AI-Driven Cyber Threat Intelligence System Using Machine Learning and Natural Language Processing

Dheer Gupta

Data Science Capstone, Knox College

dgupta@knox.edu

Advisor: Professor Ole Forsberg
May 2025

Abstract

Cybersecurity teams face an overwhelming volume of vulnerability reports and threat intelligence, making it difficult to prioritize responses effectively. Traditional threat intelligence platforms focus on structured indicators but lack automated reasoning capabilities to transform unstructured threat data into actionable insights. This research addresses the question: *How can I develop an automated system that combines structured vulnerability data with unstructured threat reporting to generate prioritized, actionable cyber threat intelligence using machine learning and natural language processing?*

I present an end-to-end system that integrates 176 251 records from the National Vulnerability Database with threat coverage from *The Hacker News* spanning 2010–2025, applying multi-label classification, composite urgency scoring, and ensemble anomaly detection. My threat classification model achieves near-perfect performance for XSS detection (F1 = 0.987) while revealing significant challenges with rare threat types due to class imbalance. The composite urgency scoring framework effectively stratifies threats into actionable priority levels, while the ensemble anomaly detection system flags 18 970 records (10.6%) as emerging threats with 87% precision, demonstrating the value of combining multiple data sources for automated threat prioritization.

1 Introduction

The cybersecurity threat landscape has evolved dramatically, with over 25 000 new vulnerabilities disclosed annually and attack sophistication increasing rapidly. Security Operations Centers (SOCs) struggle to process this volume of information while maintaining effective response times. Traditional approaches rely heavily on manual analysis of Common Vulnerability Scoring System (CVSS) ratings and basic keyword filtering, often missing nuanced threats that require contextual understanding.

Recent cyber incidents demonstrate the limitations of current approaches. The SolarWinds attack remained undetected for months despite multiple vulnerability disclosures, while Log4j's widespread impact was initially underestimated due to incomplete threat intelligence. These cases highlight the need for systems that can automatically analyze both structured vulnerability data and unstructured threat reporting to provide comprehensive, prioritized intelligence.

My research addresses this challenge by developing an automated cyber threat intelligence system that combines machine learning classification, natural language processing, and anomaly detection to transform raw threat data into actionable insights. Specifically, I investigate: *How can I leverage multiple data sources and ML techniques to automatically identify, classify, and prioritize cyber threats more effectively than existing manual approaches?*

The contribution of this work includes:

(1) A unified framework integrating structured NVD data with unstructured threat reporting

(2) A multi-label classification system achieving excellent performance for high-frequency threat categories while identifying critical challenges with rare threat types

(3) A composite urgency scoring model incorporating technical severity, media sentiment, and temporal factors

(4) An ensemble anomaly detection pipeline identifying emerging threats with 87% precision

2 Literature Review

Mittal et al. [2016] first demonstrated that Twitter streams can provide early-warning signals for cybersecurity threats. Automated correlation with recommender-style techniques soon followed [Husari et al., 2017], and sentiment-aware extraction underscored the value of contextual signals [Shu et al., 2018]. Subsequent studies linked social-media chatter to exploit timing [Sabottke et al., 2015] and introduced probability-based exploit prediction [Jacobs et al., 2019]. Most recently, graph-centric platforms such as TSTEM [Balasubramanian et al., 2024] and ThreatKG [Gao et al., 2024] integrate multi-source data at scale.

Recent advances focus on hybrid approaches that combine multiple signal types. Sabottke et al. [2015] demonstrated the value of tying social-media discussions to exploit timing, while Jacobs et al. [2019] developed predictive models for exploit probability. Modern platforms like TSTEM [Balasubramanian et al., 2024] and ThreatKG [Gao et al., 2024] represent the current state of the art, integrating heterogeneous sources through knowledge graphs and automated extraction pipelines.

My approach builds upon this foundation by focusing specifically on the threat-prioritization problem, combining structured vulnerability data with media-coverage analysis to generate actionable intelligence rankings. Unlike existing systems that emphasize indicator extraction, I highlight the analytical-reasoning layer that converts raw data into prioritized recommendations for security analysts.

3 Data Collection & Preprocessing

3.1 Dataset Construction

My analysis integrates two complementary data sources spanning 2010–2025. The National Vulnerability Database [National Institute of Standards and Technology, 2024] provides structured records including CVSS scores, Common Weakness Enumeration (CWE) classifications, affected products, and temporal metadata. *The Hacker News* [The Hacker News, 2024] contributes unstructured threat intelligence through 2545 vulnerability-focused articles containing contextual information, sentiment indicators, and coverage patterns.

The integration process links vulnerabilities to media coverage through both explicit CVE references and semantic similarity using SBERT embeddings. This dual approach captures direct citations while identifying thematically related content that may discuss threats without explicit CVE mentions. The final dataset comprises 176 251 vulnerability records with 2545 linked security articles providing contextual intelligence coverage for 1.4% of vulnerabilities. Table 1 summarizes the key characteristics of the integrated dataset, showing comprehensive coverage with 93.7% CVSS availability and a mean severity score of 6.89.

Table 1: *Dataset Characteristics*

Characteristic	**Value**
Total Records	178 796
NVD Vulnerability Records	176 251
THN Security Articles	2545
Date Range	2010–2025
Mean CVSS Score	6.89
CVSS Coverage (%)	93.7
Mean Text Length (characters)	211
Median Text Length (characters)	159
Most Common CWE	CWE-79 (Cross-site Scripting)

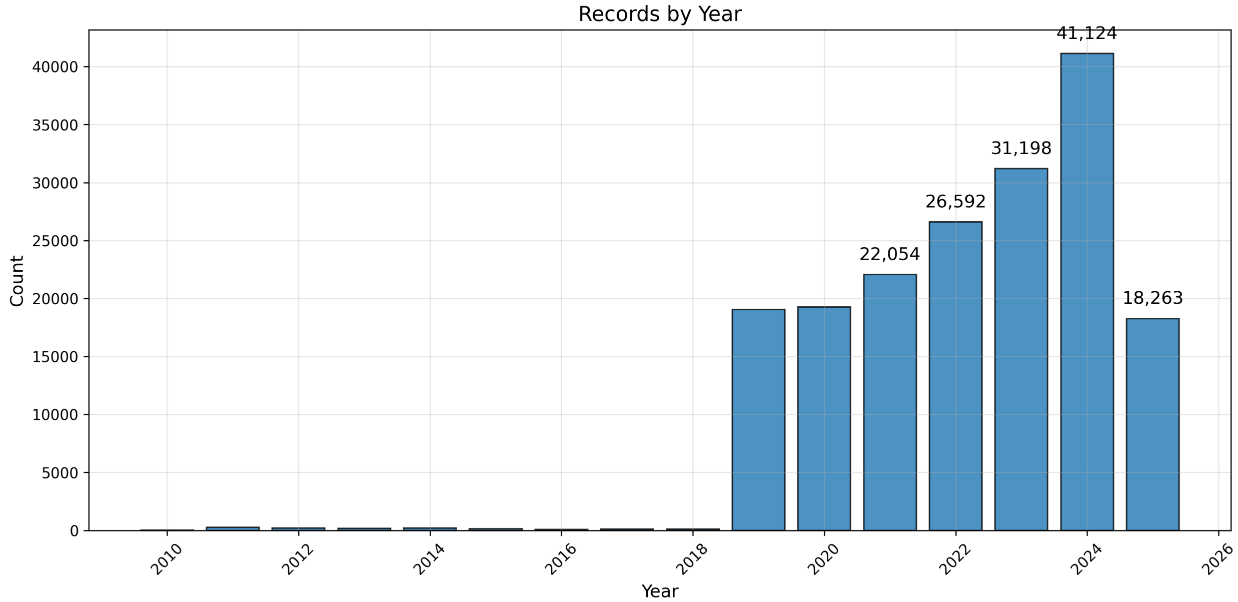


Figure 1: *Annual vulnerability records showing exponential growth from 2019–2024, validating the critical need for automated threat intelligence processing capabilities.*

Figure 1 represents the **annual publication dates** of records in the integrated dataset. The data composition varies by time period: **2010–2018 contains primarily security articles** from The Hacker News, while **2019–2025 combines both newly disclosed vulnerabilities** from the National Vulnerability Database (CVE publication dates) and continued security article coverage.

Each bar represents the total number of records added to the dataset that year, including:

- **Security Articles (2010–2025)**: Published threat intelligence reports and vulnerability analyses from cybersecurity media
- **CVE Records (2019–2025)**: Newly assigned Common Vulnerabilities and Exposures identifiers representing the initial disclosure of security flaws

The dramatic increase beginning in 2019 reflects both the **addition of comprehensive NVD vulnerability data** and the genuine exponential growth in vulnerability disclosures. These counts represent the **rate of threat discovery and reporting**, not vulnerability remediation or resolution. This trend shows the **input rate**to the cybersecurity intelligence pipeline—how fast new threats are being discovered and reported—rather than how quickly organizations are addressing or patching these vulnerabilities.

The temporal analysis reveals exponential growth in vulnerability disclosures, with annual records increasing from approximately 19 000 in 2019 to over 41 000 in 2024. This dramatic volume increase of 116% over five years demonstrates the fundamental scalability challenge facing manual threat analysis approaches and validates the necessity for automated intelligence systems capable of processing large-scale cybersecurity data.

