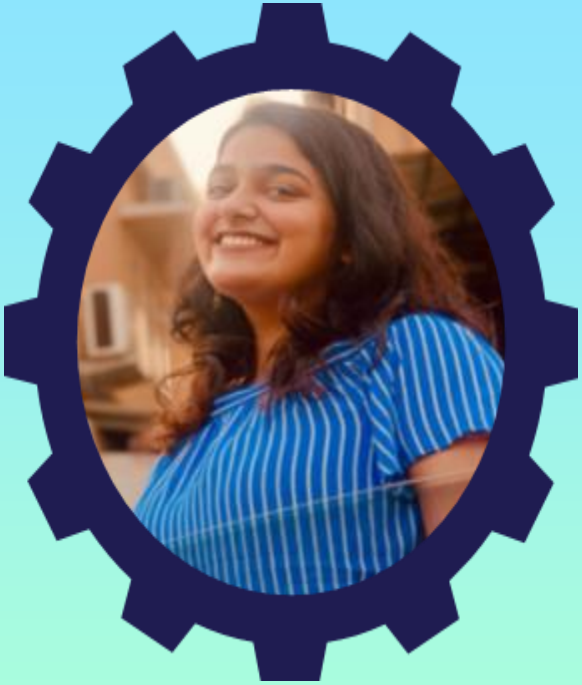
Cybersecurity Threat Analysis & Dataset Integration

Data Acquisition and Pre-Processing - DSCI-511-900
Professor Dr. Milad Toutounchian

TEAM MEMBERS



Georgina George

Jessica Delgado

Jilu James

OVERVIEW

Dataset

40,000 records, 25 attributes, detailing cyberattacks

Key Features

Attack types, Ips, protocols, anomaly scores, severity levels

Fosus

- Analyze severity, anomaly scores, attack types, and geographic origins
- Highlight evolving tactics of cyber threats

Objective

- Uncover insights on current cybersecurity trends
- Discuss strategies to mitigate emerging risks

DATA SOURCES

Kaggle

We obtained one dataset from Kaggle via an Excel file

Data References

- ✓ vizsec.org
- ✓ csr.lanl.gov
- ✓ Github: Real CyberSecurity Datasets
- ✓ UNB Datasets
- ✓ Github.com: Cybersecurity Datasets
- ✓ csis.org

