Anomaly detection in Time series data using Transformers

Mini-Project Synopsis

submitted to

Mr. Sudarsan N S Acharya

Assistant Professor

Manipal School of Information Sciences, MAHE, Manipal

Reg. Number	Name	Branch
241057022	**Samarth Prathap M**	**AI-ML**
241057015	**Sanjana Y Alva**	**AI-ML**

DD/MM/YYYY

MANIPAL SCHOOL OF INFORMATION SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)

Introduction

In today's industrial world, especially in important sectors like oil and gas, it is very important to continuously monitor machinery and processes to ensure efficiency, safety, and reliability. In various industries, particularly on oil rigs, sensor networks are being widely used to gather real-time data on key factors like temperature, pressure, and flow rates. However, as the amount of data increases, analysing it becomes more difficult. Detecting abnormalities in these data streams is crucial because they can indicate problems like equipment failures, safety risks, or inefficiencies.

Traditionally, detecting anomalies has depended a lot on labelled datasets. But in many real-world situations, getting such labelled data can be tough and costly. Moreover, traditional methods often find it hard to handle the complex and high-dimensional nature of time series data, especially the kind produced by industrial sensors on oil rigs. This project aims to explore how Transformer models, which are known for their success in natural language processing, can be used to detect anomalies in time series data.

Transformers are well-suited for this task because they can understand long-term patterns and relationships in sequences. The goal of this project is to develop a strong anomaly detection system that does not rely on labelled data, using unsupervised learning methods. This system will aim to spot even the smallest and most complex irregularities in sensor data, making it a crucial tool for preventive maintenance, fault detection, and improving operational efficiency.

This project intends to show how Transformer models can be useful in a new field, providing insights into how they can be adjusted for non-labelled time series data, and paving the way for their practical use in industrial settings, both in India and beyond, particularly in the oil and gas industry.

Project Objective

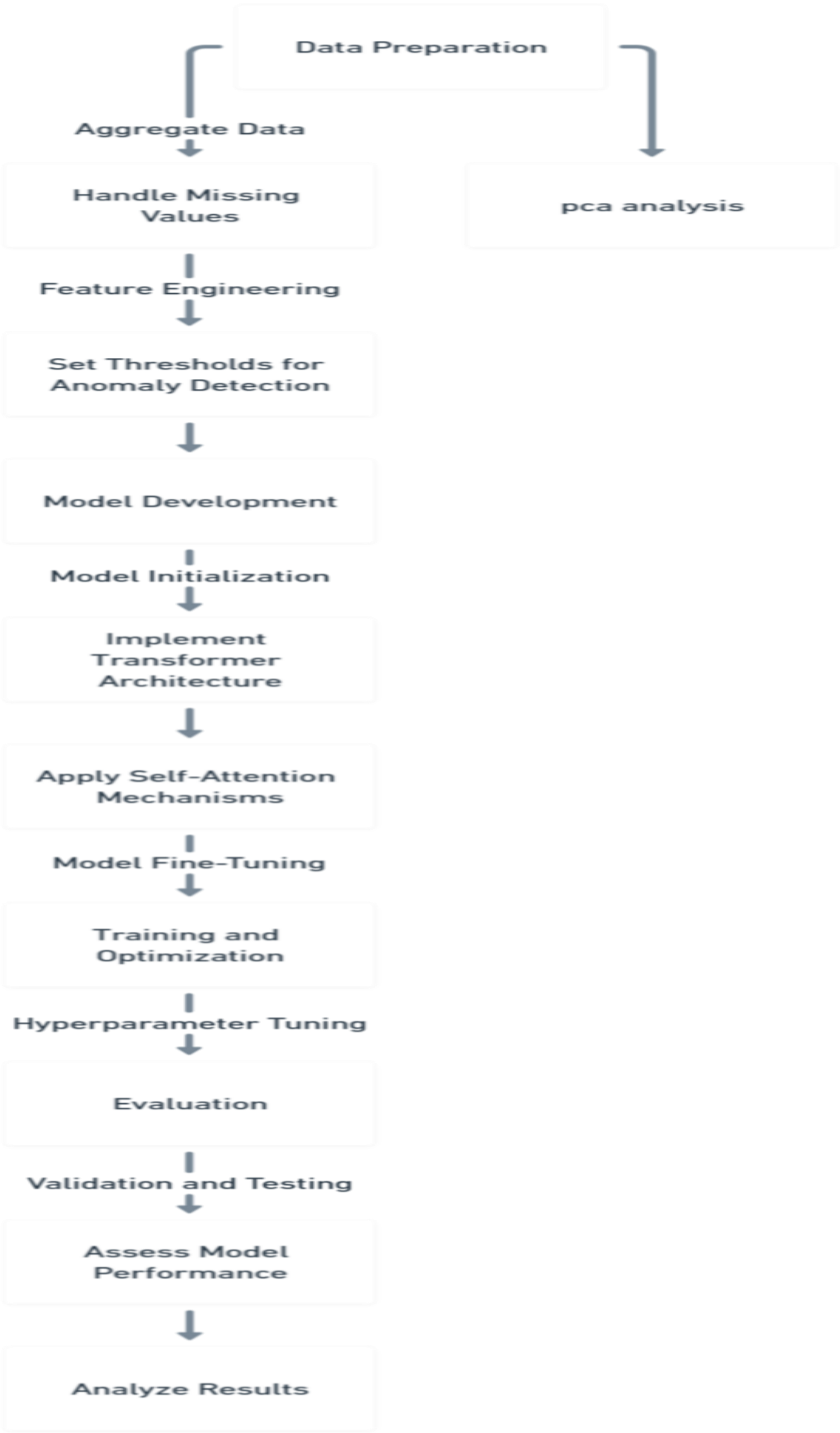
1. Data Preparation: Aggregate 5-minute interval data to a 30-minute time frame, handle missing values using interpolation, and set thresholds for anomaly detection.