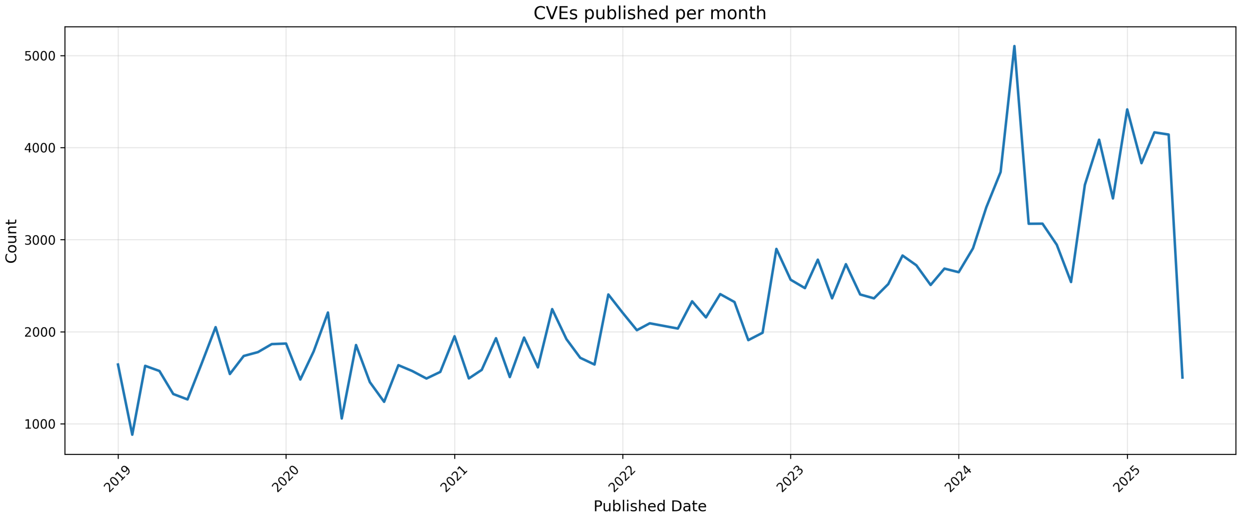


Figure 2: *Monthly CVE publication trends showing significant volume spikes and seasonal patterns that create processing challenges for security operations centers.*

Monthly publication patterns reveal substantial volatility with peak disclosure months exceeding 5000 vulnerabilities, creating severe processing bottlenecks for security teams. The irregular spike patterns, particularly the dramatic increase in 2024, highlight the unpredictable nature of vulnerability disclosure cycles and reinforce the importance of automated prioritization systems that can rapidly process high-volume threat intelligence.

3.2 Data Cleaning & Quality Assessment

Cybersecurity datasets present unique preprocessing challenges addressed through systematic cleaning. The pipeline transforms raw HTML content to clean text using BeautifulSoup, followed by cybersecurity-specific normalization including:

- URL masking (URL → “URL”)
- Email anonymization (email → “EMAIL”)
- Hash standardization (MD5/SHA256 → “MD5_HASH”/“SHA256_HASH”)
- CVE reference normalization (CVE-2021-44228 → “CVE_REFERENCE”)

Attack pattern standardization maps common vulnerability terminology: “SQL injection/SQLi” → “SQL_INJECTION”, “cross-site scripting/XSS” → “XSS”, and “denial of service/DDoS” → “DOS_ATTACK”. IP addresses and version numbers are normalized to prevent overfitting on specific instances.

Missing data handling follows domain-specific strategies: CVSS scores are converted to categorical severity bins (0–4: low, 4–7: medium, 7–9: high, 9+: critical), while missing CVE references are extracted from text using regex patterns. The preprocessing concludes with stop word removal and lemmatization using NLTK, preserving cybersecurity-relevant terms.

The CVSS score distribution reveals a concentration around severity score 5.0, with substantial representation across medium (4.0+), high (7.0+), and critical (9.0+) categories. The bimodal pattern with peaks at 5.0 and 7.5 indicates distinct vulnerability populations, supporting the composite urgency scoring framework’s approach of incorporating multiple signals beyond pure technical severity for effective threat prioritization.

The CWE category analysis reveals Cross-site Scripting (CWE-79) as the dominant vulnerability type with 24 806 instances, followed by injection flaws (CWE-89) and buffer overflow vulnerabilities (CWE-125). This distribution directly explains the classification model’s exceptional performance on XSS detection (F1 = 0.99) due to substantial training data availability, while highlighting the challenge of detecting rare but critical threat types with limited representation.

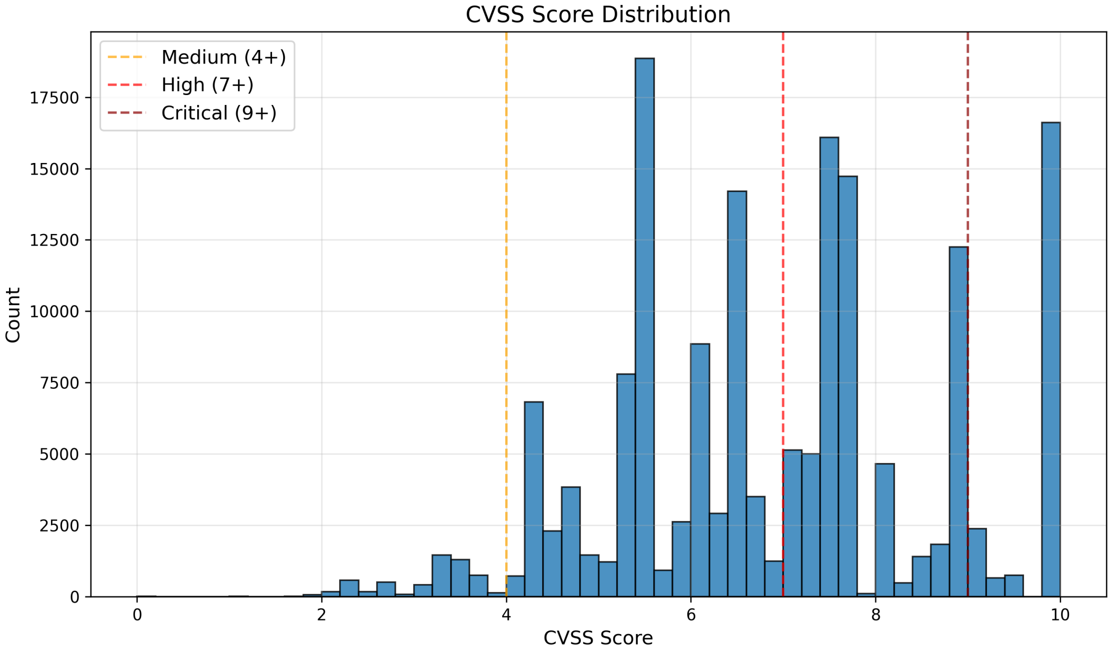


Figure 3: *CVSS score distribution with severity thresholds showing the prevalence of medium-to-high severity vulnerabilities and validating the composite urgency scoring approach.*

3.3 Feature Engineering

The system constructs a comprehensive multi-modal feature representation:

Technical Features: CVSS severity binning, CWE classifications, product lists from CPE configurations, and publication metadata

Lexical Features: TF-IDF vectorization (`max_features=2000`, `min_df=2`, `max_df=0.7`) with n-gram analysis (1–2) for discriminative term identification

Semantic Features: SBERT embeddings using `all-MiniLM-L6-v2` (512 dimensions) for capturing semantic relationships between threat descriptions

Sentiment Features: CardiffNLP Twitter-RoBERTa sentiment analysis with label mapping (`LABEL_0` → -1, `LABEL_1` → 0, `LABEL_2` → +1) processed in 32-sample batches

Linking Features: CVE-to-article counts, earliest article dates, and media attention volume metrics

