Mazen's Sprint:

Dataset Selected: https://arxiv.org/abs/2103.05767

Analysis:

- **Source**: This dataset was collected from the raw output of firewalls in a real network environment.
- **Description**: ZYELL-NCTU NetTraffic-1.0 is a large-scale, real dataset designed to advance network security research. It includes network traffic data captured from firewalls, containing both normal and anomalous activities. The dataset provides a realistic environment for developing and testing network anomaly detection models.
- **Usability**: The dataset is made for supervised and unsupervised machine learning intrusion detection system tests. It offers a view of network traffic, which is good for training and evaluating detection models.
- **Comparative Analysis**: Compared to older datasets, ZYELL-NCTU NetTraffic-1.0 reflects the current network conditions, making it more applicable for modern network security research. Its real world data collection enhances its relevance to contemporary network environments.

Research Paper Selected: https://www.nature.com/articles/s41598-025-88050-z

Analysis:

- **Problem Statement**: The paper addresses the challenge of efficiently detecting anomalies in tabular cybersecurity data, which is needed for identifying security threats in network logs.
- **Dataset Used**: The study uses a dataset comprising tabular cybersecurity data, but specific details about the dataset aren't provided, most likely due to privacy reasons.
- **Methodologies Used**: The research introduces a new method for anomaly detection based on LLMs specifically designed for tabular data. This approach leverages the capabilities of LLMs to identify anomalies in complex data.
- **Findings and Contributions**: The study presents an innovative method for tabular data anomaly detection, demonstrating the effectiveness in identifying anomalies in cybersecurity data.
- **Relevance to Project**: The paper's findings are relevant to our project because they explore advanced AI techniques for anomaly detection in network security. The use of LLMs for tabular data could better prepare ourselves when we develop our AI based Anomaly Detection Software, possibly enhancing its ability to detect anomalies in network logs and other features as well.

Vadim's Sprint:

Dataset Selected: https://www.unb.ca/cic/datasets/ids-2017.html

Analysis:

- **Source**: Collected by the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick.

- **Description**: This dataset includes network traffic data simulating both normal and attack scenarios, such as brute force, botnets, DoS, and web-based attacks. It provides flow-based and packet-based features extracted using tools like Wireshark.
- **Usability**: Suitable for supervised learning, requiring preprocessing such as feature selection, encoding categorical variables, and handling class imbalances. Useful for training and evaluating intrusion detection models.
- **Comparative Analysis**: CICIDS2017 offers a realistic network environment with up-to-date attack patterns. Unlike older datasets (e.g., KDDCup'99), it provides labeled, structured data with modern network attack types, improving its relevance for contemporary intrusion detection research.

Dataset Selected: https://research.unsw.edu.au/projects/toniot-datasets
Analysis:

- **Source**: Collected by the Cyber Range Lab at the University of New South Wales (UNSW) for IoT security research.
- **Description**: This dataset consists of network traffic, operating system logs, and telemetry data from IoT devices, including normal and attack behaviors. It captures cyber threats targeting smart home and industrial IoT environments.
- **Usability**: Supports both supervised and unsupervised machine learning approaches. Requires preprocessing to integrate multiple data types (network packets, system logs, and telemetry) and handle potential class imbalances.
- **Comparative Analysis**: Unlike traditional network datasets like CICIDS2017, TON_IoT provides a broader view of cyber threats by including IoT device vulnerabilities. This makes it valuable for researchers focusing on IoT-specific security challenges.

Research Paper Selected: https://arxiv.org/abs/2212.04546
Analysis:

- **Problem Statement**: The paper addresses the challenge of efficiently detecting anomalies in tabular cybersecurity data, which is crucial for identifying security threats in network logs.
- **Dataset Used**: The study utilizes a tabular cybersecurity dataset, but specific details are not disclosed, likely due to privacy concerns.
- **Methodologies Used**: Introduces a novel anomaly detection approach using Large Language Models (LLMs) optimized for tabular data, leveraging their ability to identify complex anomalies.
- **Findings and Contributions**: Demonstrates the effectiveness of LLM-based methods in detecting anomalies in cybersecurity data, presenting a new direction for AI-driven security analysis.
- Relevance to Project: The study provides valuable insights into using LLMs for anomaly detection, which could inform the development of AI-based security solutions in our project, enhancing network log analysis and threat detection.

