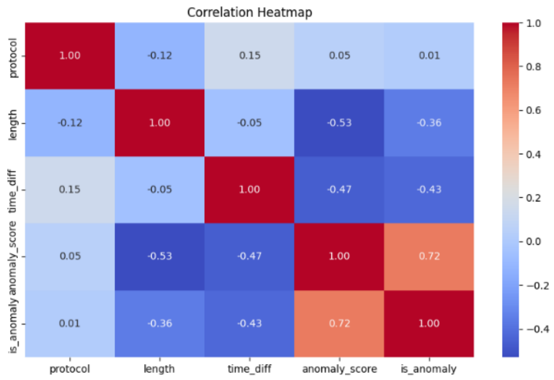


Figure 4.1 Dataset 8 - Correlation Heatmap

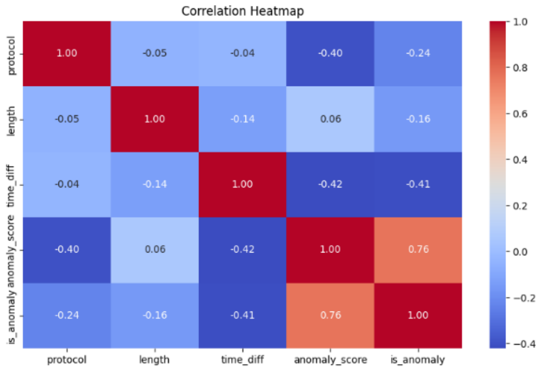


Figure 4.2 Dataset 12 - Correlation Heatmap

In Dataset 8, the strongest correlation is observed between ***anomaly_score*** and ***is_anomaly*** (0.72), highlighting the central role of the ***anomaly_score*** in detecting anomalous activities. Other features, such as protocol and ***time_diff***, exhibit weaker correlations with ***anomaly_score*** and ***is_anomaly***, indicating their limited contribution to the anomaly detection process. Notably, the protocol feature demonstrates negligible correlation with ***is_anomaly*** (0.01), suggesting that protocol type has little influence in this dataset.

In contrast, the heatmap for **Dataset 12** reveals a stronger correlation between ***anomaly_score*** and ***is_anomaly*** (0.76), suggesting a more robust anomaly detection model for this dataset. Furthermore, the protocol feature exhibits a more pronounced correlation with ***is_anomaly*** (0.24) compared to Dataset 8. This finding implies that the protocol type plays a more significant role in detecting anomalies within Dataset 12, possibly due to variations in network configurations or traffic patterns.

While the length and ***time_diff*** features show weak correlations with ***anomaly_score*** and ***is_anomaly*** in both datasets, their relative consistency suggests limited variability in their influence across datasets. These observations indicate that while some features have dataset-specific importance, others, like length and ***time_diff***, remain consistent in their limited contribution to anomaly detection.

The comparison of the two heatmaps highlights the differences in feature relationships between Dataset 8 and Dataset 12. The stronger correlation between ***anomaly_score*** and ***is_anomaly*** in Dataset 12 suggests a more refined anomaly detection process, while the increased relevance of the protocol feature reflects potential differences in traffic protocols or underlying network behaviors

DATA MINING TECHNIQUE SECTION AND APPLICATION

The Isolation Forest algorithm is one of the most suitable methods for detecting anomalies, especially in high-dimensional and complex datasets. In this study, a processed dataset based on the CTU-13 dataset has been used. In the context of cybersecurity, the Isolation Forest algorithm is preferred due to its efficiency, speed, ease of distinguishing abnormal data points, flexibility in parameters, and its insensitivity to distribution. This algorithm does not rely on any distribution assumptions and provides effective results with large datasets. The user can adjust the model's sensitivity by setting the contamination parameter to determine the anomaly rate.

The Isolation Forest algorithm isolates data points by creating random decision trees. Points that are isolated in a short time are considered anomalies. In the study, the model parameters were set as follows: n_estimators=100, meaning the model creates 100 isolation trees; contamination=0.1, assuming that 10% of the data contains anomalies; and random_state=42, ensuring consistency of results. During model training, an anomaly score is calculated for each data point, and each data point is classified as 1 (normal) or -1 (abnormal).

