Final Project Report: Anomaly Detection in Network Traffic using a Transformer Autoencoder

Group Name: Bumblebee
Members: Alexios C Papazoglou, Geetika Khanna, Mohit Saluru, Pragati Rao
Date: Aug 21, 2025 | **Github Link:** https://github.com/geetikak13/dlproject

---

1. Abstract

Distributed Denial of Service (DDoS) attacks pose a persistent and evolving threat to the availability of online services. Traditional supervised detection methods struggle to identify novel, zero-day attack vectors due to their reliance on pre-labeled data. This project addresses this challenge by developing an unsupervised anomaly detection system using a Transformer-based autoencoder. By training the model exclusively on benign network traffic from the CIC-DDoS2019 dataset, it learns a deep, contextual representation of normal behavior. Anomalies, such as DDoS attacks, are then identified by their high reconstruction error when processed by the model. A key contribution is the development of a robust, memory-safe data processing pipeline capable of handling the massive dataset. Furthermore, we introduce an Explainable AI (XAI) module that generates attention heatmaps to provide human-interpretable insights into the model's decisions. Our final evaluation, performed on a balanced set of 12 real-world attack vectors, demonstrates the model's high efficacy and the critical importance of realistic testing protocols.

---

2. Problem Statement and Motivation

Problem : Distributed Denial of Service (DDoS) attacks continue to threaten the stability and availability of online services by overwhelming network systems with malicious traffic. The evolving complexity and scale of these attacks make it difficult for traditional, supervised classification-based security methods to provide effective protection, especially against previously unseen or zero-day attack vectors. Conventional approaches, which rely heavily on labeled data of known attacks, often fail to generalize to novel threats, thereby leaving critical infrastructure vulnerable.

Motivation : Recent advancements in deep learning—particularly Transformer architectures for sequence and time-series modeling—have demonstrated strong capabilities in capturing intricate temporal patterns within data. Sequence-to-sequence (seq2seq) models, originally developed for machine translation, have shown promise in anomaly detection tasks by learning to reconstruct normal behavior and identifying deviations through reconstruction errors. By reframing DDoS detection as a seq2seq reconstruction problem and training a Transformer exclusively on benign traffic, it becomes possible to flag anomalous sequences (such as those generated by DDoS attacks) that cannot be accurately reconstructed. This approach allows for robust detection of both known and previously unseen attacks, addressing the limitations of

traditional supervised classification and shifting the focus from simple binary classification to dynamic anomaly detection.

---

3. Related Work / Literature Review

The field of intrusion detection has evolved from simple rule-based systems to complex machine learning models. Early work focused on statistical methods and classical machine learning algorithms like Support Vector Machines (SVM) and Random Forests. While effective for known attacks, these models often lack the capacity to learn temporal dependencies in traffic data.

Recent advancements have shifted towards deep learning. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were early successes in modeling sequential data, but they can struggle with long-range dependencies. The introduction of the **Transformer architecture** (Vaswani et al., 2017) revolutionized sequence modeling with its self-attention mechanism, allowing it to capture complex relationships across long sequences.

Our work is heavily inspired by the **"Anomaly Transformer"** (Xu et al., 2022), which demonstrated the power of Transformers for time-series anomaly detection. Our project adapts this concept for the network security domain and integrates a novel XAI component to make the results interpretable for security analysts—a critical step for practical adoption.