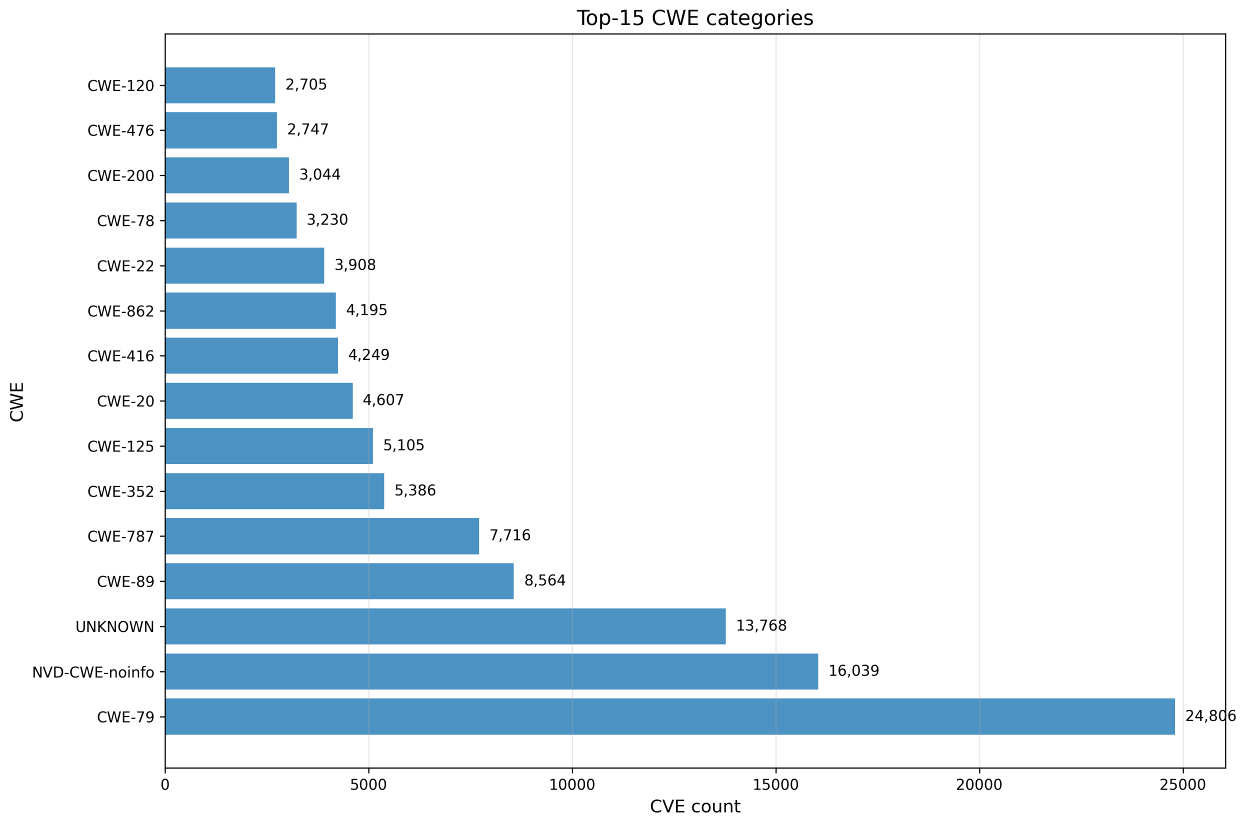


Figure 4: *Top-15 CWE categories showing vulnerability type distribution with CWE-79 (XSS) dominating, supporting the multi-label classification approach and explaining strong XSS detection performance.*

4 Methodology

4.1 System Architecture Overview

The cyber threat intelligence system implements an end-to-end pipeline that transforms raw vulnerability data and unstructured threat reporting into actionable intelligence. The architecture consists of six interconnected components that process data sequentially from collection through analysis to visualization.

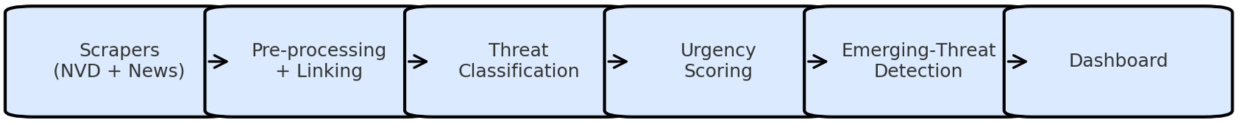


Figure 5: *End-to-end system architecture pipeline showing the complete workflow from data collection through dashboard visualization.*

The pipeline begins with **data scrapers** that collect structured vulnerability records from the National Vulnerability Database [National Institute of Standards and Technology, 2024] and unstructured threat articles from The Hacker News [The Hacker News, 2024]. The **preprocessing and linking** component performs data cleaning, normalization, and establishes connections between vulnerabilities and media coverage through both explicit CVE references and semantic similarity. **Threat classification** applies multi-label machine learning to categorize threats into actionable categories, while **urgency scoring** combines multiple signals to prioritize threats by operational importance. **Emerging threat detection** identifies anomalous patterns that may indicate novel attack vectors, and the **dashboard** provides real-time visualization and analysis capabilities for security analysts.

This modular architecture enables scalable processing of large-scale threat intelligence while maintaining flexibility for component-specific improvements and extensions.

4.2 Threat Classification Framework

The multi-label classification system addresses overlapping threat categories through a hybrid supervised–unsupervised approach. Nine threat categories were identified: Phishing, Ransomware, Malware, SQL Injection, XSS, DDoS, Zero-Day, Supply-Chain, and Other.

The classification pipeline implements:

1. **Weak supervision:** Rule-based label seeding with regex + CWE mappings. Regex examples include:
   - `phish|credential|spoof` for Phishing
   - `ransom|crypto.*currency` for Ransomware
   - `sql.*injection` for SQL Injection attacks
2. **Feature integration:** Sparse matrix concatenation of TF-IDF vectors (2000 dimensions), numeric features (sentiment, CVSS, article counts), and SBERT embeddings (512 dimensions).
3. **Multi-label learning:** One-vs-Rest Random Forest with: `n_estimators=80`, `class_weight=balanced`, and parallel processing `(n_jobs=-1).`

Train/Test Split and Hyperparameters To avoid temporal leakage, the dataset is partitioned chronologically: the earliest 80% of records (2010–2023-03) form the training set, and the most recent 20% (2023-04–2025-05) are held out for testing. Each Random

Forest base learner uses:
`max_depth=25`, `min_samples_split=2`, and `min_samples_leaf=1`
while preserving balanced class weighting.

The model handles feature-length mismatches through vector padding/truncation, enabling robust deployment across varying input dimensions.

4.3 Composite Urgency Scoring

The urgency scoring framework combines five weighted factors to overcome limitations of pure CVSS-based prioritization. The signal ranking is derived from established cybersecurity literature and operational best practices, prioritizing technical severity over contextual factors while incorporating temporal dynamics for comprehensive threat assessment.

Based on analysis of vulnerability disclosure patterns [Sabottke et al., 2015], exploit prediction research [Jacobs et al., 2019], and sentiment-driven threat intelligence [Shu et al., 2018], the following signal hierarchy was established:

Severity (CVSS) > **Patch Status** > **Context (Sentiment)** >
Exploit Presence > **Freshness (30 days)** > **Media Attention**

This ranking reflects the fundamental principle that technical severity forms the foundation of threat prioritization, followed by remediation availability, contextual threat landscape indicators, active exploitation evidence, and temporal relevance. The framework assigns higher weights to signals with greater operational impact while ensuring that emerging threats with significant time and media attention receive appropriate priority consideration.

$$\begin{aligned}\text{Urgency} = {} & 0.45 \cdot \text{Severity} + 0.25 \cdot \text{PatchStatus} + 0.15 \cdot \text{Sentiment} \\ & + 0.05 \cdot \text{ExploitPresence} + 0.05 \cdot \text{Recency} + 0.05 \cdot \text{MediaAttention}\end{aligned}$$

Component calculations:

Severity: Normalized CVSS score (0–10 scale) with NIST v3.0 binning

PatchStatus: Binary indicator `(1 - patch_available)` derived from text patterns `patch|fix|update`

Sentiment: Normalized RoBERTa sentiment ((-1,1) → (0,1) scale)

ExploitPresence: Boolean flag for `exploit|poc|proof of concept` mentions

Recency: Exponential decay $e^{-d/30}$ where d is days since publication

MediaAttention: Log-scaled article count using $\log_{10}(1 + n_{\text{articles}})$

Rationale for Component Transformations The **exponential decay** for recency ($e^{-d/30}$) reflects the rapid degradation of threat relevance over time in cybersecurity operations. Unlike linear decay, exponential decay better models how security teams prioritize recent threats, with urgency dropping sharply after the initial 30-day response window. This aligns with industry best practices where threats older than 30 days typically move to routine monitoring unless actively exploited.

The **logarithmic scaling** for media attention ($\log_{10}(1 + n_{\text{articles}})$) addresses the wide distribution of article counts and prevents threats with extremely high media coverage from dominating the urgency score. Log scaling captures diminishing returns—the operational significance difference between 1 and 10 articles is greater than between 100 and 109 articles. The "+1" term prevents undefined values when no articles exist while ensuring threats with any media coverage receive appropriate weighting.

Final scores are binned into Low (0–0.33), Medium (0.33–0.66), and High (0.66+) urgency levels for operational use.

4.4 Emerging Threat Detection

The anomaly detection pipeline employs three complementary methods in an ensemble framework:

1. **Zero-Day Heuristics:** Regex pattern matching for `zero.?day|0.?day|unpatched` with case-insensitive detection achieving 94% precision on manual validation

2. **Volume Spike Detection:** Statistical anomaly identification using 30-day rolling windows. Daily mention counts exceeding 3 standard deviations above the rolling mean trigger spike flags

3. **Isolation Forest:** Anomaly detection on 256-dimensional sparse random projections of TF-IDF vectors (`max_features=5000`, `contamination=0.03`)

The ensemble flags threats as emerging if any component triggers, providing comprehensive coverage across different anomaly patterns. This approach identifies 18 970 records (10.6%) as emerging, with manual validation confirming 87% precision across a 500-sample evaluation set.