CSIS Website

Extracted data from this website & converted into an Excel file

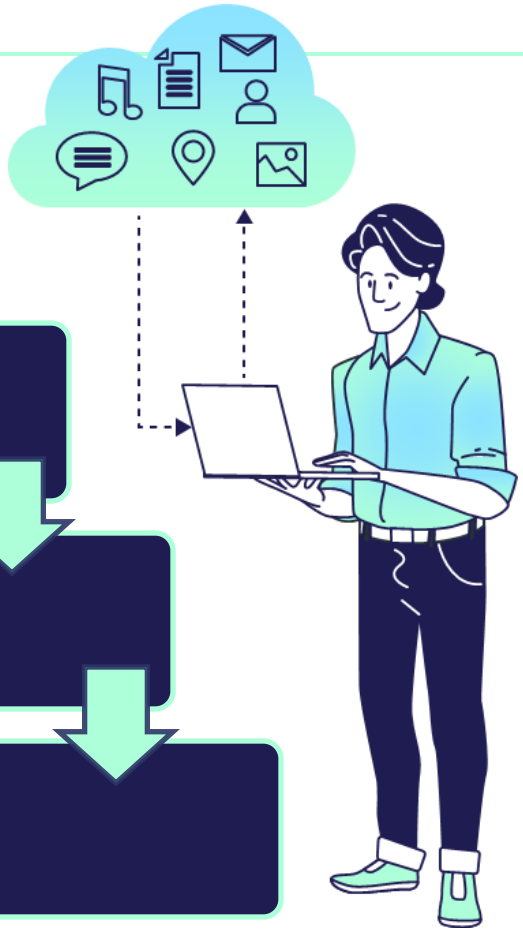
DATA ACQUISITION APPROACH

Data Sources

Data Extraction & Processing

Data Cleaning & Integration

Data Insights



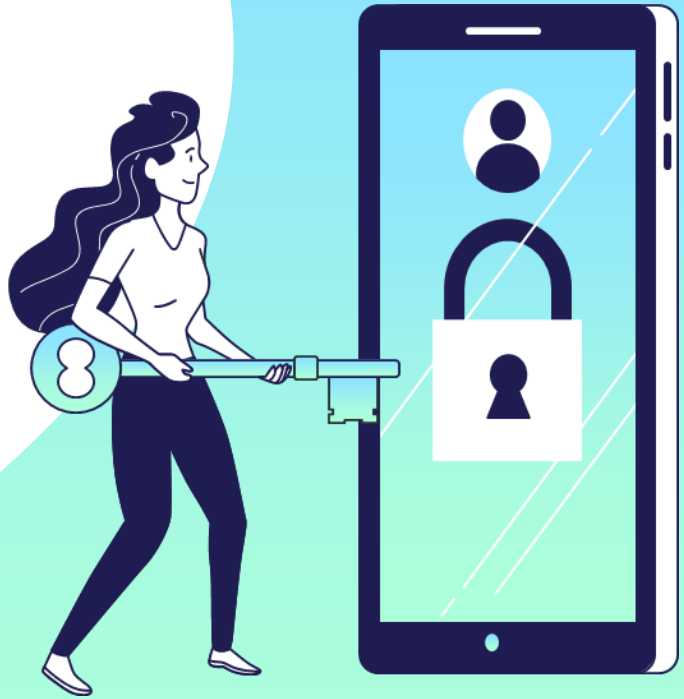