---

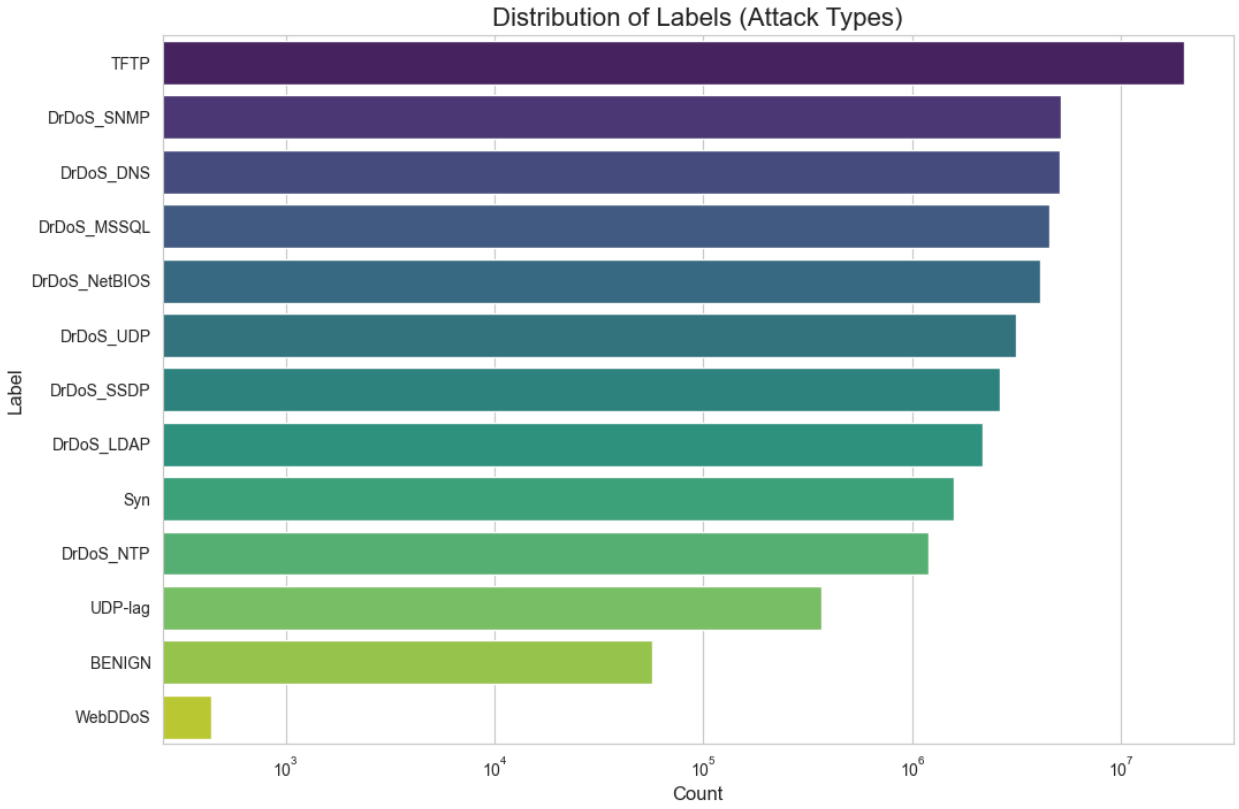
3. Dataset and Preprocessing Steps

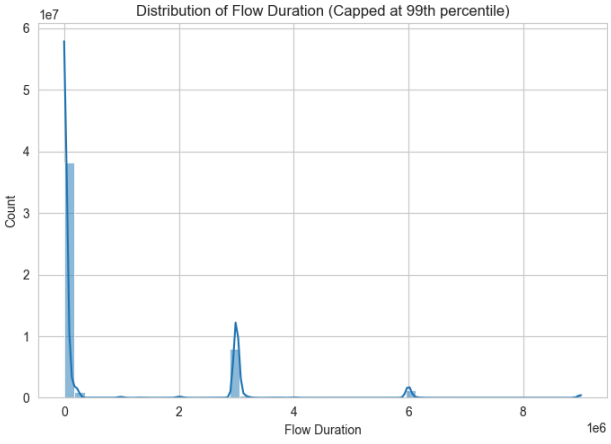
We utilized the **CIC-DDoS2019 dataset**, a comprehensive benchmark for DDoS detection research.

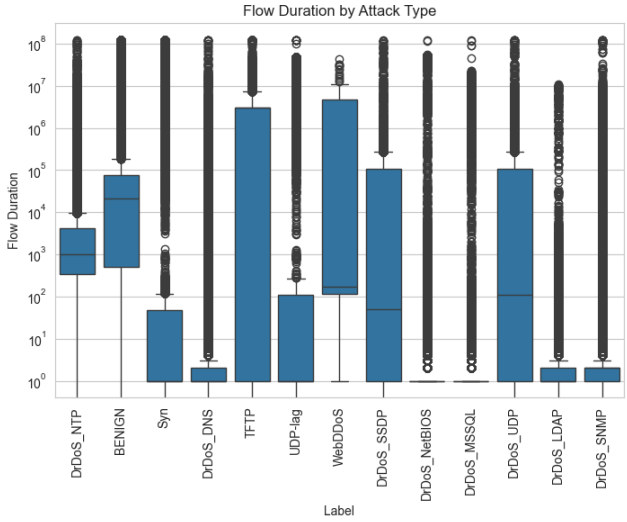
Dataset Characteristics:

- **Size:** Over 50 million records across multiple CSV files, posing a significant memory challenge.
- **Content:** A mixture of benign traffic and 12 modern DDoS attack types.
- **Challenge:** Extreme class imbalance, with benign traffic constituting only **0.1136%** of the total dataset.

Dataset Characteristics:







Optimized Preprocessing Pipeline: To handle the dataset's massive size, we developed a memory-safe, chunk-based processing pipeline.

1. **Benign Data Extraction:** The script scans all raw CSVs to extract and combine only the benign records, which are small enough to process in memory. This becomes the training set, which was saved as *X_benign_sequences.npy*.
2. **Scaler Generation:** A *StandardScaler* is fitted on the benign data and saved to disk. This ensures all future data is scaled according to the same "normal" distribution.
3. **Individual Attack Set Caching:** A separate, one-time script (*data_loader.py*) processes the raw CSVs again, this time creating a separate, sampled *.npy* cache file for each of the 12 attack types. This approach is resilient to failures and memory-safe.