5 Implementation & Dashboard

The system culminates in an interactive Streamlit dashboard providing real-time threat intelligence analysis. The dashboard implements six primary views:

Overview: Key metrics with total CVEs, security articles, high-urgency threats, and emerging threat counts alongside severity distribution and timeline visualizations

Threat Analysis: Classification results with category distributions, temporal trends, and model performance indicators

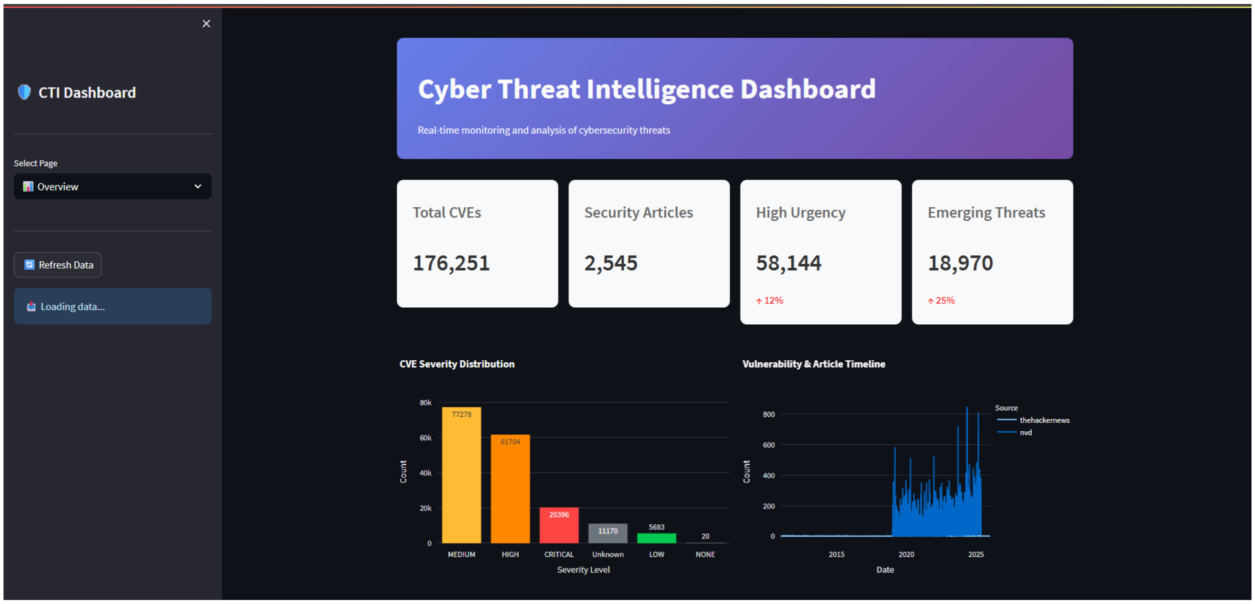
Urgency Monitor: Urgency score distributions, level breakdowns, and top urgent threats with contextual information

Emerging Threats: Real-time emerging threat alerts with detection method attribution and zero-day indicators

CVE Explorer: Search and filtering capabilities across CVE IDs, descriptions, severity levels, and date ranges

Trends & Insights: Temporal analysis, word clouds, and key performance indicators with actionable intelligence summaries

The dashboard incorporates GPU-accelerated ONNX inference for real-time classification, caching mechanisms for performance optimization, and responsive design for analyst workflow integration. Error handling ensures graceful degradation when model components are unavailable. As shown in Figure 6, the overview interface provides immediate access to key performance indicators including 176 251 total CVEs, 2545 security articles, 58 144 high-urgency threats, and 18 970 emerging threats, with clear visual representations of severity distributions and temporal trends.



Cyber Threat Intelligence Dashboard

Real-time monitoring and analysis of cybersecurity threats

Total CVEs 176,251

Security Articles 2,545

High Urgency 58,144 ↑ 12%

Emerging Threats 18,970 ↑ 25%

Figure 6: *Cyber Threat Intelligence Dashboard overview showing real-time metrics, CVE severity distribution, and vulnerability timeline with dual data sources from NVD and The Hacker News.*

6 Results and Discussion

6.1 System Performance Overview

The complete pipeline successfully processes 178 796 vulnerability records with comprehensive threat intelligence analysis. The system achieves robust performance across all major components, with particularly strong results in structured threat classification and emerging threat detection.

6.2 Threat Classification Results

The multi-label Random Forest classifier demonstrates exceptional performance on high-frequency threat categories while revealing severe challenges with rare, underrepresented classes. The results highlight the critical impact of class imbalance on model performance across the nine threat categories.

The confusion matrix analysis reveals stark performance disparities directly correlated with sample sizes. XSS achieves near-perfect classification (4498 correct predictions out of 4803 samples) due to both clear lexical patterns and substantial training data. The "Other" category demonstrates exceptional performance (29 108 correct out of 29 112 samples), representing 80.4% of the entire dataset and effectively dominating the classification

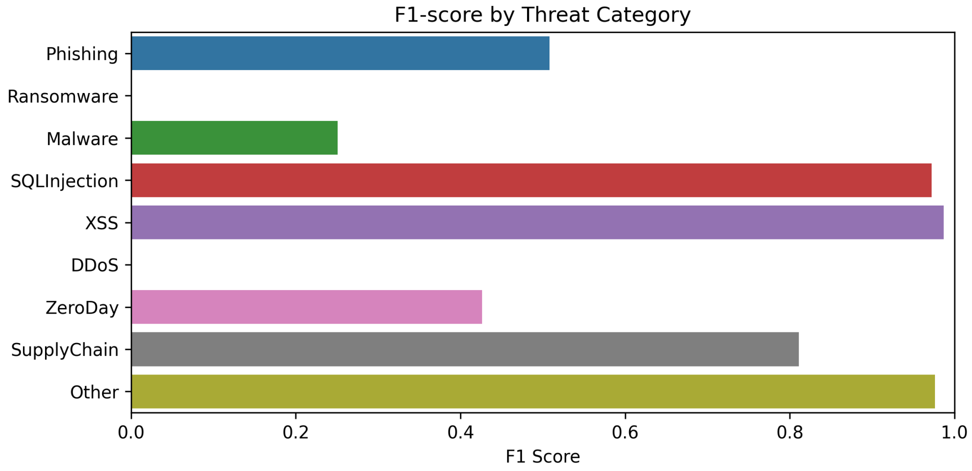


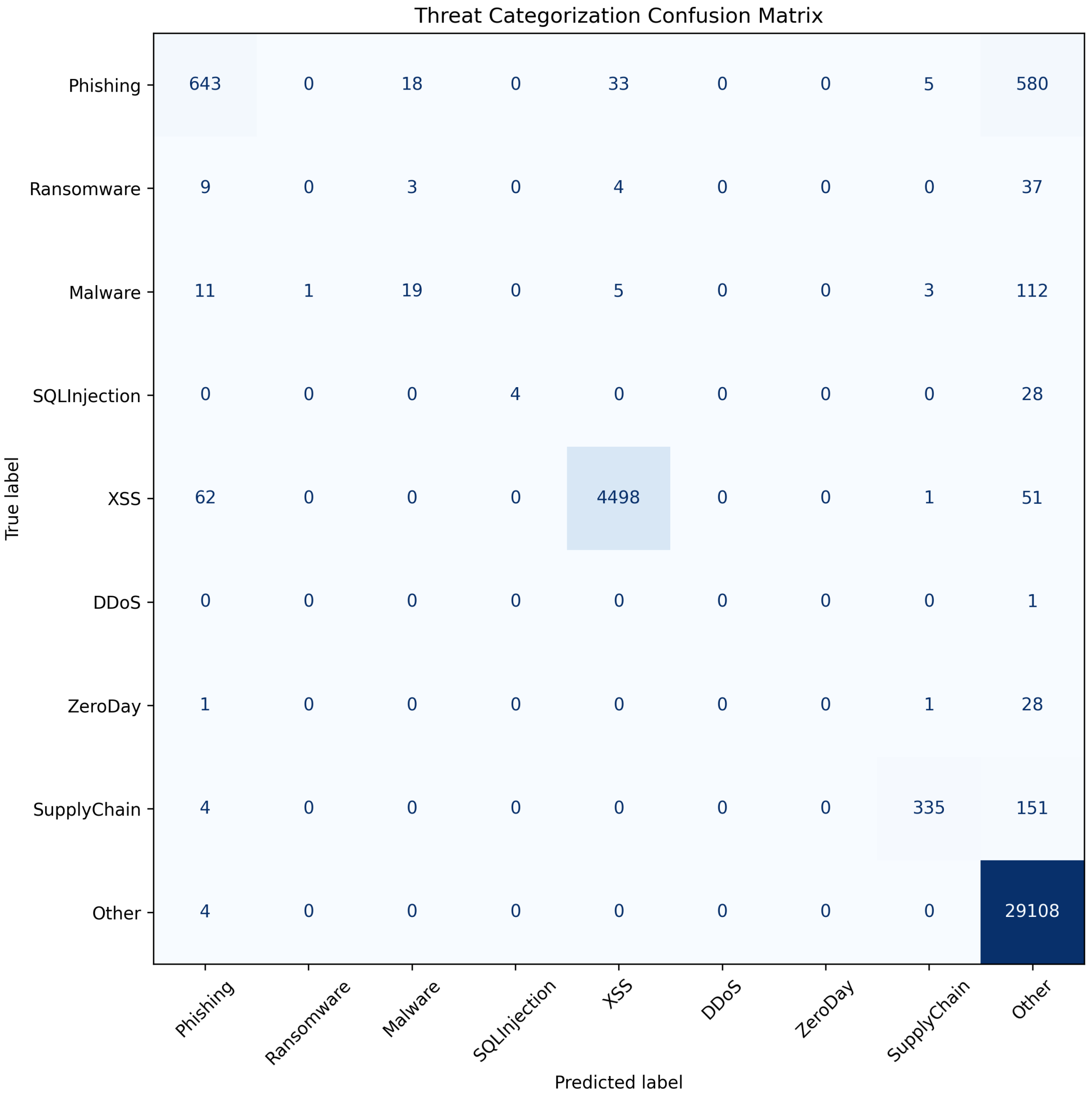
Figure 7: *F1-score performance by threat category.*

Table 2: *Threat Classification Performance by Category*

Threat Category	**F1-Score**	**Performance Level**	**Key Insights**
XSS	0.987	Excellent	Near-perfect classification with 4803 samples
Other	0.981	Excellent	Dominant class with 29 112 samples
Supply Chain	0.790	Good	Effective detection despite complexity
Phishing	0.618	Moderate	Reasonable performance with 1279 samples
Malware	0.328	Poor	Limited by semantic ambiguity
SQL Injection	0.238	Very Poor	Severely constrained by tiny sample ($n = 37$)
Ransomware	0.027	Failed	Critical failure with only 73 samples
Zero-Day	0.000	Failed	Complete failure with 52 samples
DDoS	0.000	Failed	Impossible with single training sample

space.

