

# RESEARCH PAPER ON CYBER THREAT INTEL

1<sup>st</sup> Mitesh Yadav  
B.tech CSE - Cyber Security  
Acropolis Institute of technology,  
Indore, India  
[miteshyadav220174@acropolis.in](mailto:miteshyadav220174@acropolis.in)

4<sup>th</sup> Prof. Ashish Anjana  
B-tech CSIT & Cyber Security  
Acropolis Institute of technology,  
Indore, India  
[ashishanjana@acropolis.in](mailto:ashishanjana@acropolis.in)

2<sup>nd</sup> Yogesh Panchole  
B.tech CSE- Cyber Security  
Acropolis Institute of technology,  
Indore, India  
[yogeshpanchole221106@acropolis.in](mailto:yogeshpanchole221106@acropolis.in)

5<sup>th</sup> Prof. Satyam Srivastava  
B-tech CSIT & Cyber Security  
Acropolis Institute of technology,  
Indore, India  
[satyamshrivastava@acropolis.in](mailto:satyamshrivastava@acropolis.in)

3<sup>rd</sup> Khushi Patel  
B.tech CSE- Cyber Security  
Acropolis Institute of technology,  
Indore, India  
[yogeshnagar221176@acropolis.in](mailto:yogeshnagar221176@acropolis.in)

**Abstract—** Cyber Threat Intelligence (CTI) is integrated into the system design to identify, analyze, and mitigate emerging cyber threats. CTI enables early detection of suspicious activities, strengthens authentication mechanisms, improves log analysis, and ensures secure transmission of occupancy data. By mapping adversary tactics and vulnerabilities through frameworks such as MITRE ATT&CK and STRIDE, the system enhances its resilience against data tampering, spoofing, and privacy breaches. Incorporating CTI practices not only secures live .

**Keywords—** Cyber Threat Intelligence, Visual Counting Security, Threat Modeling, MITRE ATT&CK, STRIDE, Real-Time Monitoring, Secure Video Transmission, Data Integrity , Attack Mitigation, Privacy Protection, Secure Dashboard Architecture.

## INTRODUCTION

Cyber threats are increasing in frequency, sophistication, and global impact, making it essential to develop systems that can automatically collect, analyze, and interpret threat intelligence. This research presents an open-source Cyber Threat Intelligence (CTI) platform capable of scraping data from multiple cybersecurity sources, performing automated keyword and severity analysis, detecting India-related incidents, and presenting the information through an interactive dashboard with AI-assisted explanations. The methodology integrates web scraping, threat analysis, structured database storage, full-text search, and a Gemini-based AI chatbot to support users in understanding complex threat data. The system aims to provide a cost-effective, accessible, and real-time threat monitoring solution for researchers, students, and organizations.

**Keywords—** Cyber Threat Intelligence, Web Scraping, Threat Analysis, Severity Classification, Cybersecurity Automation, AI Chatbot, MySQL Database, Real-time Monitoring.

## PROBLEM STATEMENT

The rise of cyberattacks such as ransomware, phishing, data breaches, and zero-day vulnerabilities has created an urgent need for real-time cybersecurity monitoring systems. Threat intelligence enables organizations to understand emerging risks, anticipate attacks, and respond effectively. However, most existing Cyber Threat Intelligence (CTI) platforms are expensive, complex, or require expert-level knowledge, making them inaccessible to small organizations, students, or developing regions. Threat information is often scattered across blogs, news sites, and advisories in unstructured formats, making it difficult to manually monitor or analyze.

## Review & literature

Cyber Threat Intelligence (CTI) research highlights the importance of continuously collecting and analyzing threat information from diverse cybersecurity sources. Many studies emphasize automated scraping and RSS-based monitoring as reliable methods for obtaining real-time updates from news portals, security blogs, and advisories. These techniques reduce manual effort and help users stay aware of emerging incidents. Literature also shows that keyword-based classification is widely used in lightweight and open-source CTI systems because it is simple, transparent, and easy to maintain. It allows effective categorization of threats—such as ransomware, phishing, or vulnerabilities—without requiring complex algorithms.

Research on CTI dashboards stresses the need for clear visualization, searchable databases, and user-friendly interfaces to support fast decision-making. Recent works also note the growing use of external AI services to provide explanations and summaries, making cybersecurity information more accessible to non-technical users.