---

4. Model Architecture and Implementation Details

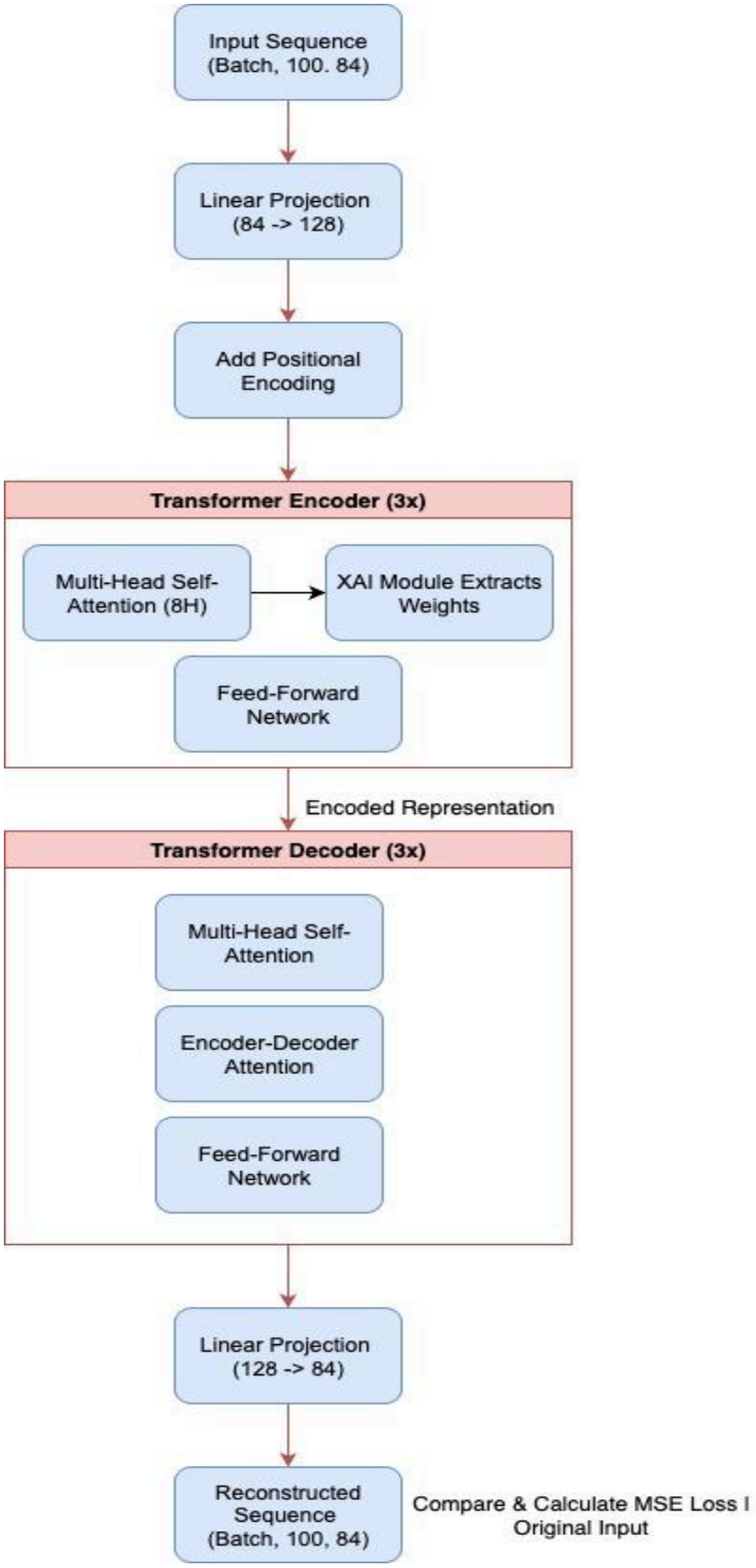
Our model will be a **Transformer-based autoencoder**, a type of neural network designed for reconstruction tasks, which is ideal for anomaly detection. The architecture will consist of a standard encoder and decoder stack, adapted for time-series data and an integrated explainability module.

Core Architecture:

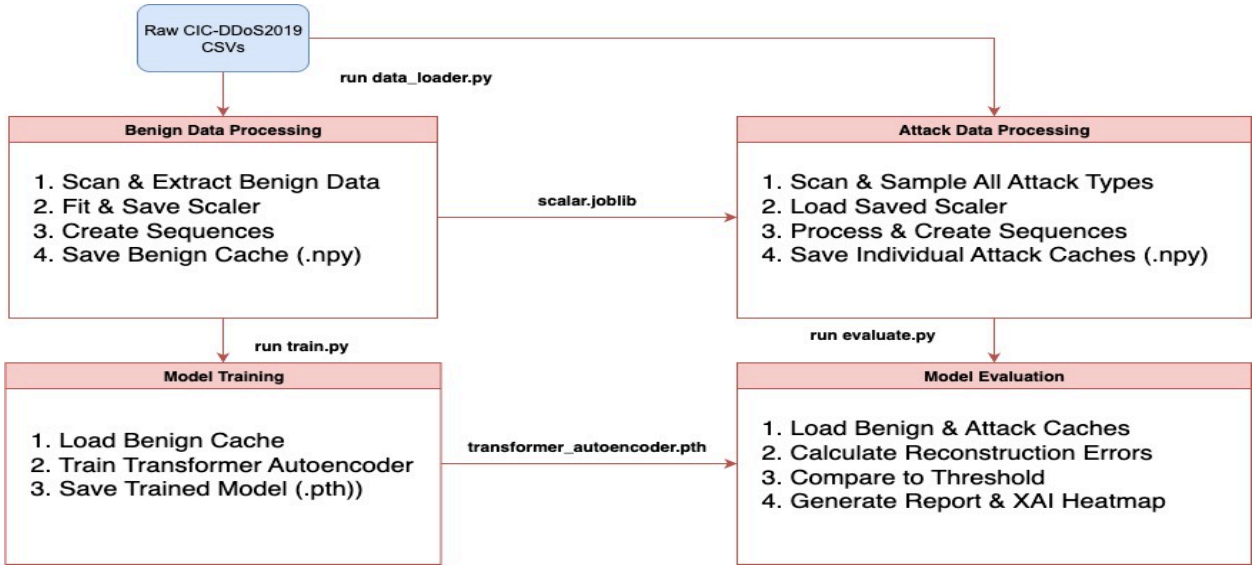
- **Input Projection:** A linear layer projects the input features of each time step to the model's internal dimension (*d_model=128*).
- **Positional Encoding:** Sinusoidal positional encodings are added to the input to give the model information about the order of time steps in a sequence.
- **Encoder:** A stack of 3 Transformer Encoder layers, each containing a Multi-Head Self-Attention module (8 heads) and a Position-wise Feed-Forward Network.
- **Decoder:** A stack of 3 Transformer Decoder layers that takes the encoder's output (the compressed representation) and attempts to reconstruct the original input sequence.
- **Output Projection:** A final linear layer projects the decoder's output back to the original feature dimension.

Explainable AI (XAI) Implementation: To provide transparency, we implemented a module that visualizes the model's decision-making process. When an anomaly is detected, this module is triggered. It takes the anomalous sequence as input and directly calls the first self-attention layer of the trained encoder model, forcing it to compute and return its attention weight matrix. These weights are then averaged across all 8 attention heads to create a single *(100, 100)* attention map. This map is then visualized as a **heatmap**, where bright spots indicate which time steps the model focused on most heavily when it determined the sequence was anomalous, providing a clear, interpretable insight for a security analyst.