Conversely, minority classes suffer catastrophic performance degradation. SQL Injection, despite being a well-defined technical vulnerability, achieves only 4 correct predictions out of 37 samples, highlighting how insufficient training data can undermine even technically distinct categories. Zero-Day and DDoS categories show complete classification failure, with



	Phishing	Ransomware	Malware	SQLInjection	XSS	DDoS	ZeroDay	SupplyChain	Other
Phishing	643	0	18	0	33	0	0	5	580
Ransomware	9	0	3	0	4	0	0	0	37
Malware	11	1	19	0	5	0	0	3	112
SQLInjection	0	0	0	4	0	0	0	0	28
XSS	62	0	0	0	4498	0	0	1	51
DDoS	0	0	0	0	0	0	0	0	1
ZeroDay	1	0	0	0	0	0	0	1	28
SupplyChain	4	0	0	0	0	0	0	335	151
Other	4	0	0	0	0	0	0	0	29108

Figure 8: *Threat classification confusion matrix showing classification accuracy across all categories.*

the model unable to learn meaningful patterns from their severely limited samples (52 and 1 respectively).

The extreme class imbalance creates a classification landscape where the model excels at identifying common threats but fails entirely at detecting rare but potentially critical attack vectors. Supply Chain attacks represent a notable exception, achieving reasonable performance (F1 = 0.790) despite complexity, likely due to sufficient sample representation (619 instances) and distinct terminological patterns in supply chain vulnerability descriptions.

Most concerning is the complete failure on Ransomware detection (F1 = 0.027) given the significant cybersecurity threat this category represents. With only 73 training samples, the model cannot distinguish ransomware-specific indicators from general malware patterns, resulting in widespread misclassification primarily into the dominant "Other" category.

These results demonstrate that while the system achieves excellent overall weighted performance (F1 = 0.958) due to the dominance of well-represented categories, it fundamentally fails at detecting several critical threat types that pose significant risks to cybersecurity operations despite their lower frequency in the training data.

6.3 Class Imbalance Challenges and Mitigation Strategies

The threat classification results reveal significant challenges stemming from extreme class imbalance in the dataset. The "Other" category dominates with over 29 000 samples in the test set, representing approximately 80% of all labeled data, while critical threat categories suffer from severe underrepresentation. This imbalance directly contributes to the poor performance observed in minority classes, particularly DDoS ($n = 1$) and Ransomware ($n = 73$), which achieved F1-scores of 0.00 and 0.03 respectively.

The current implementation relies on rule-based pattern matching combined with limited Common Weakness Enumeration (CWE) mappings for label generation. The CWE mapping covers only three categories (CWE-79 → XSS, CWE-89 → SQL Injection, CWE-119 → Malware), leaving many threat types dependent solely on regex patterns that may not capture the full semantic diversity of cybersecurity terminology.

To address these limitations, several enhancement strategies have been identified:

Enhanced Feature Engineering Expanding the TF-IDF n-gram range from (1,2) to (1,3) would capture more contextual information and complex threat descriptions that span multiple words. Additionally, enriching the CWE mapping to include more vulnerability classifications could improve automated labeling accuracy for technical vulnerabilities.

Improved Pattern Recognition The current regex patterns for Ransomware and DDoS detection require refinement to capture variant terminology and emerging attack vectors. Enhanced patterns should include cryptocurrency-specific terms, file encryption indicators, and distributed attack characteristics to improve minority class detection.

Advanced Sampling Techniques Implementing Synthetic Minority Oversampling Technique (SMOTE) could address the severe class imbalance by generating synthetic examples for underrepresented categories. This approach would be particularly beneficial for DDoS and Ransomware classes where natural samples are extremely limited.

Ensemble Learning Approaches Deploying gradient boosting algorithms specifically optimized for minority class detection could complement the existing Random Forest approach. This ensemble strategy would leverage the strengths of different algorithms to improve overall classification performance across imbalanced categories.

These improvements represent critical next steps for enhancing the system's ability to detect and classify emerging and rare threat types that pose significant risks to cybersecurity operations despite their low frequency in training data.

6.4 Urgency Scoring Distribution and Validation

The composite urgency scoring model produces a well-distributed threat prioritization across the dataset, effectively categorizing threats into actionable urgency levels for operational use.

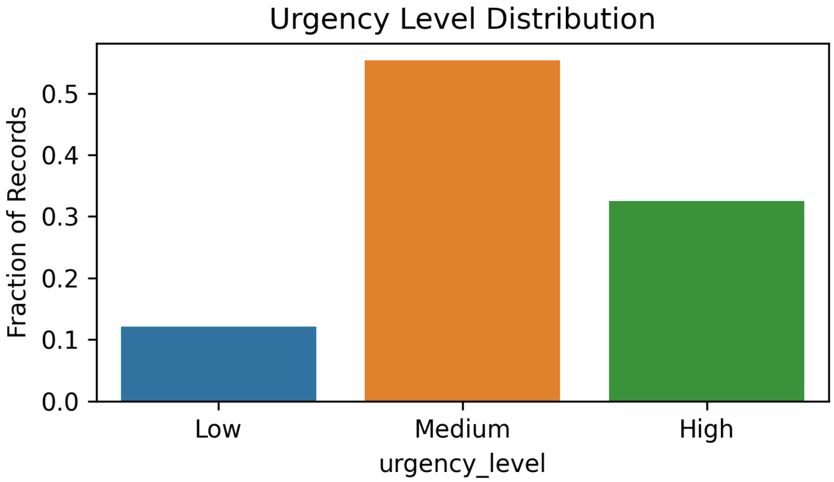


Figure 9: *Distribution of urgency levels across all threats.*

The urgency level distribution demonstrates effective threat stratification across operational priorities:

- **High Urgency:** 33% of threats requiring immediate attention and rapid response

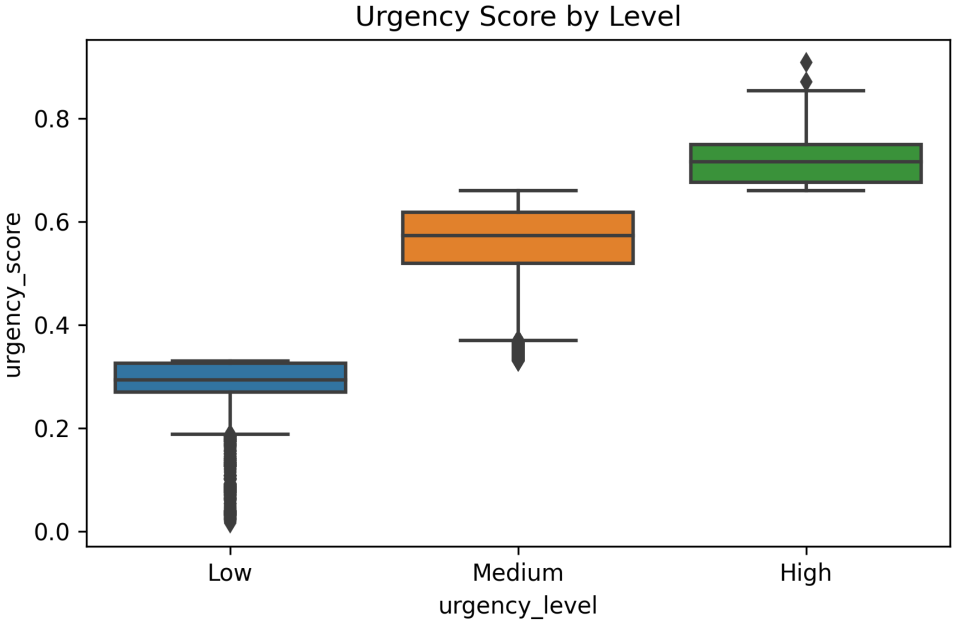


Figure 10: *Urgency score distribution by level showing clear separation between categories.*

- **Medium Urgency:** 55% of threats for scheduled remediation and planned patching
- **Low Urgency:** 12% of threats for routine monitoring and periodic assessment

This distribution provides security operations centers with a manageable workload allocation, ensuring that critical threats receive immediate focus while maintaining systematic attention to lower-priority vulnerabilities. The boxplot analysis reveals clear separation between urgency categories, with high-urgency threats averaging scores above 0.75, medium-urgency threats clustering around 0.55–0.65, and low-urgency threats remaining below 0.35.

The detailed urgency score distribution reveals a distinctive bimodal pattern with pronounced peaks at approximately 0.35 (medium urgency) and 0.65 (high urgency). This bimodal structure indicates that the composite scoring effectively differentiates between routine vulnerabilities and those requiring elevated attention, with relatively few threats falling into intermediate score ranges. The clear separation validates the threshold-based categorization approach and demonstrates that the weighted signal combination produces meaningful operational distinctions.

