**CS 7643 Group Project Proposal**

**Team Name: Data Anomaly Detector**

**Project Title:** Deep Learning Approach for Anomaly Detection of Time Series Data

**Team Members:** Lily Chebotarova, Raga Lasya Munagala, Ligeng Peng, Brian Zhang

1. **Project summary**

Anomaly detection in time series data is an intriguing problem with profound implications across multiple domains. The ability to accurately identify anomalies enables us to proactively detect irregular patterns, diagnose potential issues, and enhance the quality and reliability of various services. In domains like web traffic management, anomalies could indicate potential issues such as cyber-attacks, system failures, or unexpected spikes in user traffic. Timely detection of these anomalies is crucial to allow for prompt response, minimizing any adverse impact on the system's performance and user experience.

Our project revolves around the challenge of accurate anomaly detection in large-scale time series data, specifically targeting a dataset comprising both real and synthetic time series with labeled anomaly points. This dataset serves as a comprehensive testing ground to evaluate the accuracy of anomaly detection for various anomaly types, including outliers and change-points.

The synthetic dataset in our possession features time series with diverse characteristics, encompassing varying trends, noise, and seasonality. On the other hand, the real dataset offers a collection of time series that encapsulate the metrics of various Yahoo services.

1. **Approach**

Our core approach revolves around employing Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM), to model time series data for anomaly detection. Time series data inherently exhibit both long-term and short-term trends, like seasonal patterns or weekly variations, which make LSTM models a natural fit for capturing such temporal intricacies. The ability of these models to capture long-term dependencies in time-series data also makes them more adaptable to evolving data patterns, which is crucial for accurately identifying anomalies over time.

Our project will kick off with meticulous data preprocessing. This phase involves addressing missing data points by either imputing them suitably or removing them. Additionally, we intend to remove noise and normalize features as part of preprocessing.

Based on our preliminary research, there are multiple approaches that have been tested on time series data for anomaly detection. We will start by replicating some of the most successful existing approaches such as LSTM and autoencoder to reproduce the results in [4] for the available dataset.

As a stretch goal, we want to explore possible improvements with more novel approaches to anomaly detection such as Generative Adversarial Networks.

1. **Datasets**

The dataset we will be working with denoted as “S5 - A Labeled Anomaly Detection Dataset, version 1.0(16M)”, serves as an excellent starting point for our project. It offers a diverse range of real and synthetic time-series data, specifically curated for the purpose of anomaly detection.

The dataset tests the detection accuracy of various anomaly-types including outliers and change-points. The synthetic dataset consists of time-series with varying trend, noise and seasonality. The real dataset consists of time-series representing the metrics of various Yahoo services.

The dataset is publicly accessible through the following link:

<https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70>

1. **Resources/Related Work:**

[1] R. Chalapathy and S. Chawla, “Deep Learning for Anomaly Detection: A Survey.” arXiv, Jan. 2019. doi: [10.48550/arXiv.1901.03407](https://doi.org/10.48550/arXiv.1901.03407).

[2] T. Hagemann and K. Katsarou, “Reconstruction-based anomaly detection for the cloud: A comparison on the Yahoo! Webscope S5 dataset,” in *Proceedings of the 2020 4th International Conference on Cloud and Big Data Computing*, in ICCBDC ’20. New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 68–75. doi: [10.1145/3416921.3416934](https://doi.org/10.1145/3416921.3416934).

[3] T.-Y. Kim and S.-B. Cho, “Web traffic anomaly detection using C-LSTM neural networks,” *Expert Systems with Applications*, vol. 106, pp. 66–76, Sep. 2018, doi: [10.1016/j.eswa.2018.04.004](https://doi.org/10.1016/j.eswa.2018.04.004).

[4] L. Wong, D. Liu, L. Berti-Equille, S. Alnegheimish, and K. Veeramachaneni, “AER: Auto-Encoder with Regression for Time Series Anomaly Detection.” arXiv, Dec. 2022. doi: [10.48550/arXiv.2212.13558](https://doi.org/10.48550/arXiv.2212.13558).

[5] Z. Kang, A. Mukhopadhyay, A. Gokhale, S. Wen, and A. Dubey, “Generative Anomaly Detection for Time Series Datasets.” arXiv, Jun. 2022. doi: [10.48550/arXiv.2206.14597](https://doi.org/10.48550/arXiv.2206.14597).

[6] R. Raturi, A. Kumar, N. Vyas, and V. Dutt, “A Novel Approach for Anomaly Detection in Time-Series Data using Generative Adversarial Networks,” in 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Jun. 2023, pp. 1352–1357. doi: [10.1109/ICSCSS57650.2023.10169365](https://doi.org/10.1109/ICSCSS57650.2023.10169365).

[7] K. Yoshihara and K. Takahashi, “A simple method for unsupervised anomaly detection: An application to Web time series data,” *PLOS ONE*, vol. 17, no. 1, p. e0262463, Jan. 2022, doi: [10.1371/journal.pone.0262463](https://doi.org/10.1371/journal.pone.0262463).

[8] “Design of a Large-Scale Anomaly Detection Algorithm for Time Series in the Context of Wireless Network Operations - ProQuest.”: <https://www.proquest.com/openview/30ab3da21157f8c9c667d18d1427f708/>