Overall, existing literature indicates a gap in simple, open-source CTI platforms designed for students, small organizations, and general users. The proposed system directly addresses this need by combining multi-source scraping, keyword-based analysis, structured storage, and an AI-assisted interface

## METHODOLOGY

The methodology of the proposed system follows a structured, multi-stage pipeline consisting of data collection, preprocessing, analysis, storage, retrieval, and AI-assisted interaction.

### 1. Data Collection and Scraping

Threat data is collected from cybersecurity news portals, RSS feeds, and blogs using tools like BeautifulSoup, feedparser, and Firecrawl. Each scraped item is converted into a structured format containing title, description, source, URL, publication date, and raw content.

### 2. Preprocessing and Content Extraction

Scraped content is cleaned to remove HTML tags, redundant text, and irrelevant information. Preprocessing ensures accurate keyword extraction and consistent analysis across different articles.

### 3. Threat Classification and Keyword Analysis

The system uses predefined keyword sets to identify threat categories such as ransomware, phishing, vulnerabilities, and data breaches. Severity is calculated based on keyword importance, while India-related terms are detected to highlight region-specific incidents. Output metadata includes category, severity score, extracted keywords, and India relevance.

#### 4. Database Storage and Structuring

All processed threats and analysis results are stored in a MySQL database. The schema includes tables for threats, keyword analysis, and system logs. Full-text indexing enables fast searching and efficient storage.

#### 5. Search and Filtering Mechanism

Users can retrieve threats using filters such as category, severity, and India-related relevance. Full-text search ensures quick access to relevant reports, supporting analysts and researchers in threat investigation.

#### 6. Dashboard and Visualization

A responsive interface displays threat cards, severity labels, keywords, and India-specific alerts. Statistical summaries and daily insights allow users to track emerging trends and patterns.

#### 7. AI-Driven Interaction Layer

An integrated Gemini-powered chatbot helps users interpret threats, generate summaries, and receive expert-level guidance. The AI uses stored threat context to provide accurate and relevant answers to user queries.

#### 8. Logging and Monitoring

System activities such as scraping events, user searches, and chat interactions are recorded in the System\_Log table, supporting debugging, auditing, and performance evaluation.

### USER AND SYSTEM WORKFLOW

The workflow of the proposed Cyber Threat Intelligence system describes how users interact with the platform and how the system processes data internally to deliver real-time threat intelligence. The workflow is divided into two parts: **User Workflow** and **System Workflow**, both operating together to ensure seamless performance.

#### 1. User Workflow

The user interacts with the platform primarily through a simple and intuitive dashboard.

The workflow includes:

- Accessing the Dashboard**  
The user opens the web interface to view the latest cyber threats, categorized by severity, source, and country relevance.
- Searching Threats**  
The user enters keywords or filters such as severity, category, or “India-only” to locate specific threats. Full-text search retrieves relevant results instantly.
- Viewing Threat Details**  
Clicking on a threat reveals its title, description, keywords, severity classification, and source link for further reading.
- Triggering Manual Scraping**  
The user may choose to manually initiate a scraping process to fetch the latest data from cybersecurity websites and RSS feeds.
- Interacting with the AI Assistant**  
For deeper understanding, the user can open the Chat Assistant, select a threat, and ask questions. The AI provides simplified explanations and impact assessments.
- Reading Statistics and Analytics**  
Users can view summarized metrics such as total threats, severity distribution, India-related threats, and recent trends.

#### 2. System Workflow

The system operates multiple automated background processes to ensure continuous, real-time data flow and analysis:

- Automatic Scraping Process**  
At predefined intervals, the system scrapes data from supported cybersecurity sources using RSS feeds, BeautifulSoup, and Firecrawl.
- Raw Data Extraction**  
The scraped content is cleaned, structured, and prepared for analysis by removing HTML tags, unnecessary text, and noise.
- Keyword and Severity Analysis**  
The analyzer identifies relevant cybersecurity keywords and determines the threat category (e.g., ransomware, phishing, vulnerability). Severity is calculated using predefined scoring rules.
- India-Related Detection**  
The system checks for terms such as “India,” “CERT-In,” or major city names to classify threats relevant to Indian cyberspace.
- Database Insertion**  
The processed threat is stored in the MySQL database with its associated attributes including category, severity, keywords, raw content, and timestamp.
- Logging System Events**  
Every scraping action, error, user access, or chat interaction is stored in the System\_Log table for auditing and monitoring.
- Search and Retrieval Engine**  
When a user performs a search or applies filters, the system retrieves results using full-text indexing for rapid response.
- AI Response Generation**  
During user-AI interaction, the system fetches the selected threat’s context and sends it to the Gemini model via the CodeWords API, returning a concise answer.