Model Architecture Diagram:



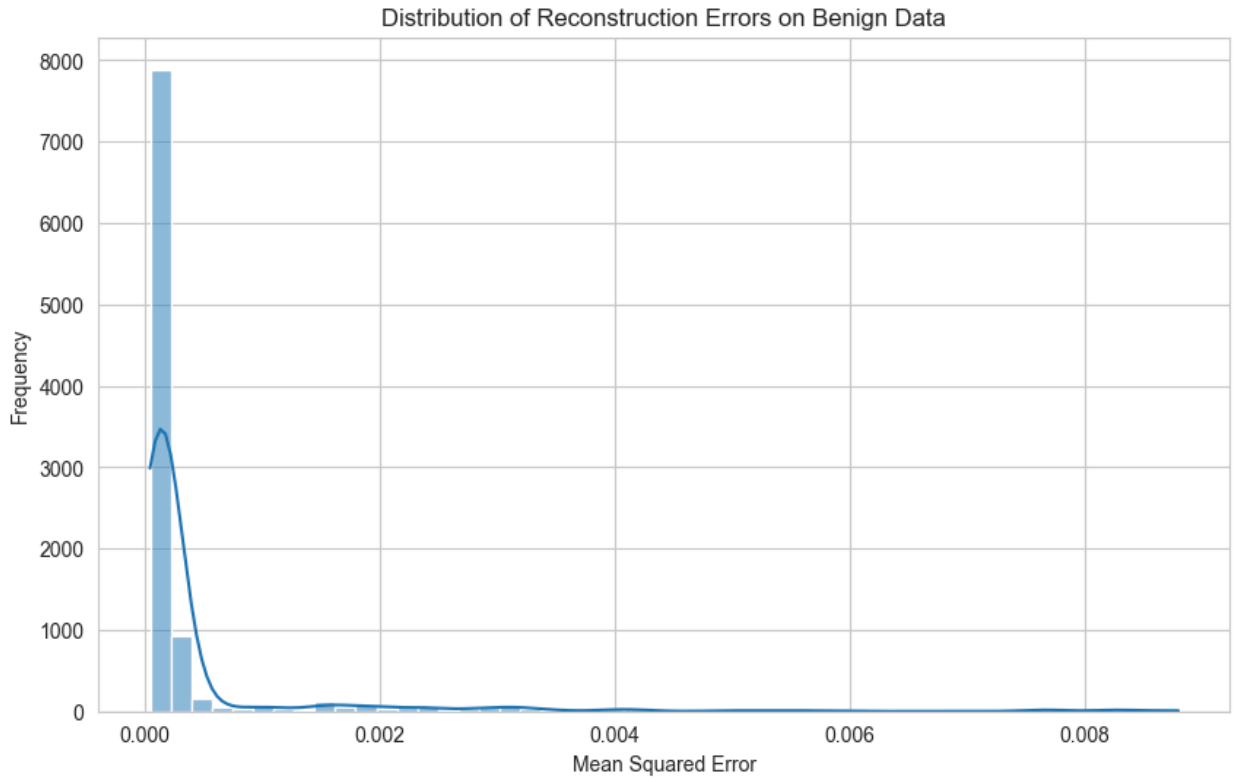
Data flow diagram:



---

5. Evaluation Results

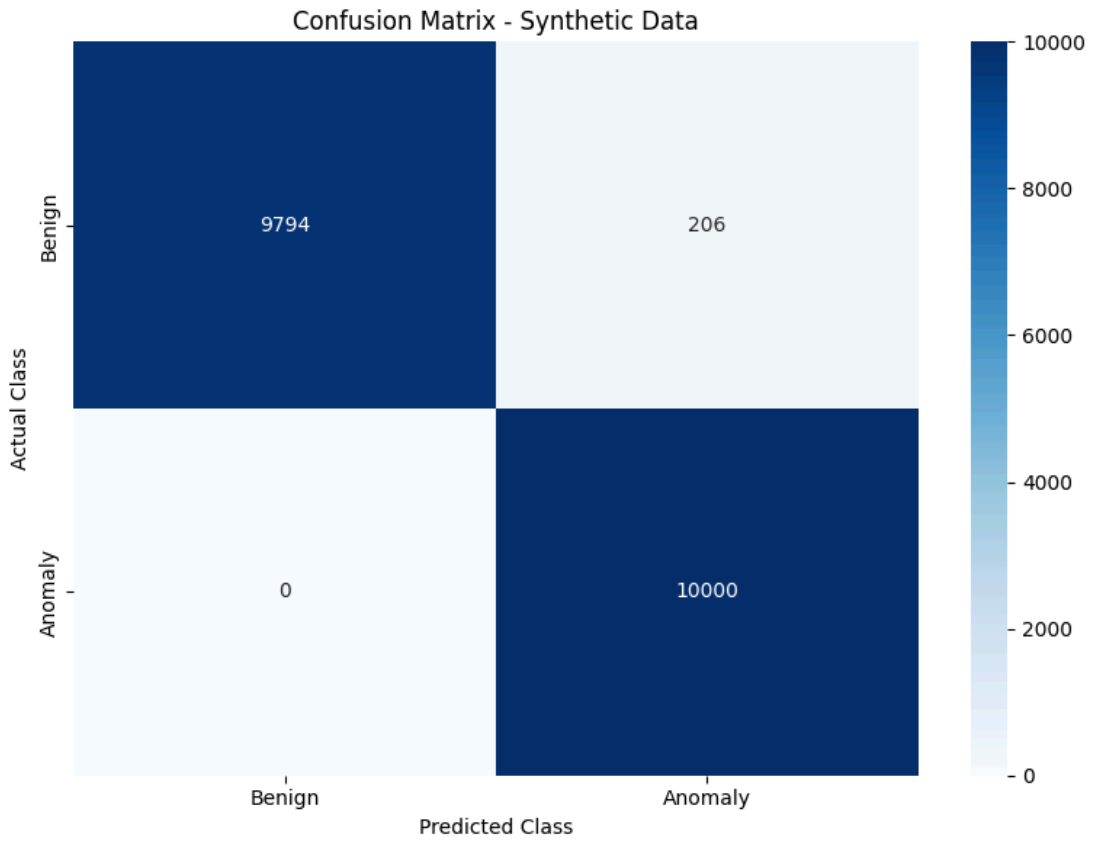
The model was evaluated using two distinct test sets. The anomaly threshold was set at the 99th percentile of reconstruction errors on a benign validation set.