When evaluating the model's performance, it was found that out of 22,764 data points, 22,288 were normal and 2,476 were abnormal. This shows that the model works with 90% accuracy and results in a 10% anomaly rate. The 90% accuracy rate obtained by the model is in line with the generally accepted success criteria for anomaly detection. In cybersecurity applications, especially with large and dynamic datasets such as network traffic, accuracy rates above 85% are generally considered successful [9]. This rate demonstrates the model's ability to both identify real anomalies and accurately classify normal traffic. However, the results obtained also include certain limitations and limitations in the application context. For example, the false positive rate of the model is a parameter that needs to be analyzed in depth, especially in sensitive security systems. When visualizing the distribution of anomaly scores, it is clearly seen that values at the extremes are considered anomalies. Furthermore, in the scatterplot analysis based on the "hour" and "time_diff" features, it was observed that anomalies were concentrated at specific hours.

The types of anomalies detected include network traffic issues, IP-based anomalies, and protocol-based anomalies. High "time_diff" values indicate delays or congestion in the network, while isolated, unusual IP addresses point to unauthorized access attempts. Protocol-based anomalies indicate unusual protocol usage. These findings point to critical threats in cybersecurity. In particular, unauthorized access attempts, botnet activities, system performance degradation, and DDoS attacks are potential threats.

In practice, the detected anomalies can be used in network security management. These anomalies allow network administrators to quickly respond to potential attacks. Additionally, anomaly analyses help strengthen security policies and can be used to prevent performance issues in the network. In the visualizations, the time frames and time differences where anomalies are concentrated are clearly observed. These visual results provide critical information for network management and security experts.

The CTU-13 dataset, which realistically simulates network traffic scenarios, was effectively used in detecting anomalies with the Isolation Forest algorithm. This study forms an important foundation for improving network security and responding more quickly to cyber threats. In the future, optimizing parameters, adding new features to the dataset, and comparing different algorithms can make the model more precise and robust.

In conclusion, the Isolation Forest algorithm effectively detected anomalies in network traffic data and provided critical information for cybersecurity and network management. This study offers an analysis that can be used as a basis for improving network security and solving performance problems.

Summary of Analysis Results and Discussion on Accuracy and Effectiveness of Applied Data Mining Techniques In this study, we applied the Isolation Forest algorithm to detect anomalies in the CTU-13 dataset, which simulates network traffic, including botnet activities. The Isolation Forest model effectively identified anomalous patterns, with the results showing that 90% of the data were classified as normal, while 10% were deemed anomalous. The anomaly detection was based on various network traffic features, including timestamps, IP addresses, and protocols. The model's ability to detect outliers and deviations in network behavior was evaluated using anomaly scores and visualizations, which demonstrated that anomalies were concentrated in specific times and exhibited unusual traffic patterns.

The model's accuracy, based on the anomaly distribution, is consistent with the parameters set for contamination, showing a 90% correct classification of normal data points and a 10% classification of anomalous data points. The anomaly scores were effectively utilized to distinguish between normal and abnormal data, with the distribution indicating that extreme values were correctly identified as anomalies. The graphical visualizations, including the scatterplot and histogram, further supported the accuracy of these findings by clearly delineating the differences between normal and anomalous data.

Regarding the effectiveness of the Isolation Forest algorithm in the context of anomaly detection for cybersecurity, the technique proved to be highly efficient in detecting irregularities in large-scale network traffic datasets like CTU-13. The algorithm's ability to operate without any distributional assumptions makes it well-suited for complex datasets with diverse data distributions, such as the ones found in network traffic analysis. Additionally, the speed and scalability of Isolation Forest allowed it to process a large dataset efficiently, which is crucial in real-time anomaly detection systems.

However, while the model performed well in this context, there are opportunities for further enhancement. For instance, optimizing hyperparameters such as n_estimators and contamination could improve detection accuracy, especially in more diverse or complex network traffic scenarios. Additionally, integrating other techniques, such as feature engineering or using hybrid models (combining graph mining and deep learning), could increase the model's robustness in detecting subtler anomalies.