This internal workflow ensures the platform functions continuously, updates threats automatically, and responds intelligently to user actions..

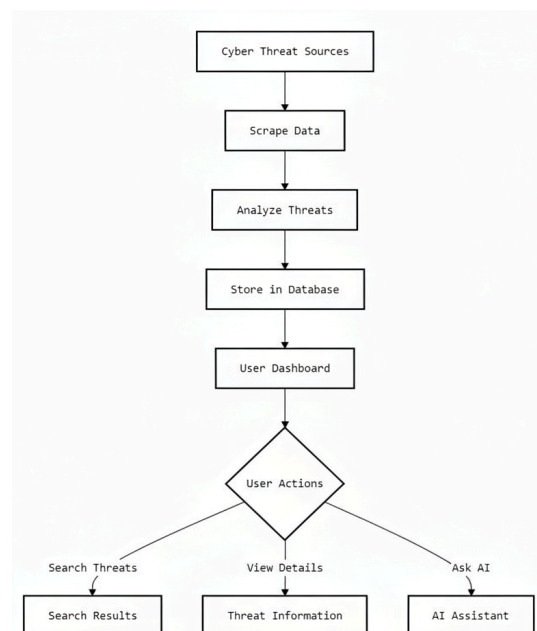


Fig. 1. User and System Workflow

## **BENEFITS**

The proposed Cyber Threat Intelligence platform provides several significant benefits that collectively enhance cybersecurity awareness, operational efficiency, and informed decision-making across various user groups. By automating the collection, analysis, and presentation of threat data, the system removes the dependency on manual monitoring and ensures that users receive timely and accurate intelligence from multiple reputable cybersecurity sources. These benefits extend to a wide range of stakeholders—including students, researchers, small and medium-sized organizations, Security Operations Center (SOC) teams, educators, and the general public—by transforming complex cyber threat information into clear, structured, and actionable insights.

For students and academic institutions, the platform serves as a valuable learning tool that introduces real-world threat patterns, attack techniques, and emerging vulnerabilities in an understandable manner. Researchers benefit from the system's structured database, keyword extraction, and severity metrics, which support advanced studies and experimentation in fields such as threat detection, machine learning, and cyber behavior analysis.

### **BETTER USE OF TIME**

One of the most significant advantages of the system is its ability to drastically reduce the time spent on manual cyber threat monitoring. Traditional methods require users to visit multiple cybersecurity websites, blogs, and advisories, manually extract information, and interpret details for relevance. The automated scraping engine eliminates this repetitive process by continuously collecting threat data in the background. Users are presented with a structured, categorized, and severity-based overview of the latest threats the moment they access the dashboard. This ensures that analysts, students, and organizations can focus on decision-making and response strategies instead of spending valuable time gathering data from scattered sources.

### **UNREFINED GROWTH**

The system's open-source architecture encourages continuous development and expansion. Since the platform is fully transparent and community-driven, new contributors can easily introduce additional threat feeds, integrate machine learning-based classifiers, or implement real-time anomaly detection algorithms. The modular design allows independent enhancement of scraping tools, analytics components, UI features, and AI capabilities. Over time, this fosters natural growth, enabling the platform to evolve alongside emerging cybersecurity trends. This adaptability ensures long-term sustainability and makes the system suitable for academic research, enterprise integration, and nationwide cybersecurity awareness initiatives.

### **TRACKING AND REPORTING**

The platform offers powerful tracking and reporting functionalities through a visually organized dashboard. It automatically generates summaries based on severity levels, threat categories, and India-specific relevance. Daily and weekly reports enable users to identify trends, understand attack patterns, and observe fluctuations in the cyber threat landscape. The inclusion of charts, metrics, and real-time statistics supports quick assessment of emerging threats and provides valuable insights useful for SOC teams, security researchers, and policy-makers. These reporting capabilities help organizations stay proactive rather than reactive by enabling faster risk identification and mitigation.