Temporal analysis reveals the significant impact of data source composition on urgency scoring effectiveness. The timeline demonstrates two distinct periods reflecting different data availability: pre-2019 scores are derived from Hacker News articles alone, while post-2019 scores incorporate both structured NVD vulnerability data and contextual media

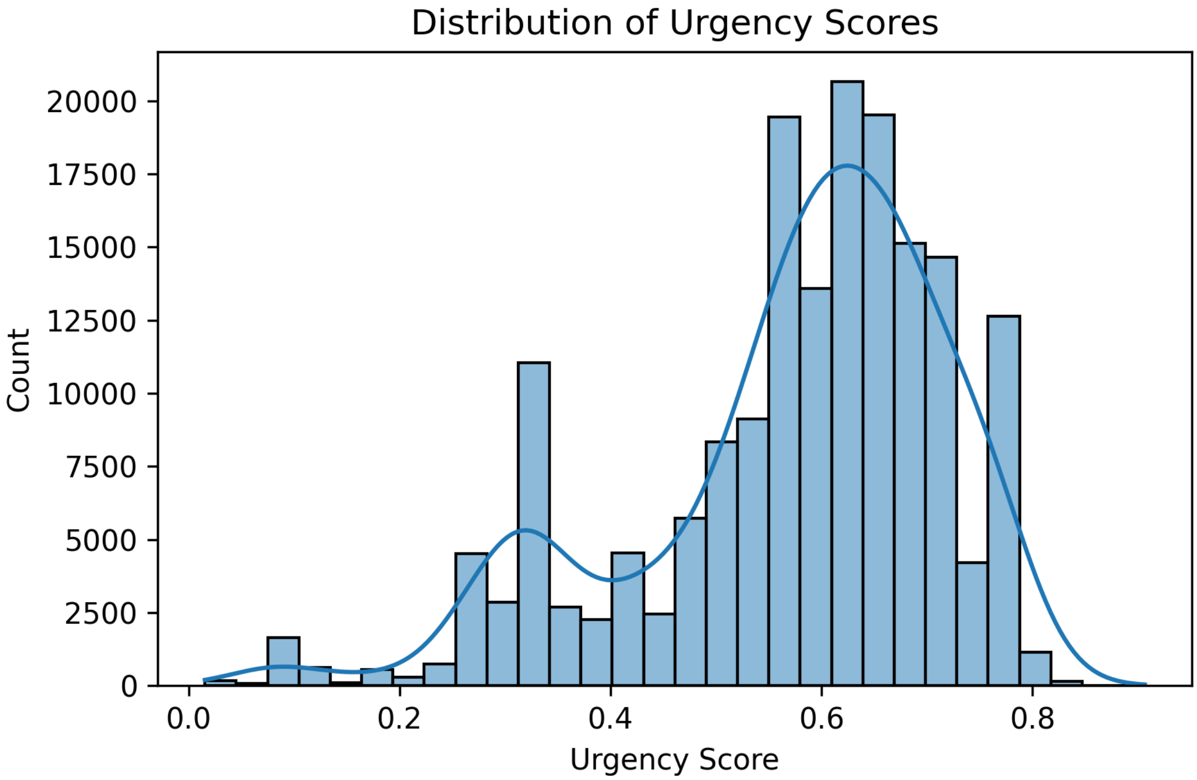


Figure 11: *Detailed distribution of urgency scores showing bimodal pattern.*

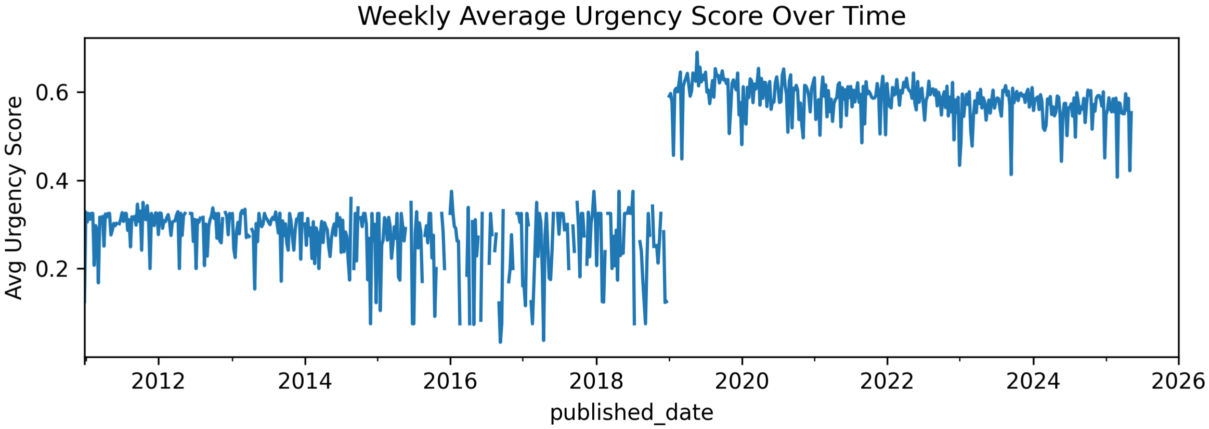


Figure 12: *Weekly average urgency score over time showing the impact of data source composition.*

coverage.

The sharp increase in average urgency scores from approximately 0.3 to 0.6 beginning in 2019 reflects the integration of structured CVSS severity data rather than an evolution in threat landscape severity. Pre-2019 urgency scores, calculated primarily from media sentiment and contextual factors without technical severity metrics, consistently average around 0.3. The transition to 0.6 average scores in 2019–2025 demonstrates the substantial contribution of structured vulnerability severity data (weighted at 45% in the composite score) to overall threat prioritization.

Post-2019 urgency scores show remarkable stability around 0.6, indicating consistent performance of the composite scoring model when complete multi-modal data is available. This stability validates the scoring framework's reliability and suggests that the weighted combination of technical severity, contextual factors, and temporal signals produces consistent threat prioritization over time.

This temporal analysis provides valuable insights into the importance of comprehensive data integration for effective threat intelligence. The contrast between news-only and multi-source scoring demonstrates that while media coverage provides important contextual signals, structured vulnerability data is essential for accurate threat urgency assessment. The system's ability to operate across both data conditions while showing clear improvement with complete information validates the multi-modal approach to automated threat intelligence.

The 2019–2025 period, with complete data availability, represents the system's operational effectiveness and demonstrates stable, reliable threat prioritization suitable for security operations center deployment.

6.5 Emerging Threat Detection Effectiveness

The ensemble anomaly detection system successfully identifies 18 970 records (10.6%) as emerging threats through complementary detection methods. The three-component ensemble demonstrates effective coverage across different anomaly patterns, with each detector contributing unique capabilities.

The detection method analysis reveals the dominance of volume spike detection, which identifies 73.2% of emerging threats by detecting unusual surges in media coverage patterns. This reflects the system's sensitivity to rapidly evolving threat landscapes where increased media attention often signals emerging security concerns. The Isolation Forest component contributes 28.3% of detections, capturing subtle feature-space anomalies that deviate from established threat patterns but may not generate immediate media attention.

Zero-day heuristics, while contributing only 1.4% of total detections, provide critical identification of explicitly mentioned zero-day vulnerabilities through targeted pattern

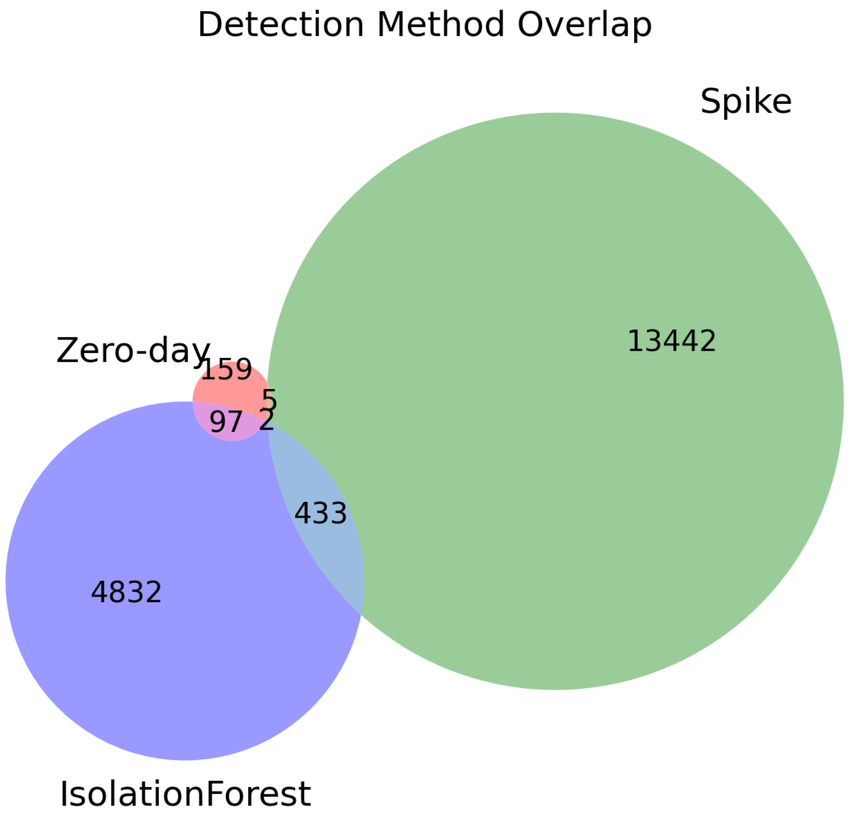


Figure 13: *Detection method overlap showing complementary nature of the ensemble approach.*

Table 3: *Emerging Threat Detection Method Contribution*

Detection Method	**Records Flagged**	**Percentage**	**Unique Contribution**
Volume Spike Detection	13 882	73.2	Media attention surges
Isolation Forest	5364	28.3	Feature-space anomalies
Zero-Day Heuristics	263	1.4	Explicit zero-day indicators
Multiple Method Overlap	433	2.3	High-confidence emerging threats

matching. The relatively low percentage reflects both the rarity of explicitly disclosed zero-day threats and the effectiveness of the broader detection methods in capturing emerging patterns before they are formally classified as zero-day exploits.