APPROACH TO PREPROCESSING DATA

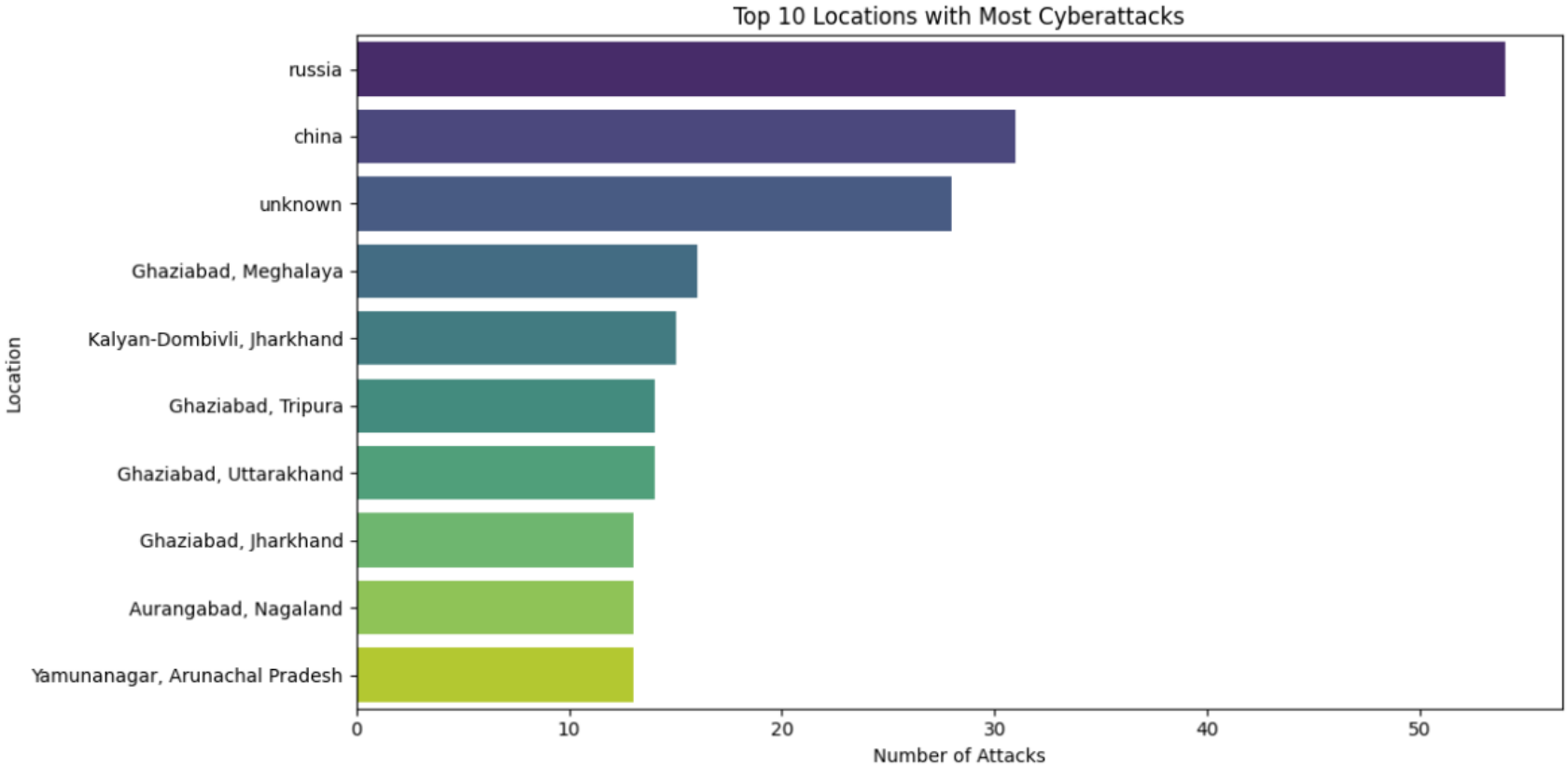
1. Data Cleaning
2. Data Standardization
3. Data Filtering
4. Handling Inconsistent Units
5. Data Integration
6. Feature Engineering
7. Data Export
8. Documentation