## **TECHNICAL LIMITATIONS**

The system relies heavily on web scraping, which is an inherently unstable and sensitive process. Even minor changes in website structures, HTML layouts, or content loading patterns can break scraping scripts and disrupt data collection. Websites may also implement anti-bot mechanisms such as CAPTCHAs, dynamic content loading, or IP rate limiting, all of which reduce scraping reliability. JavaScript-heavy pages often require more advanced tools like headless browsers to extract content accurately. Furthermore, scraping multiple sources continuously increases server load, bandwidth usage, and processing time, which may necessitate more robust infrastructure. To maintain consistent performance, the platform must be regularly updated with fallback mechanisms such as RSS feeds, APIs, cached snapshots, or adaptive scraping logic. These challenges make long-term maintenance technically demanding.

### **USER EXPERIENCE CHALLENGES**

Although the system attempts to simplify threat intelligence, many users—especially beginners—may still find cybersecurity terminology difficult to interpret. Concepts such as exploits, malware strains, CVE identifiers, or attack vectors often require foundational knowledge that non-technical users may not possess. If the interface does not provide clear explanations, tooltips, or intuitive navigation, users may feel overwhelmed. Ensuring accessibility requires thoughtful UI/UX design, minimal cognitive load, and AI-generated guidance to help users understand technical content. Additionally, users with varying levels of digital literacy may interact differently with the system, making it essential to design a flexible and inclusive interface. Failure to address these issues could hinder widespread adoption.

### **Cost and Maintenance**

Although the platform is open-source, maintaining it over time involves financial and resource considerations. Continuous scraping requires server uptime, network bandwidth, storage expansion, and periodic hardware upgrades. AI-based features such as chatbot interactions may incur costs related to API usage or compute time. Enterprises or large-scale deployments may require database replication, load balancing, and backup systems, all of which add to long-term operational costs. Moreover, maintenance includes regular updating of scraping scripts, patching vulnerabilities, monitoring logs, and conducting security audits. Without proper cost planning and sustainable resource management, the platform may face operational challenges, especially under high usage.

### **Cultural and Behavioral Barriers**

Organizations, especially small or traditional institutions, may resist adopting automated cybersecurity systems due to cultural norms or lack of awareness. Many teams rely on manual monitoring or outdated practices and may distrust automated tools, fearing false information or system-level errors. Employees may also lack adequate cybersecurity training, making it difficult for them to interpret the platform's outputs or integrate them into decision-making workflows. Overcoming these barriers requires training programs, awareness workshops, and demonstration of the system's accuracy and usefulness. Without organizational buy-in, even the most advanced tools may not achieve their desired impact.

## Security and Privacy Concerns

Handling cyber threat data introduces several privacy and security considerations. Logs, scraped content, URLs, metadata, and user interaction history must be stored securely to prevent unauthorized access. Weak security configurations can expose sensitive data or allow attackers to manipulate system behavior. Improper access controls, unencrypted connections, or insecure database practices may result in data leakage. Additionally, scraping malicious websites may unintentionally retrieve harmful payloads or infected scripts, requiring strict sanitization procedures. The platform therefore must employ encryption, firewall policies, secure authentication for administrators, and robust data validation to ensure safety and integrity.

## Domain-Specific Challenges

Cybersecurity is a rapidly evolving field in which new vulnerabilities, attack vectors, and threat actors appear frequently. Static keyword lists and predefined categories may not accurately capture emerging threats, leading to misclassification or missed detections. New malware strains, exploit kits, and social engineering patterns require continuous updates to the system's analysis models. The platform must stay aligned with global cybersecurity standards, advisories, and threat databases such as MITRE ATT&CK, CVE lists, and CERT reports. This demands ongoing research, expert involvement, and regular refinement of classification rules to maintain high accuracy and relevance.

## Analytics and Feedback Issues

As the volume of stored threats increases, analytics generation becomes more computationally intensive. Generating charts, severity trends, category distributions, and India-specific statistics may slow down if the database is not properly indexed or optimized. Inefficient storage practices or outdated data management strategies can disrupt dashboard performance and degrade user experience. Additionally, feedback loops—user corrections, error reporting, or classification refinement—must be incorporated effectively to improve system accuracy.

Periodic data cleaning, caching strategies, and indexing enhancements are essential to maintain responsiveness and ensure that analytics remain meaningful and accurate over time.

## CONCLUSION

The development of the proposed Cyber Threat Intelligence (CTI) platform demonstrates a practical and effective approach to addressing the growing complexity of cyber threats in today's digital environment. By integrating automated scraping, keyword-based threat classification, structured data storage, and AI-driven analysis, the system provides a comprehensive, real-time view of the global and Indian cyber threat landscape. The platform successfully transforms unstructured and scattered threat information into actionable intelligence that can be easily interpreted by users across different levels of technical expertise, ranging from students and researchers to security analysts and organizational decision-makers.