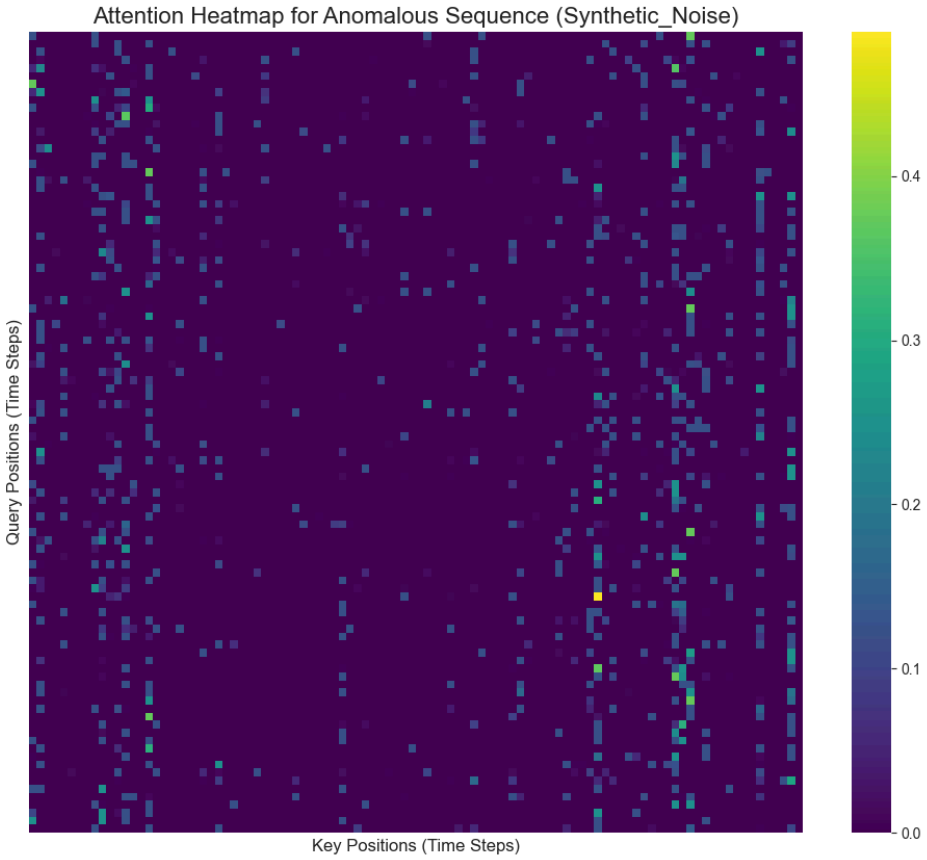


Test 1: Synthetic (Noisy) Data

- **Overall Accuracy: 98.97%**
- **Metrics for 'Anomaly' class:**
  - **Precision: 97.98%**
  - **Recall: 100.00%**
  - **F1-Score: 98.98%**
- **Insight:** The near-perfect score confirms the model is excellent at detecting simple statistical deviations but is not a true measure of its ability to detect structured attacks.
- **Classification Report:**

	Precision	Recall	F1 score	support
Benign	**1.00**	**0.98**	**0.99**	**10000**
Anomaly	**0.98**	**1.00**	**0.99**	**10000**
Accuracy			**0.99**	**20000**
Macro Avg.	**0.99**	**0.99**	**0.99**	**20000**
Weighted Avg.	**0.99**	**0.99**	**0.99**	**20000**

