Data Sources for AI-Driven Cyber Security Threat Prediction

1. In this project, both static and real-time datasets are used to train and evaluate models for detecting various types of cyber threats.
2. Static datasets contain pre-collected data such as malware files, phishing URLs, emails, and source code, while real-time datasets capture live or simulated network traffic and ongoing attacks.

The tables below summarize the best and most widely accepted datasets for each threat type, along with their descriptions and data sources:

STATIC THREATS:

🧩 Data Sources

S. NO	Threat Type	Dataset	Description
1.	Malware Classification	EMBER (by Endgame / Elastic)	Large-scale dataset of Windows PE (Portable Executable) malware and benign files. Each file has 2,381 static features such as entropy, imported functions, and header information. Ideal for ML models like Random Forest and XGBoost.
2.	Phishing URL Detection	UCI Phishing Websites Dataset	Structured dataset with 30 lexical and host-based URL features (domain length, HTTPS usage, presence of "@", IP in URL, etc.). Great for classical ML models.
3.	Malicious Email / Spam Detection	Enron Email Dataset (by CMU)	Contains 500,000+ emails labeled as spam or ham (legitimate). Includes headers, subject lines, and message text. Useful for NLP-based spam/phishing detection.
4.	Static Code Vulnerability Detection	Juliet Test Suite (by NIST)	Collection of C/C++/Java programs containing both vulnerable and secure code snippets. Used to train AI models for static vulnerability analysis.
5.	Malware Image Classification	Malimg Dataset (by Vision Research Lab)	Malware binaries converted into grayscale images — enables CNN-based image recognition for malware family classification.

🔗 Static Datasets Source Links

1. EMBER (Malware Classification) – https://github.com/elastic/ember
2. UCI Phishing Websites Dataset (Phishing URL Detection) – https://archive.ics.uci.edu/ml/datasets/phishing+websites
3. Enron Email Dataset (Malicious Email / Spam Detection) – https://www.cs.cmu.edu/~enron/
4. Juliet Test Suite (Static Code Vulnerability Detection) – https://samate.nist.gov/SARD/test-suites/112

5. Malimg Dataset (Malware Image Classification)-
https://www.kaggle.com/datasets/ashwinsharmaaa/malimg-dataset

REAL-TIME THREATS:

🧩 Data Sources

S. NO	Threat Type	Dataset	Description
1.	Intrusion Detection (IDS / DDoS / PortScan / Brute Force)	CICIDS2017 (by Canadian Institute for Cybersecurity)	Realistic modern network dataset containing normal and attack traffic (DoS, DDoS, PortScan, Brute Force, Botnet). Each network flow has 80+ statistical features.
2.	IoT / Botnet Threat Detection	BoT-IoT (by UNSW Canberra)	IoT-based network dataset simulating normal and malicious traffic including DDoS, Keylogging, and Data Theft. Ideal for deep learning models (LSTM, CNN).
3.	Real-Time IP / Domain Threat Intelligence	AlienVault OTX (Open Threat Exchange)	Real-time threat intelligence feed with constantly updated malicious IPs, domains, URLs, and file hashes. Accessed via API for live detection.
4.	Network Anomaly Detection	UNSW-NB15 (by UNSW Canberra)	Dataset combining synthetic and real traffic representing modern attacks like Fuzzers, Exploits, Worms, and Shellcode. Provides 49 network features for ML.
5.	DDoS Attack Detection (Advanced IDS)	CSE-CIC-IDS2018 (by Canadian Institute for Cybersecurity)	Updated IDS dataset with newer attack types: DDoS, Web Attacks, Infiltration, and Botnet. Contains CSVs suitable for ML/DL pipelines.

🔗 Real-Time Datasets Source Links

1. CICIDS2017 (Intrusion Detection / DDoS / PortScan / Brute Force) –
https://www.unb.ca/cic/datasets/ids-2017.html
2. BoT-IoT Dataset (IoT / Botnet Threat Detection) –
https://research.unsw.edu.au/projects/bot-iot-dataset
3. AlienVault OTX (Real-Time IP / Domain Threat Intelligence) – https://otx.alienvault.com/
4. UNSW-NB15 Dataset (Network Anomaly Detection) –
https://research.unsw.edu.au/projects/unsw-nb15-dataset
5. CSE-CIC-IDS2018 (Advanced IDS / DDoS Attack Detection) –
https://www.unb.ca/cic/datasets/ids-2018.html