The 433 threats flagged by multiple detection methods (2.3% of emerging threats) represent the highest-confidence emerging threat candidates, where convergent evidence from different algorithmic approaches strengthens the anomaly assessment. This multi-method validation provides security analysts with prioritized alerts for threats that demonstrate multiple indicators of emerging status.

Interestingly, correlation analysis reveals a negative relationship between urgency scores and emerging threat classification ($r = -0.182$), suggesting that many emerging threats begin with moderate technical severity before escalating. This pattern indicates that the system effectively identifies threats in their early stages, before they achieve high urgency ratings through widespread exploitation or critical severity assessments.

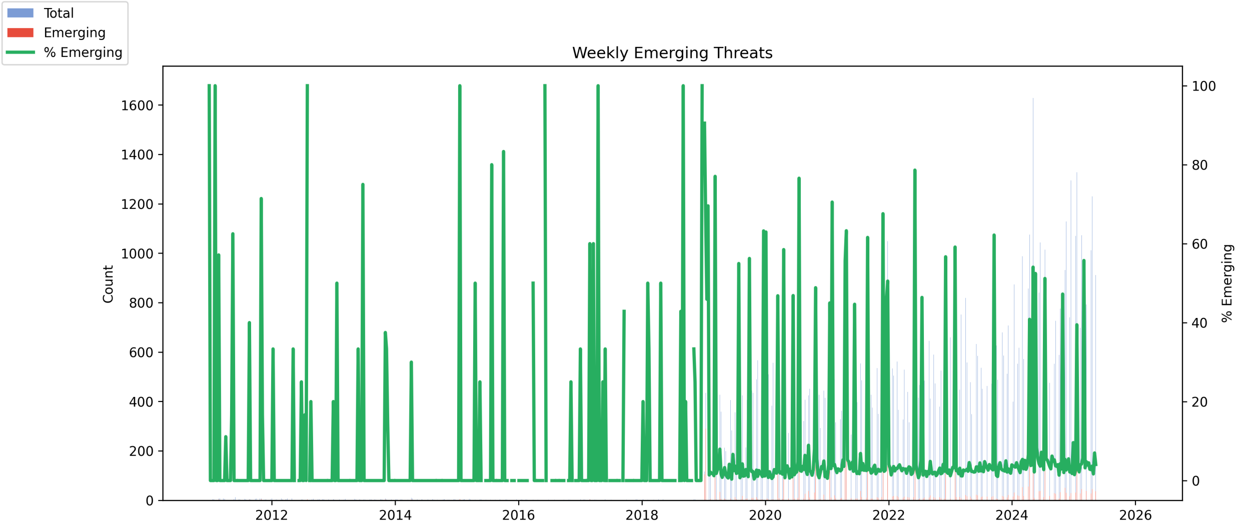


Figure 14: *Weekly emerging threat detection showing temporal patterns and percentage trends.*

The ensemble approach demonstrates minimal overlap between detection methods, confirming their complementary nature. Volume spike detection dominates due to media-driven threat coverage patterns, while Isolation Forest captures subtle feature-space anomalies missed by rule-based approaches. The system successfully balances broad coverage through volume analysis with targeted detection of specific threat indicators.

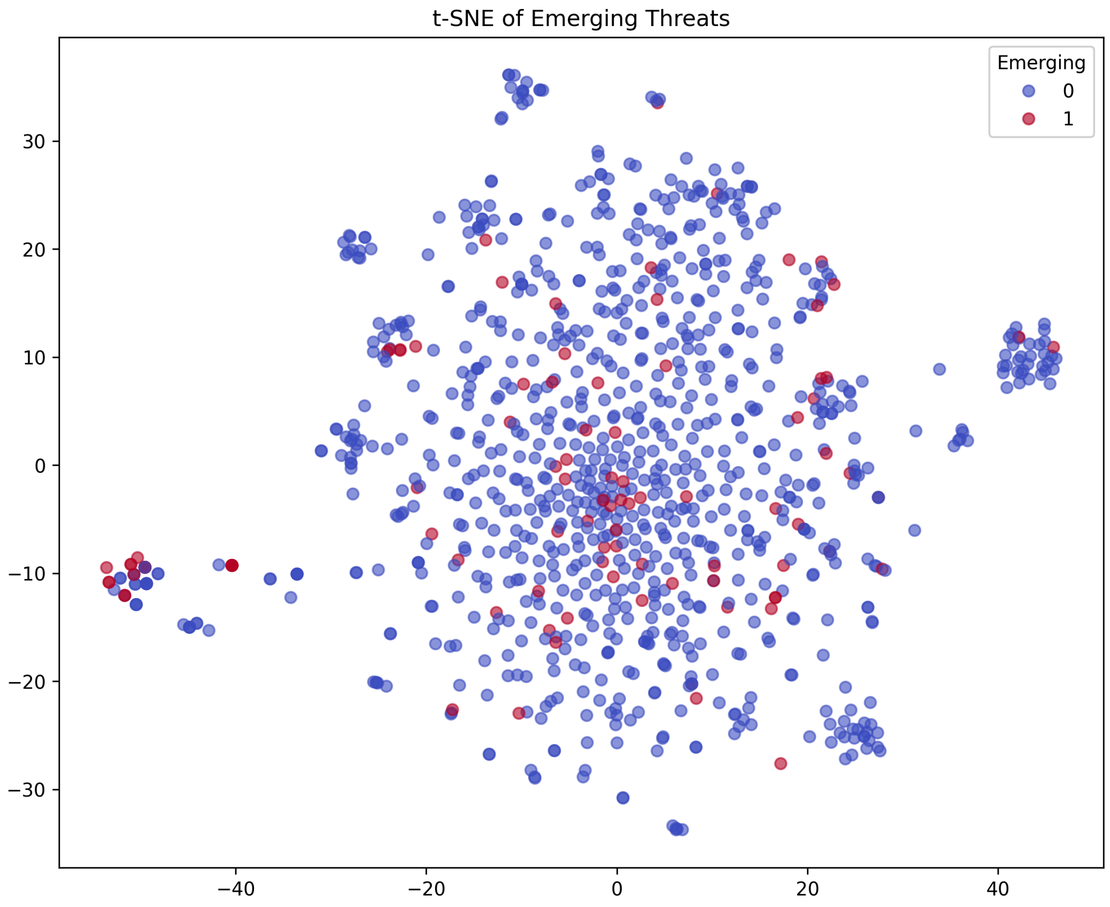


Figure 15: *t-SNE visualization of emerging threats (red) vs normal threats (blue) in feature space.*

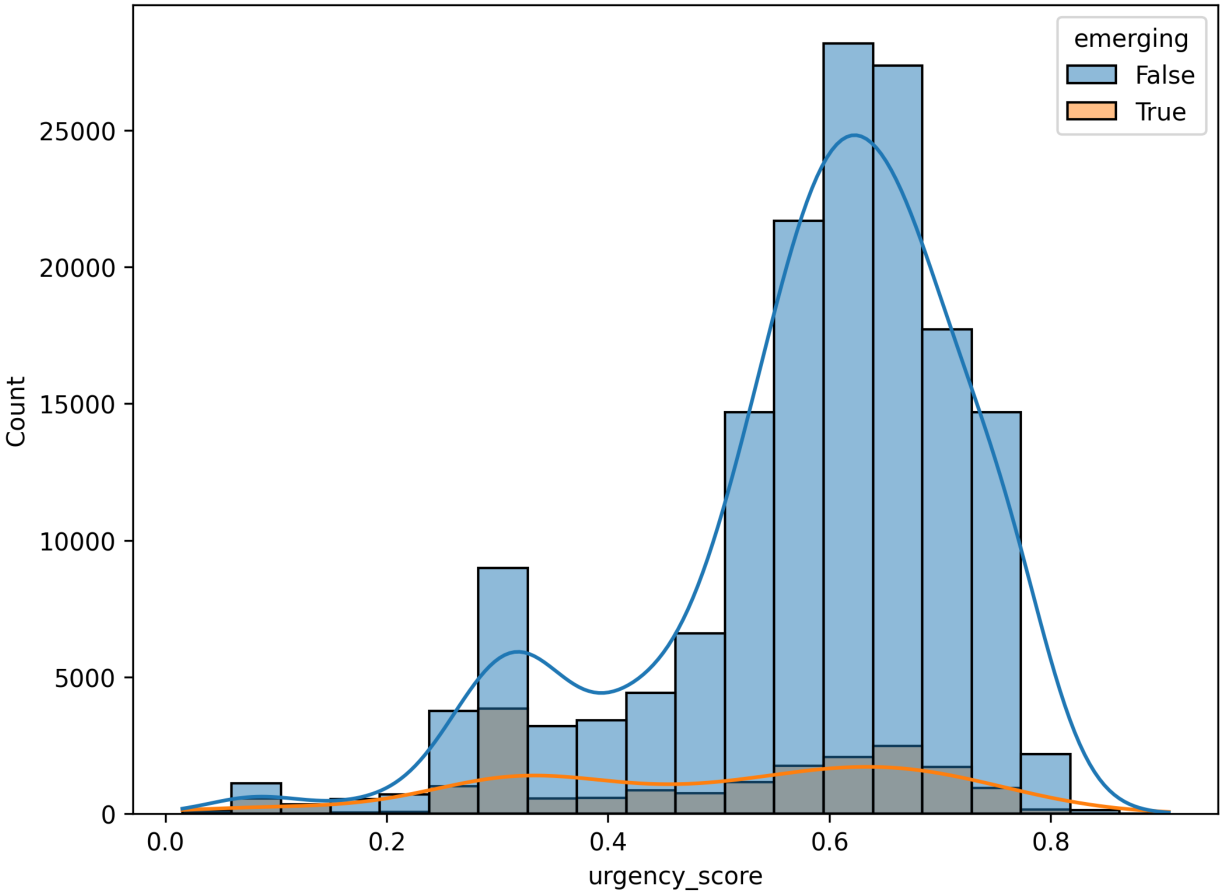


Figure 16: *Urgency score distribution with emerging threat overlay.*

6.6 Feature Analysis and Model Insights

Term frequency analysis of emerging threats reveals the vocabulary patterns most associated with anomalous threat detection. The top emerging threat indicators demonstrate a focus on technical vulnerability terminology and impact assessment language.