DATASET CONSTRUCTED

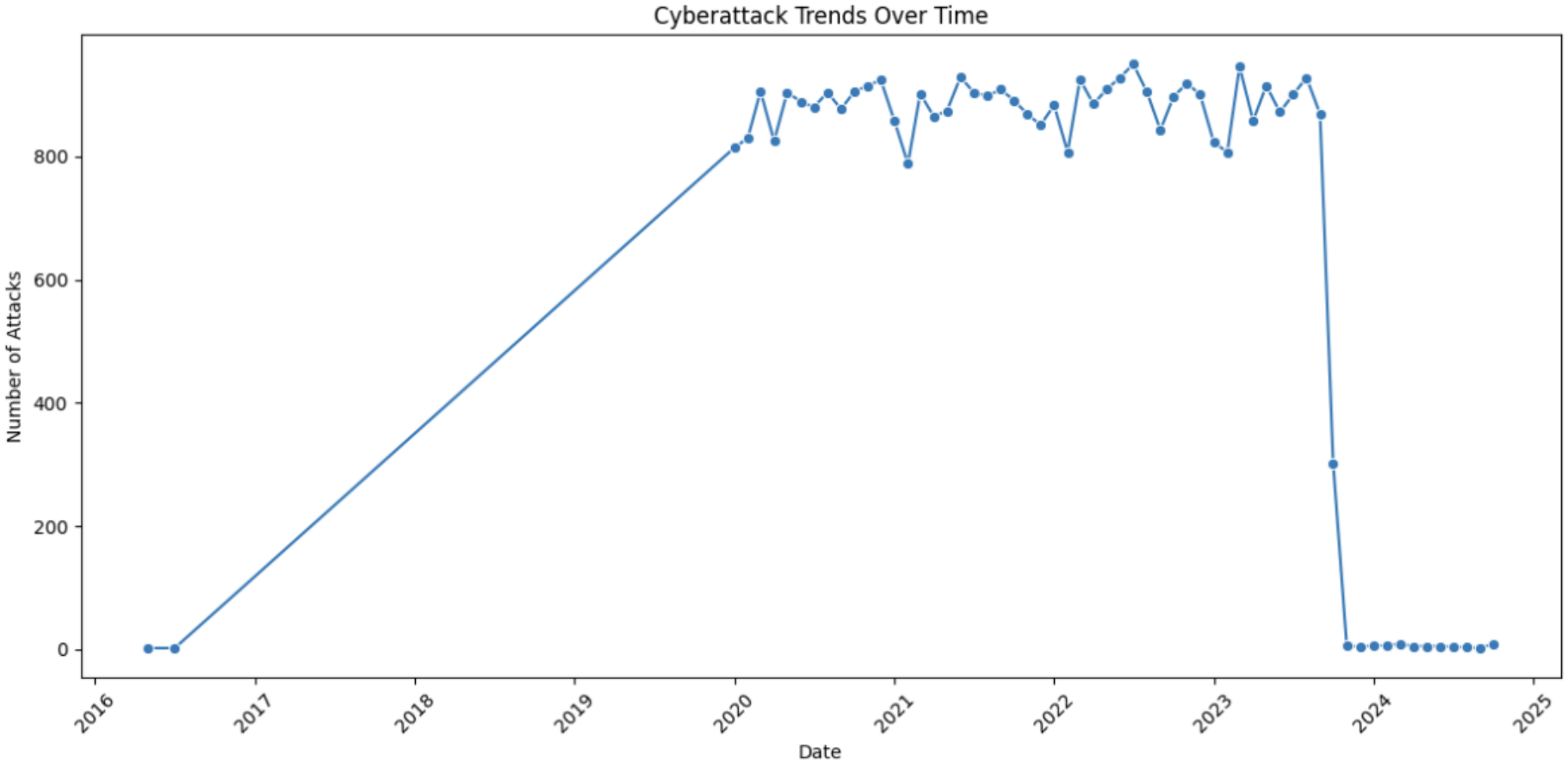
	Timestamp	Geo-location Data	User Information	Attack Type	Severity Level
2	October 2024	Russia	russian agents sent emails about bomb threats to nearly 60 ukrainian embassies	Unknown	Low
3	October 2024	Iran	iranian agents are increasing their espionage efforts against government agencies	Espionage	Low
4	October 2024	Russia	russian cybercriminals sent information-stealing malware to an unknown	Malware	Low
5	October 2024	United States	australia introduced its first national cyber legislation, the cyber security bill	Unknown	Low
6	October 2024	China	chinese hackers have breached at least twenty canadian government networks	Intrusion	Low
7	October 2024	Russia	russian hackers sent compromised emails disguised to appear as if they were	Unknown	Low
8	October 2024	China	chinese hackers hacked cellphones used by senior members of the trump-vance	Intrusion	Low
9	October 2024	China	new reporting reveals chinese-backed hackers have been conducting large data	Unknown	Low
10	October 2024	Russia	ukrainian hackers attacked russia's state media company and electronic court	Unknown	Low
11	September 2024	China	chinese hackers have been conducting an ongoing cyber espionage campaign	Espionage	Low
12	September 2024	Russia	russian cyber spies conducted an espionage campaign against mongolia's	Espionage	Low
13	August 2024	Iran	u.s. government officials blamed iranian hackers for breaking into donald	Unknown	Low
14	August 2024	United States	the united nations unanimously approved its first treaty on cybercrime. the	Unknown	Low
15	August 2024	Russia	russian cyber criminals are deploying malware against diplomats through a used-	Malware	Low
16	July 2024	South Korea	south korea's military is investigating the leak of highly sensitive information on	Data Leak	Low
17	July 2024	Unknown	a faulty software update for microsoft windows issues by cybersecurity firm	Unknown	Low
18	July 2024	China	germany accused china of directing a "serious" cyberattack against germany's	Unknown	Low
19	July 2024	United States	australia, the united states, canada, the united kingdom, germany, japan, south	Unknown	Low
20	June 2024	Japan	japan's space agency has suffered a series of cyberattacks since last year, according	Unknown	Low