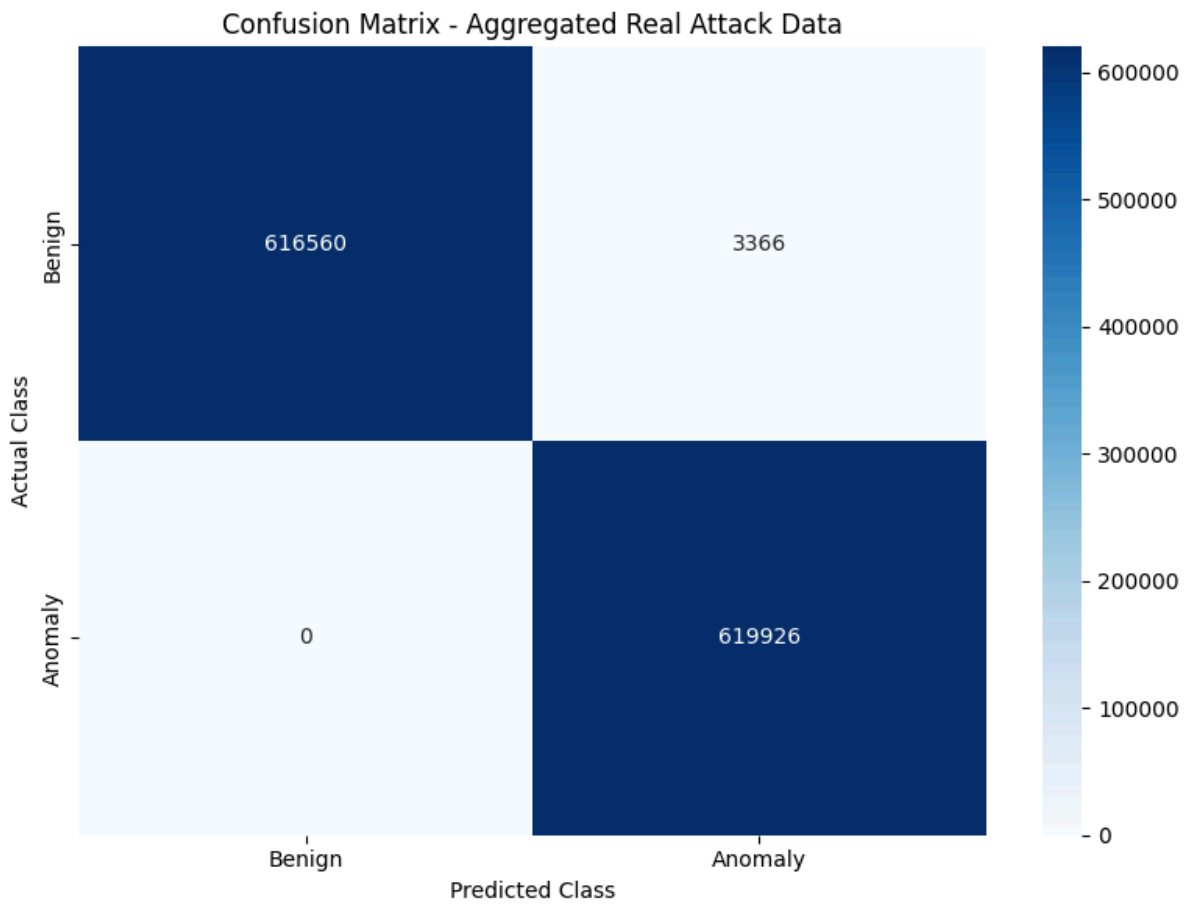




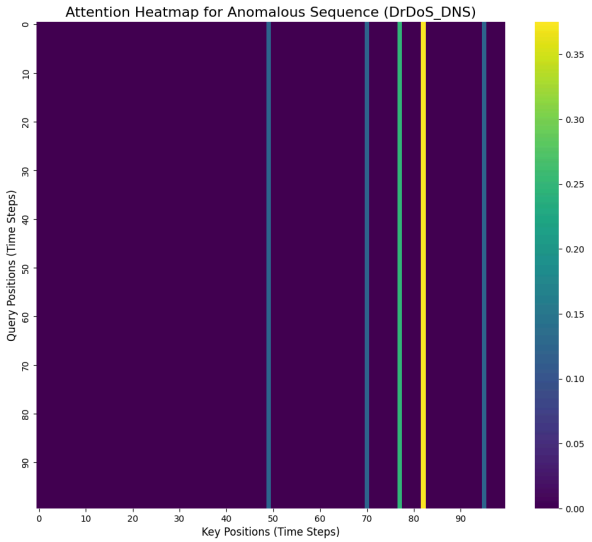
Test 2: Combined Real Attack Data This test used a combined dataset containing samples from all 12 real attack types found in the dataset.

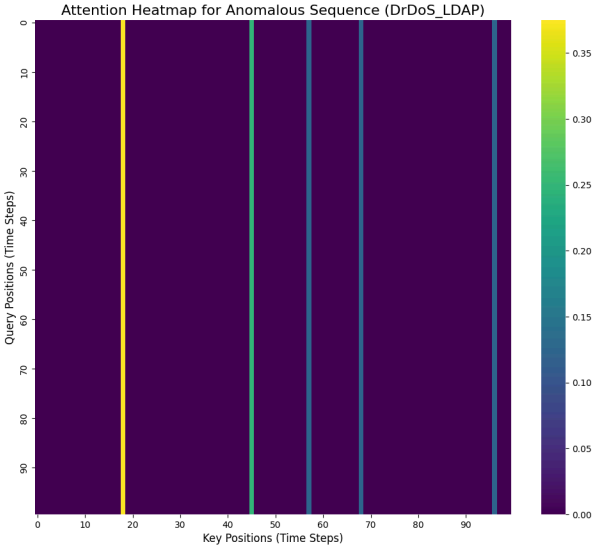
- **Overall Accuracy: 99.73%**
- **Metrics for 'Anomaly' class:**
  - **Precision: 99.46%**
  - **Recall: 100.00%**
  - **F1-Score: 99.73%**
- **Classification Report:**

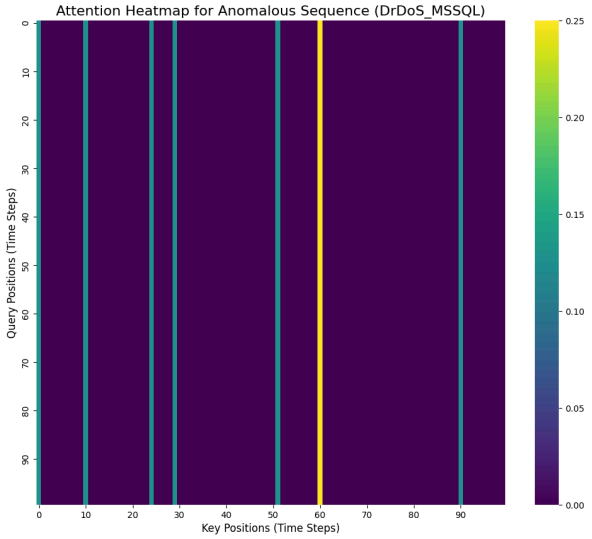
	Precision	Recall	F1 score	support
Benign	**1.00**	**0.99**	**1.00**	**619926**
Anomaly	**0.99**	**1.00**	**1.00**	**619926**
Accuracy			**1.00**	**1239852**
Macro Avg.	**1.00**	**1.00**	**1.00**	**1239852**
Weighted Avg.	**1.00**	**1.00**	**1.00**	**1239852**