The platform's open-source nature ensures accessibility and encourages long-term expansion through community-driven development. Its modular architecture allows the integration of additional threat sources, advanced analytics, and machine learning models in the future. The AI-assisted chatbot further enhances usability by simplifying complex cybersecurity concepts, helping users grasp the implications of various threat events without prior domain knowledge.

Despite the challenges identified—such as scraping limitations, evolving cybersecurity terminology, and maintenance overhead—the system establishes a strong foundation for continuous improvement. Future enhancements may include automated incident correlation, behavior-based threat detection, predictive analytics, or integration with national cybersecurity frameworks like CERT-In advisories and MITRE ATT&CK mapping. With ongoing refinement, the platform has the potential to evolve into a robust, community-supported CTI ecosystem that contributes significantly to cybersecurity readiness, awareness, and research.

In conclusion, the proposed CTI platform not only addresses the limitations of manual monitoring and expensive commercial systems but also democratizes access to threat intelligence, providing a valuable resource for individuals, academic institutions, and organizations striving to strengthen their cybersecurity posture.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to all individuals and organizations who contributed to the successful completion of this research work. Special thanks are extended to the faculty members and academic mentors whose guidance, constructive feedback, and continuous encouragement played a crucial role in shaping the direction of this project. Their expertise and insights greatly enhanced the quality of the proposed system and its implementation.

We also acknowledge the valuable contributions of the open-source developer community, whose tools, frameworks, and shared knowledge made the development of the Cyber Threat Intelligence platform possible. Platforms such as Firecrawl, BeautifulSoup, feedparser, and various cybersecurity data sources offered essential support in building an automated and reliable threat analysis environment.

The authors further express appreciation to the researchers, cybersecurity professionals, and organizations such as CERT-In, MITRE, and leading cybersecurity news portals for their publicly available threat intelligence resources. These sources provided essential data that enriched the analysis and evaluation of the system.

Lastly, we extend gratitude to classmates, peers, and family members for their encouragement, patience, and motivation throughout the research and development process. Their support was instrumental in overcoming challenges and achieving the goals of this project.

## REFERENCES

- [1] **Wagner, C., Dulaunoy, A., Wagener, G., & Iklody, A.** "Cyber Threat Intelligence Sharing: Survey and Challenges," 2019.  
*Relevance:* Provides foundational background for CTI architectures and information-sharing challenges.
- [2] **Sensors Journal.** "Systematic Review of Cyber Threat Intelligence Technologies, Machine Learning Models, and Future Trends," *Sensors*, 2025.  
*Relevance:* Supports modern approaches and future-scope planning.
- [3] **Arnold, M., et al.** "A Cyber Threat Intelligence Tool for Analyzing Threats from Dark-Web Sources," in *Proc. IEEE ISI*, 2019.  
*Relevance:* Demonstrates multi-source scraping similar to your platform.
- [4] **Wagner, C. et al.** "The Design and Implementation of MISP: An Open-Source Threat Intelligence Sharing Platform," 2016.  
*Relevance:* Key reference for understanding open-source CTI platforms.

[5] **Alzahrani, A., et al.** “Enhancing Cyber Threat Intelligence Using IoC Sharing and MISP Integration,” 2024.

*Relevance:* Helps justify India-specific IoC and threat tracking.

[6] **SLR: Text Mining in Cybersecurity.** “A Systematic Review on Applying NLP for Cybersecurity Insight Extraction,” *Year not specified.*

*Relevance:* Direct support for keyword-analysis and NLP modules.

[7] **Automating CTI with NLP.** “NLP Pipelines for Automatic Extraction of Indicators from Threat Reports,” 2023.

*Relevance:* Supports your automated keyword extraction and IoC parsing logic.

[8] **NLP for Threat Detection (2024).** “Using GRU Models and Topic Modeling for Cyber Threat Classification,” 2024.

*Relevance:* Helpful for ML-based severity prediction.

[9] **NLP-Based CTI Framework (2025).** “An Integrated Framework for Threat Extraction and Classification Using NLP Models,” 2025.

*Relevance:* Matches your planned smart ML enhancements.

[10] **MISP Collection Study (2025).** “Automated CTI Ingestion and Sanitization Using the MISP Framework,” 2025.

*Relevance:* Supports database-level ingestion, cleaning, and incident-source refinement.