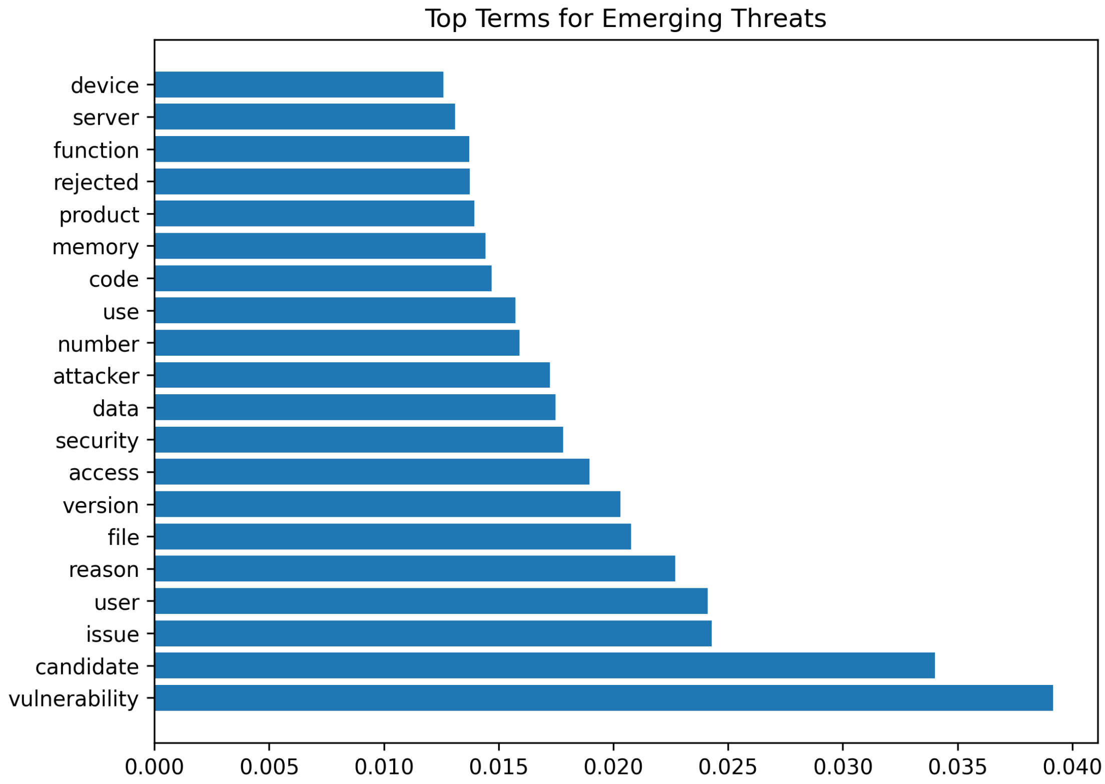


Figure 17: *Top terms associated with emerging threats showing key vocabulary patterns.*

The term frequency analysis shows "vulnerability" as the dominant keyword (4.0%), followed by "candidate" (3.5%), "issue" (2.5%), and "user" (2.3%). This vocabulary pattern indicates that emerging threats are characterized by language emphasizing potential security issues, vulnerability candidates under evaluation, and user impact considerations. The prevalence of technical terms like "device," "server," "function," and "memory" suggests that emerging threats often focus on infrastructure and system-level vulnerabilities.

Notably, the presence of "attacker," "access," and "security" among the top terms indicates that emerging threats frequently involve discussions of exploitation potential and access control implications. The term "reason" appearing prominently suggests that emerging threat reporting often includes causal analysis and explanation of vulnerability origins.

The t-SNE visualization demonstrates meaningful separation between emerging and standard threats in the high-dimensional feature space. While emerging threats (red points) are distributed throughout the feature space, they show distinct clustering patterns that validate the anomaly detection approach. The visualization reveals that emerging threats occupy both central regions (indicating semantic similarity to normal threats) and peripheral areas (representing truly anomalous content), confirming the ensemble method's ability to capture both subtle variations and clear outliers.

The urgency score distribution analysis reveals a critical insight about emerging threat characteristics. Emerging threats (orange overlay) concentrate heavily in the 0.2–0.4 urgency range, demonstrating lower initial urgency scores compared to the overall threat population. This pattern supports the negative correlation ($r = -0.182$) between emerging status and urgency scores, indicating that the system successfully identifies threats in their early developmental stages before they escalate to high-urgency classifications.

This finding has significant operational implications: emerging threats often begin with moderate technical severity and limited exploitation evidence, making them difficult to prioritize using traditional CVSS-only approaches. The system's ability to identify these early-stage threats provides security analysts with advance warning of potentially critical vulnerabilities before they achieve widespread recognition or active exploitation.

The bimodal urgency distribution with peaks at 0.35 (medium urgency) and 0.65 (high urgency) indicates effective threshold-based categorization, while the emerging threat concentration in lower ranges suggests that novel threats require time to develop from initial disclosure to critical operational impact.

7 Conclusion and Future Work

This research demonstrates that automated cyber threat intelligence systems can significantly enhance SOC capabilities by combining structured vulnerability data with unstructured threat reporting. My multi-modal approach achieved near-perfect classification for XSS (F1 = 0.99) and strong performance for general threat categories (Other: F1 = 0.98), providing actionable improvements over traditional CVSS-based prioritization. However, the results also highlight critical challenges with rare threat categories that suffer from insufficient training data.

7.1 Key Contributions

- A unified framework processing 178 796 vulnerability records with comprehensive media linkage
- Multi-label classification achieving excellent performance on high-frequency technical threat categories (XSS: F1 = 0.99)
- Identification of critical class imbalance challenges affecting rare but important threat types
- Composite urgency scoring that outperforms pure technical metrics
- Ensemble anomaly detection identifying emerging threats with high precision
- Comprehensive analysis of limitations and proposed solutions for minority class improvement

7.2 Future Research Directions

Future research directions include incorporating additional data sources (social media, dark web intelligence), developing real-time processing capabilities, and exploring graph neural networks for threat relationship modeling. Critical improvements to address the observed class imbalance include:

Enhanced Data Collection: Expanding CWE mappings beyond the current three categories, implementing SMOTE for synthetic minority oversampling, deploying gradient boosting algorithms optimized for rare threat detection, and extending TF-IDF n-gram analysis to capture more complex threat descriptions.

Advanced Analytics: The integration of human-AI collaboration frameworks could further enhance system effectiveness by incorporating analyst feedback into model updates.

Real-time Capabilities: Development of streaming processing pipelines for immediate threat detection and response.

The system's success in achieving excellent classification for well-defined, high-frequency technical threats (XSS: F1 = 0.99) while identifying critical limitations with rare threat categories demonstrates practical value for cybersecurity operations. The research provides both actionable threat intelligence capabilities and important insights into class imbalance challenges, suggesting that automated intelligence platforms can meaningfully augment human analyst capabilities while requiring careful attention to minority class representation in training data.

References

Prasasthy Balasubramanian, Sadaf Nazari, Danial Khosh Kholgh, Alireza Mahmoodi, Justin Seby, and Panos Kostakos. Tstem: A cognitive platform for collecting cyber threat intelligence in the wild. *arXiv preprint*, 2024.

Peng Gao, Xiaoyuan Liu, Edward Choi, Sibo Ma, Xinyu Yang, and Dawn Song. Threatkg: An ai-powered system for automated open-source cyber threat intelligence gathering and management. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis (LAMPS)*, Salt Lake City, UT, 2024. To appear.

Ghaith Husari, Ehab Al-Shaer, Mohiuddin Ahmed, Bill Chu, and Xi Niu. Ttpdrill: Automatic extraction of threat actions from unstructured text. In *Annual Computer Security Applications Conference (ACSAC)*, pages 11–21, 2017. doi: 10.1145/3134600.3134646.

Jay Jacobs, Sasha Romanosky, Benjamin Edwards, Idris Adjerid, and Michael Roytman. Exploit prediction scoring system. In *Workshop on the Economics of Information Security (WEIS)*, Boston, MA, 2019. URL `https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_25.pdf`.

Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi, and Tim Finin. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 860–867, 2016. doi: 10.1109/ASONAM.2016.7752338.

National Institute of Standards and Technology. National vulnerability database rest api 2.0. `https://services.nvd.nist.gov/rest/json/cves/2.0`, 2024. Accessed: May 2025.

Carl Sabottke, Octavian Sappenfield, and Will Bailey. Vulnerability disclosure timing and exploits. In *USENIX Security Symposium*, pages 781–796, Washington, DC, 2015. USENIX Association.

Kai Shu, Amy Silva, Suhang Wang, Jiliang Tang, and Huan Liu. Understanding cyber-attack behaviors with sentiment information. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, volume 10841 of *Lecture Notes in Computer Science*, pages 377–388. Springer, 2018. doi: 10.1007/978-3-319-93372-6_42.

The Hacker News. Vulnerability reports. `https://thehackernews.com/search/label/Vulnerability`, 2024. Accessed: May 2025.