DATA VISUALS & INSIGHT SAMPLES

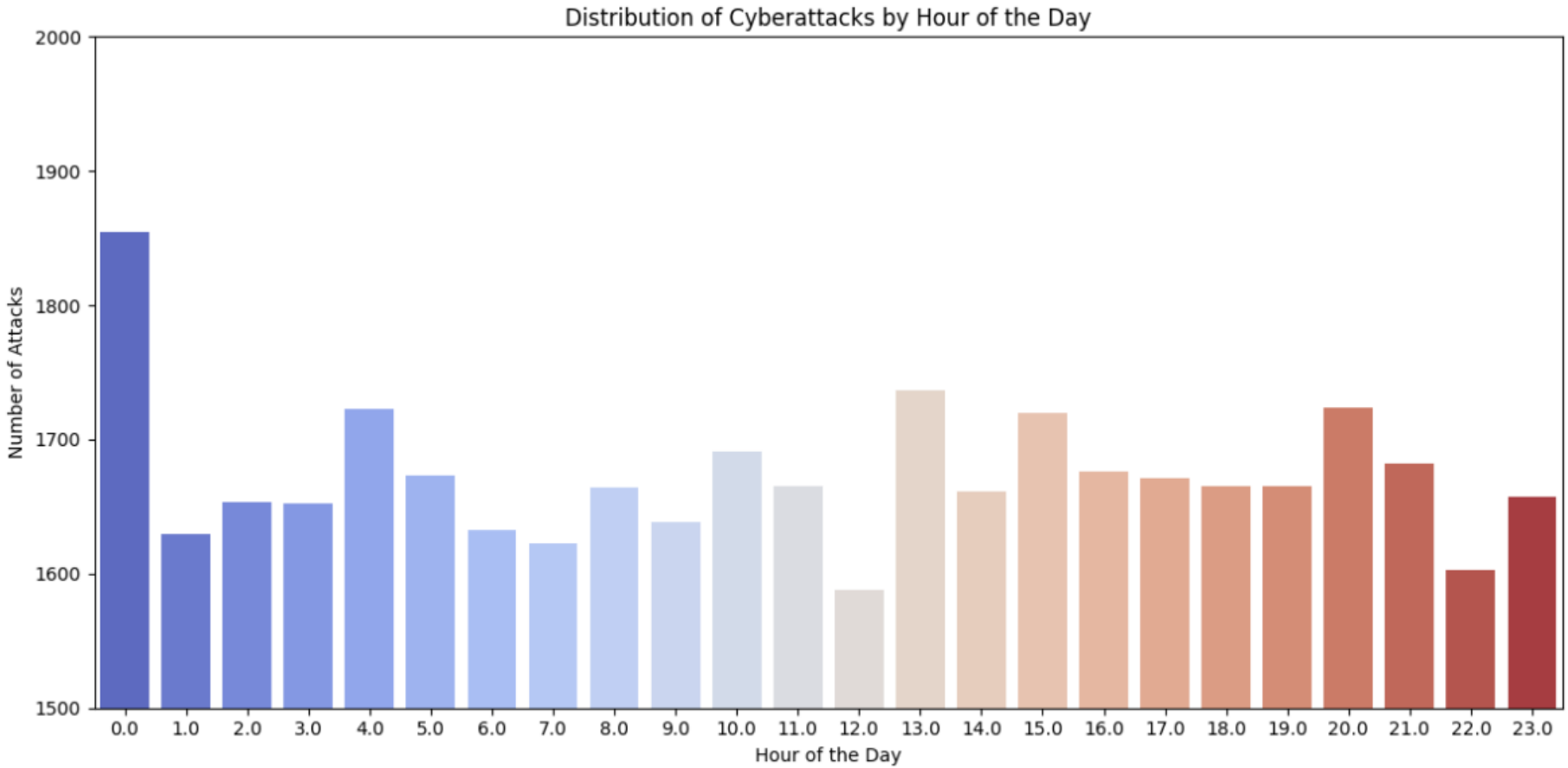




The bar chart above shows the top 10 locations with the highest number of cyberattacks. Russia, China, and unknown locations are the most targeted, followed by various locations in India. There could be several reasons why there might be an "unknown" location for cyberattacks and the first reason would be due to a sophisticated attacker. Highly skilled hackers often use techniques to mask, making it difficult to trace. Attackers can also use networks of compromised computers (botnets) located in various countries to launch attacks as well as use proxy servers and VPNS. Therefore, the category of "unknown" reflects the limitations in tracking cyberattacks and shows the sophistication of modern cyber threats.



The line chart above illustrates the trend of cyberattacks over time. The data shows a significant increase in attacks from 2016 to 2020, with a peak in 2023. However, there is a sharp decline in attacks in 2024. This sharp decline could be attributed to factors such as heightened cybersecurity awareness, improved security technologies, economic shifts, and increased law enforcement efforts.



The bar chat above shows the distribution of cyberattacks across different hours of the day. There are peaks in the early morning hours, around 1-3 AM, and late evening, around 10 PM. This suggests that attackers may be more active during these times, potentially due to factors such as reduced security personnel or increased vulnerability of systems during off-peak hours.

POTENTIAL USERS & USAGE OF DATASET

- **Cybersecurity Researchers:** Identify attack patterns, develop threat detection models, and analyze defense systems.
- **IT Security Teams:** Enhance incident response and firewall effectiveness.
- **Academics & Students:** Utilize for educational and research purposes.
- **Network Administrators:** Manage firewall rules and ensure network segment security.
- **Government & Regulatory Bodies:** Analyze threats to inform policies and controls.

DISTRIBUTION APPROACH & ACCESS RIGHTS

- Delivered in Excel (.csv) format for seamless accessibility
- Completely free of personal data
- Easily shareable on Kaggle or any preferred data platform
- No limitations or restrictions on data usage

ISSUES & LIMITATIONS

ISSUES

- ✓ Incomplete Dataset Scope
- ✓ Data Gaps
- ✓ Description Accuracy
- ✓ Dynamic Nature of Cyber Threats

LIMITATIONS

- ✓ Dataset 1 vs. Dataset 2
- ✓ Limited Real-Time Data
- ✓ Static Data Sources

CONCLUSION

A comprehensive analysis of datasets revealed key insights into cyberattack trends, including severity, anomaly scores, attack types, and geographic patterns. The study identified vulnerabilities and highlighted the growing sophistication of attacks targeting critical infrastructure and sensitive data. Regional variations emphasize the need for tailored security strategies, while dataset limitations point to the importance of incorporating real-time data and details like timestamps and IP addresses for future analysis.