Tyler's Sprint:

Dataset Selected: https://www.unb.ca/cic/datasets/ids-2018.html

Analysis:

- **Source**: Collected by the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick.
- **Description**: The final dataset includes seven different attack scenarios: Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and infiltration of the network from inside. The attacking infrastructure includes 50 machines and the victim organization has 5 departments and includes 420 machines and 30 servers. The dataset includes the captured network traffic and system logs of each machine, along with 80 features extracted from the captured traffic using CICFlowMeter-V3.
- **Usability**: Detailed descriptions of intrusions and abstract distribution models for applications, protocols, or lower level network entities. These profiles can be used by agents or human operators to generate events on the network. Due to the abstract nature of the generated profiles, we can apply them to a diverse range of network protocols with different topologies. Profiles can be used together to generate a dataset for specific needs.
- **Comparative Analysis**: This dataset is set apart from other datasets by being modular. Profiles can be used by agents or human operators to generate events on the network. Due to the abstract nature of the generated profiles, we can apply them to a diverse range of network protocols with different topologies. Profiles can be used together to generate a dataset for specific needs. We will build two distinct classes of profiles.

Dataset Selected: https://research.unsw.edu.au/projects/unsw-nb15-dataset

Analysis:

- **Source**: Created by the IXIA PerfectStorm tool in the Cyber Range Lab of UNSW Canberra
- **Description**: This dataset has nine types of attacks, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. The Argus, Bro-IDS tools are used and twelve algorithms are developed to generate totally 49 features with the class label. These features are described in the UNSW-NB15_features.csv file.
- **Usability**: Suitable for developing Intrusion Detection, Network Forensics, and Privacy-preserving, and Threat Intelligence approaches in different systems, such as Network Systems, Internet of Things (IoT), SCADA, Industrial IoT, and Industry 4.0.
- **Comparative Analysis**: UNSW-NB15 has these three advantages over older datasets: the presence of modern low footprint attack styles, modern normal traffic scenarios, and a similar distribution of training and testing sets.

Research Paper Selected:

https://ajpojournals.org/journals/index.php/EJT/article/view/1486

Analysis:

- **Problem Statement**: The paper addresses the growing challenges of cybersecurity threats and the limitations of traditional security mechanisms. It highlights how machine learning (ML) can enhance cybersecurity by providing automated, adaptive, and intelligent threat detection systems.
- **Dataset Used**: The paper does not focus on a single dataset but discusses multiple cybersecurity applications where ML is applied. It references datasets for:
    - Intrusion Detection: KDD Cup 1999, NSL-KDD.
    - Phishing Detection: Labeled phishing emails and legitimate email datasets.
    - Spam Detection: Email classification datasets.
- **Methodologies Used**: The paper outlines key ML techniques used in cybersecurity:
    - Supervised Learning: Decision Trees, Support Vector Machines (SVM), Neural Networks.
    - Unsupervised Learning: K-Means Clustering, Isolation Forests, DBSCAN for anomaly detection.
    - Deep Learning: CNNs and RNNs for malware analysis and intrusion detection.
    - Ensemble Methods: Random Forests, boosting techniques for improved accuracy.
    - Feature Engineering: Extraction of key features from email content, URLs, and network traffic logs.
- **Findings and Contributions**: ML models can significantly improve the accuracy of cybersecurity threat detection compared to traditional rule-based systems. Intrusion detection systems (IDS) benefit from deep learning models that can adapt to evolving attack patterns. Phishing detection models need continuous updates to remain effective against new attack techniques. Spam classification is best handled with Logistic Regression due to its high precision and low false positives.
    - Challenges identified: Data quality, adversarial attacks against ML models, and interpretability of deep learning models.
- **Relevance to Project**: The paper validates that ML-based anomaly detection is an effective approach for intrusion detection. It highlights unsupervised learning techniques (e.g., Isolation Forest, One-Class SVM) as key methods for detecting unknown cyber threats. It underscores the importance of feature selection and data preprocessing, which will be crucial for your model. It discusses model adaptability, which aligns with the need to continuously refine the ADNS system to detect evolving cyber threats.