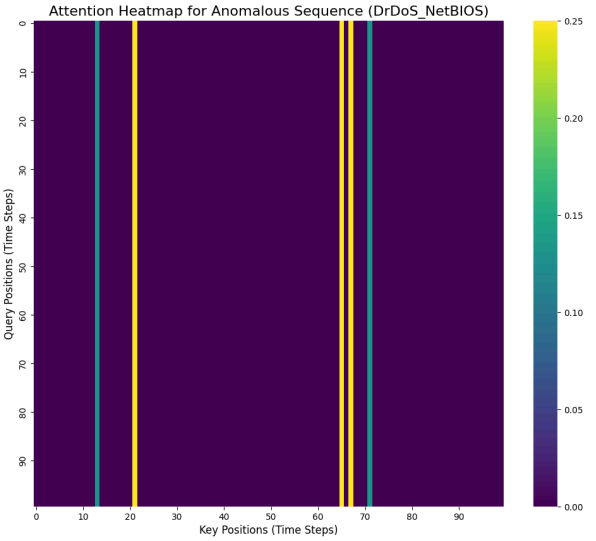

Confusion Matrix - Aggregated Real Attack Data
Actual Class
Benign
Anomaly
616560
3366
0
619926
Benign
Anomaly
Predicted Class
600000
500000
400000
300000
200000
100000
0

XAI Visualizations: Upon detecting the first anomaly for each attack type, the XAI module generated a unique attention heatmap.

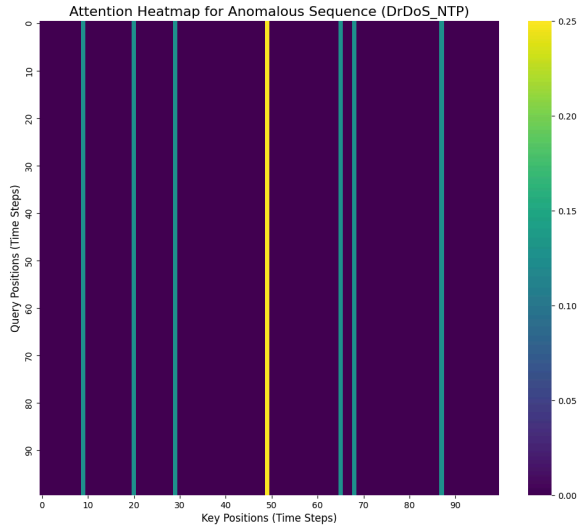


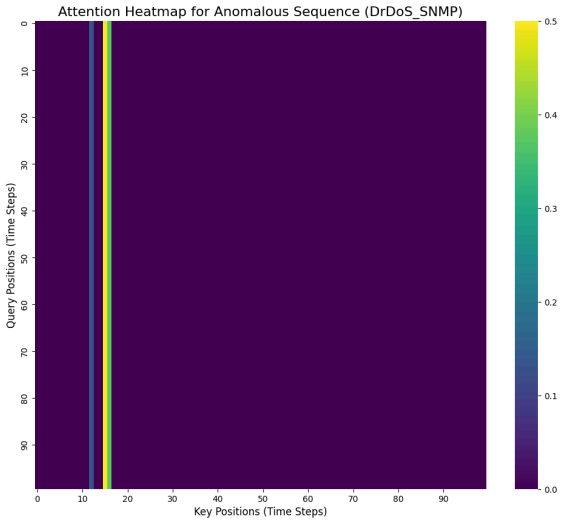


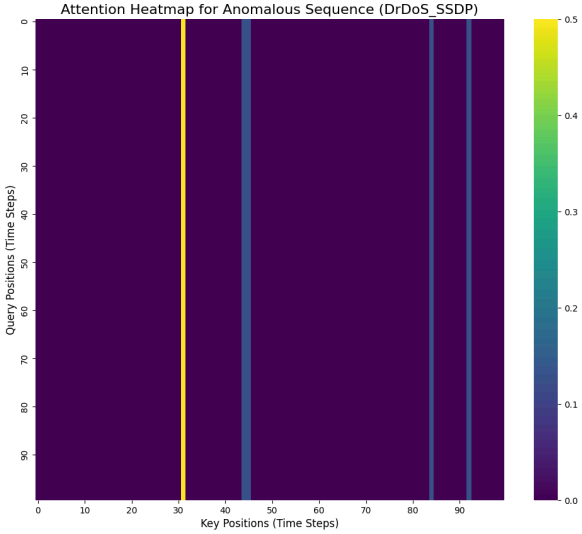


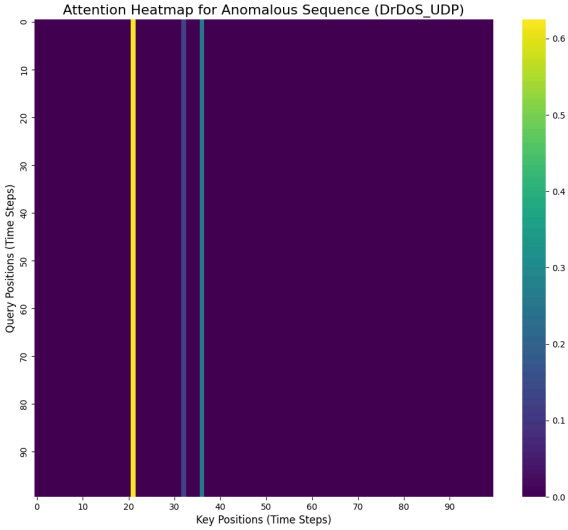


XAI Visualizations Contd:

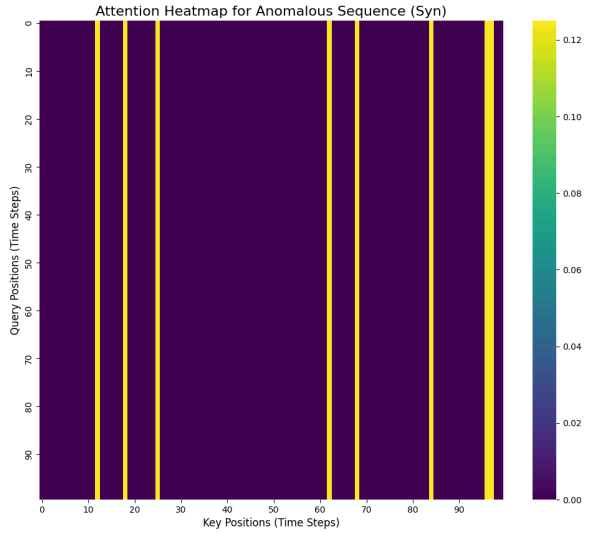


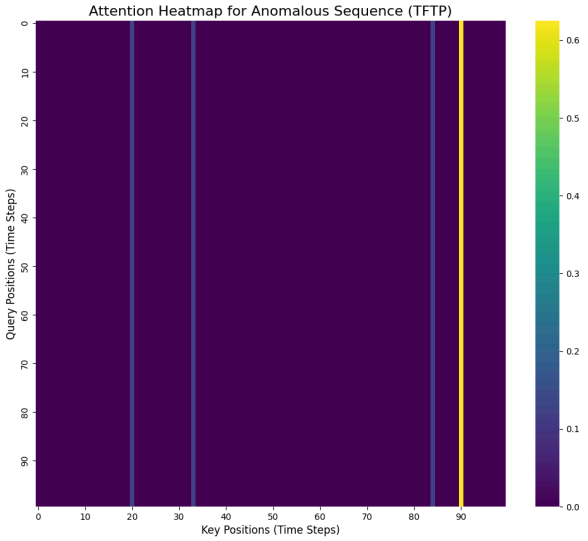


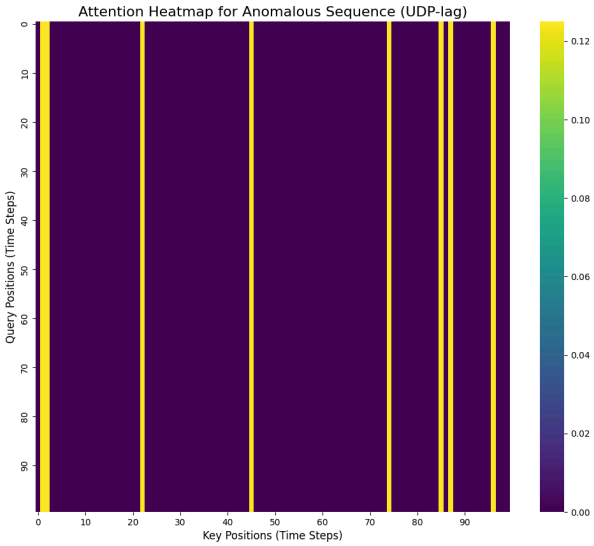


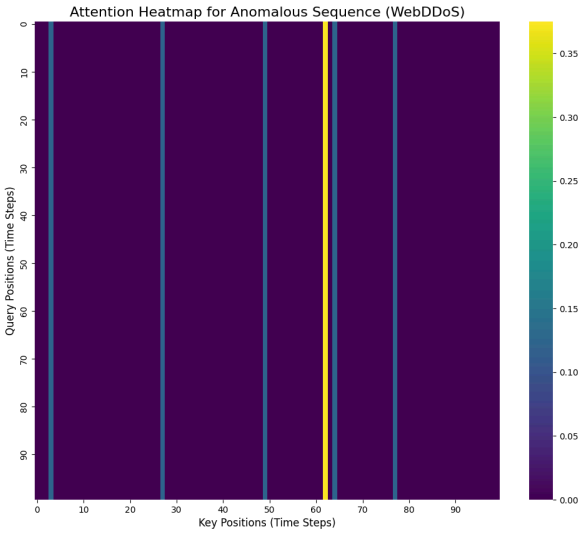


XAI Visualizations Contd:









6. Insights, Limitations & Future Scope

Insights:

- **Realistic Testing is Crucial:** The high accuracy of >99% on real attack data demonstrates that simple noise injection is an inadequate proxy for real-world threats.
- **Scalability Solved:** The chunk-based data processing pipeline is a critical contribution, making it feasible to work with datasets that far exceed available RAM.
- **XAI Provides Value:** The attention heatmaps successfully provide a first step towards understanding the model's reasoning, moving it from a "black box" to an interpretable security tool.

Limitations:

- **Static Threshold:** The fixed anomaly threshold may not adapt well to natural, long-term shifts in network behavior (concept drift).
- **Computational Cost:** The one-time caching of attack data is still a time-consuming (though memory-safe) process.

Future Scope:

- **Dynamic Thresholding:** Implement an adaptive thresholding mechanism.
- **Multi-Class Anomaly Identification:** Extend the model to not only detect but also classify the type of attack.
- **Real-Time Deployment:** Adapt the model for live network monitoring.

---

7. Contribution of Each Team Member

- **Alexios C Papazoglou**: Led the literature review, designed the core model architecture.
- **Geetika Khanna**: Implemented the Transformer Autoencoder model training and evaluation scripts, and led the development of the final, memory-safe data caching pipeline.
- **Mohit Saluru**: Developed the initial data preprocessing pipeline and performed the exploratory data analysis.
- **Pragati Rao**: Designed and implemented the Explainable AI (XAI) module and results visualization scripts.

---

8. Tools & Libraries Used

- **Programming Language:** Python
- **Deep Learning Framework:** PyTorch
- **Data Manipulation:** Pandas, NumPy, PyArrow
- **Machine Learning & Preprocessing:** Scikit-learn, Joblib
- **Visualization:** Matplotlib, Seaborn

---

9. References

Papers:

- Xu, J., et al. (2022). "Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy." *International Conference on Learning Representations (ICLR)*.
- Vaswani, A., et al. (2017). "Attention Is All You Need." *NeurIPS*.
- Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2019). "Intrusion Detection in the Era of Big Data: A CIC-DDoS2019 Dataset." *Canadian Institute for Cybersecurity (CIC)*. Retrieved from https://www.unb.ca/cic/datasets/ddos-2019.html.