AI-Based Threat Prediction System

# Data Source Reference Document

This document provides structured and reliable data sources for both **Static** and **Real-Time** datasets used in developing an **AI-based Threat Prediction System**. These datasets support the creation of intelligent agents capable of detecting, predicting, and responding to cybersecurity threats.

# Static Data Sources

Static datasets are primarily used for model training and testing. They contain labeled or structured data collected from past incidents, network traffic, or system logs, which helps in building predictive AI models.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Description** | **Link** |
| NSL-KDD | Improved version of the KDD99 dataset designed for intrusiom detection benchmarking | https://www.unb.ca/cic/datasets/nsl.html |
| CICIDS2017 | Modern network dataset featuring DDoS, brute-force, and botnet attacks | https://www.unb.ca/cic/datasets/ids-2017.html |
| UNSW-NB15 | Network dataset with 49 features covering normal and malicious traffic | https://research.unsw.edu.au/projects/unsw-nb15-dataset |
| EMBER | Dataset for malware classification based on Windows PE files. | https://github.com/elastic/ember |
| HDFS Logs | System log dataset useful for anomaly and event sequence analysis | https://github.com/logpai/loghub |
| PhishTank | Community-maintained phishing URL repository for threat identification. | https://phishtank.org/ |

# Real-Time Data Sources

Real-time data sources enable continuous monitoring and predictive analytics. These sources typically provide APIs or live feeds that an AI Agent can consume dynamically for real-time threat detection and alerting.

|  |  |  |
| --- | --- | --- |
| **Source** | **Description** | **Link** |
| AlienVault OTX | Open Threat Exchange offering live threat intelligence and IOCs feeds. | https://otx.alienvault.com/ |
| AbuseIPDB API  https://www.abuseipdb.com/ | API for identifying and reporting malicious IP addresses in real time. | https://www.abuseipdb.com/ |
| MISP Feeds | Threat intelligence feeds from the Malware Information Sharing Platform (MISP) | https://www.misp-project.org/feeds/ |
| Shodan API  https://www.shodan.io/ | Provides real-time data on connected devices and their vulnerabilities. | https://www.shodan.io/ |
| CIRCL Passive DNS | Tracks domain resolutions for suspicious or malicious behavior. | https://www.circl.lu/services/passive-dns/ |
| VirusTotal API  https://www.virustotal.com/gui/home/upload | Provides real-time scanning data and reputation reports on files and URLs. |  |

**Summary:**

Static datasets are ideal for training and validating AI models, while real-time data sources help build a continuously learning and adaptive AI Agent capable of monitoring active threats.

Combining both enables a hybrid intelligence system that evolves over time.