AI-Based Cyber Security Threat Prediction

Data Sources

I. Foundational Data Strategy and Sources for AI-Driven Threat Prediction

The AI threat prediction project relies on two primary categories of data: **Static** data for initial model training and benchmarking, and **Dynamic** (real-time/API) data for operational threat detection and continuous monitoring.

I.A. Static Data Sources (Training and Benchmarking)

Static datasets are collected offline and provide labeled instances of network flows, malware, and system logs, forming the essential foundation for supervised machine learning model development.

Category	Source	Data Type	Advantages	Limitations
Malware / Attack Logs	**CICIDS2017, CSE-CIC-IDS 2018**	Labeled Network Flow Statistics	Established benchmark, diverse attack scenarios (DDoS, Botnet), real-world malware interaction, and labeled data.	Huge volume of data, scattered presence, often contains missing values, and suffers from severe class imbalance.
Malware / Attack Logs	**UNSW-NB15**	Labeled Network Flow Statistics	Contains nine categories of modern attacks (Exploits,Reconnaissance); used for binary and multi-class analysis.	High computational cost, imbalanced classes, features require complex preprocessing.

IoT/SCADA Traffic	**TON_IoT**	Real IoT Telemetry & Network Attacks	Provides real, domain-specific IoT telemetry and attack logs.	Limited attack types compared to general network traffic, and some scenarios are simulated.
Classic IDS Data	**NSL-KDD**	Classic Intrusion Detection Dataset	Established classic dataset often used for fast, comparative baseline studies.	Represents outdated attacks and network topologies (traffic generated in 1998), making it less representative of modern threats.
Malware Samples / Hashes	**VirusShare, MalShare, Malpedia**	File Hashes, Binaries, Family Info	Crucial for file classification and forensic analysis.	Requires continuous monitoring and updating due to the high volume and constant influx of new types.
Endpoint & System Logs	**Azure Sentinel Datasets, Security Datasets Project**	Simulated Enterprise Telemetry (Sysmon, events)	Realistic attack simulations providing deep host-level context via logs.	Data is often simulated, potentially lacking the full complexity of real, live production environments.

I.B. Dynamic Data Sources (Real-Time and API Feeds)

Dynamic data sources provide streaming feeds essential for maintaining real-time situational awareness, enriching raw events, and enabling immediate threat prediction during live operations.

Category	Source / Tool	Data Type	Advantages	Limitations
Threat Intelligence APIs	**VirusTotal API, AlienVault OTX, AbuseIPDB, Cisco Talos**	Real-time Indicators of Compromise (IOCs)	Provides immediate, current context, enabling proactive defense by transforming raw observations into actionable intelligence.	Proprietary feeds can lack transparency, and the high volume of traffic generated from API calls can be challenging to manage.
SIEM / Log Feeds	**Splunk HEC, Elastic Security, Azure Sentinel**	Structured Security Events (JSON/CSV)	Centralized view of security events, aggregates data from diverse sources, and enables faster threat detection and response.	Requires specialized infrastructure (SIEM) and considerable human effort to triage and investigate high volume of alerts.
Network Traffic Stream	**Zeek (Bro), Suricata, Snort,**	Live Traffic Features	Collects and parses raw traffic features necessary for	Requires high computational power to process raw

	NetFlow		flow-based analysis and network intrusion detection.	packets and maintain low latency for real-time operation.
System Monitoring	**Sysmon, osquery, Auditd**	Local Host Activity Sequence Logs	Provides granular, host-level telemetry, which is vital for sequential behavioral modeling and forensic analysis.	Generates massive data volume that must be reliably streamed, ingested, and stored.

Works cited

1. A Comprehensive Study on CIC-IDS2017 Dataset for Intrusion Detection Systems, accessed October 15, 2025, https://www.researchgate.net/publication/378709289_A_Comprehensive_Study_on_CIC-IDS2017_Dataset_for_Intrusion_Detection_Systems
2. CIC UNSW-NB15 Augmented Dataset - University of New Brunswick, accessed October 15, 2025, https://www.unb.ca/cic/datasets/cic-unsw-nb15.html