In conclusion, the Isolation Forest algorithm demonstrated high accuracy in detecting anomalies in the CTU-13 dataset, making it a useful tool for cybersecurity applications. Its effectiveness in identifying abnormal network behaviors, such as unauthorized access, botnet activities, and system performance issues, underscores its potential in real-world cybersecurity systems. Further refinements and hybrid approaches could enhance the algorithm's performance and extend its applicability to other cybersecurity challenges.

CONCLUSION AND FUTURE WORKS

In this project, a data mining approach for anomaly detection in the cybersecurity domain is developed using the CTU-13 dataset. The Isolation Forest algorithm was used to identify unusual behaviors in network traffic and successful results were obtained with an accuracy rate of 90%. In the key stages of the study, missing values in the dataset were filled, time-based features were extracted and the data was scaled by data preprocessing. These steps played a role in improving the accuracy of the model and allowing for more accurate analysis.

The findings point to anomalies in network traffic that are concentrated in certain time periods and protocol types. For example, anomalies in Dataset 6 were concentrated in the morning hours, while anomalies in Dataset 9 occurred in the afternoon hours. Furthermore, protocol-based analysis reveals that TCP-dominated networks exhibit more consistent anomaly behavior, while UDP-dominated networks show more complex and variable anomalies. These findings provide important clues for understanding threat scenarios that differ according to network structures and traffic types.

The Isolation Forest algorithm has proven to be an important and powerful tool for cybersecurity applications by providing fast and effective anomaly detection in high-dimensional and complex data sets. However, the success of the model relies not only on the power of the algorithm, but also on properly executed data preprocessing and feature engineering steps. The study clearly illustrates the viability and potential of data mining techniques for proactive threat detection in the cybersecurity domain.

As a result, this study has made a significant contribution to data mining approaches used to detect unusual behavior in network traffic and has laid a foundation for the development of more effective solutions against cybersecurity threats. The success of the Isolation Forest algorithm opens the door for further techniques and methods to be explored in this field._

The results of this study point to the potential for further application in the field of cybersecurity. As a future work, however, although the CTU-13 dataset reflects the real world, the use of more diverse and real-time datasets could improve the generalizability of the model. Furthermore, hyper parameter optimization and hybrid approaches with techniques such as graph mining or deep learning are suggested to improve the model performance. Future work should aim to integrate this model into real-time systems and evaluate its wider applicability.

REFERENCES

[1] Stratosphere IPS. (n.d.). CTU-13 dataset. Retrieved November 17, 2024. [Online]. Avaliable: https://www.stratosphereips.org/datasets-ctu13

[2] Impact Cyber Trust. (n.d.). CTU-13 botnet dataset. Retrieved November 17, 2024. [Online] Avaliable: https://impactcybertrust.org/dataset_view?idDataset=945

[3] A. Sharma and H. Babbar, "Detecting cyber threats in real-time: A supervised learning perspective on the CTU-13 dataset," *5th International Conference for Emerging Technology (INCET)*, Karnataka, India, May 24-26, 2024..

[4] S. K. Srinarayani, Dr. B. Padmavathi, and Mrs. D. Kavitha, "Detection of botnet traffic using deep learning approach," *Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2023)*, Chennai, India, 2023.

[5] D. J. Borah and A. Sarma, "Detection of Peer-to-Peer Botnets Using Graph Mining," *International Journal of Computer Networks & Communications (IJCNC)*, vol. 15, no. 2, Mar. 2023.

[6] A. A. Ahmed, W. A. Jabbar, A. S. Sadiq, and H. Patel, "Deep learning-based classification model for botnet attack detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 10,Feb. 2020.

[7] K. Macková, D. Benk, and M. Šrotýř, "Enhancing Cybersecurity Through Comparative Analysis of Deep Learning Models for Anomaly Detection," *Proceedings of the 10th International Conference on Information Systems Security and Privacy (ICISSP 2024)*, Prague, Czech Republic, pp. 682-690, 2024

[8] K. Sinha, A. Viswanathan, and J. Bunn, "Tracking Temporal Evolution of Network Activity for Botnet Detection," Aug. 2019.

[9] H. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & Security*, vol. 28, no. 1-2, pp. 18-28, 2009. DOI: 10.1016/j.cose.2008.08.003.

[10] [1] hsila-ui, "ForAssignment," GitHub. [Online]. Available:https://github.com/hsila-ui/ForAssignment [Accessed: Jan. 7, 2025].