2. Model Development: Implement a Transformer model with encoder-decoder architecture and self-attention mechanisms to capture and reconstruct sequence patterns.

3. Training and Optimization: Train the model to recognize normal patterns and identify anomalies, fine-tune it for improved performance, and optimize hyperparameters using grid search.

4. Evaluation: Assess the model's anomaly detection capability using reconstruction error and other unsupervised metrics, given the absence of labelled data.

•



General outline of the working pipeline

Pipeline Stages*:*

1. **Data Preparation:**
    - Aggregate data from various sources.
    - Handle missing values in the data.
    - Perform Principal Component Analysis (PCA) for dimensionality reduction.
2. **Feature Engineering:**
    - Create or transform features to improve model performance.
3. **Set Thresholds for Anomaly Detection**:
    - Define criteria to distinguish between normal and anomalous data points.
4. **Model Development:**
    - Initialize the model.
    - Implement a Transformer architecture.
    - Apply self-attention mechanisms within the Transformer.
    - Fine-tune the model.
5. **Training and Optimization:**
    - Train the model on the prepared data.
    - Optimize model parameters through hyperparameter tuning.
6. **Evaluation:**
    - Validate and test the model's performance.
7. **Assess Model Performance:**
    - Analyse the model's results and evaluate its effectiveness.

Our Understanding of the data

The dataset has 377,719 entries and 7 columns.

It has the following Columns:

Time: Timestamps, though stored as strings.

Cyclone_Inlet_Gas_Temp: Temperature of the inlet gas (many entries seem to have values labelled as "Not Connect").

Cyclone_Material_Temp: Temperature of the material.

Cyclone_Outlet_Gas_draft: Draft pressure at the outlet gas.

Cyclone_cone_draft: Draft pressure at the cone

Cyclone_Gas_Outlet_Temp: Temperature of the outlet gas.

Cyclone_Inlet_Draft: Draft pressure at the inlet.

The sensors data are recorded every 5 minutes, it is observed that inlet gas temperature has no data sometimes, so we might need to handle those missing parts of data.

Here, Draft pressure refers to the pressure difference between two points in a system, often used to describe the movement of gases or air within industrial processes. It’s crucial for ensuring the proper flow and removal of gases or combustion byproducts.

- **Cyclone_Outlet_Gas_draft:** Measures the draft pressure at the outlet of the cyclone where the gas exits. This helps monitor the exhaust conditions and the efficiency of gas removal.
- **Cyclone_cone_draft:** Measures the draft pressure at the cyclone's cone section. This can indicate how well the cyclone is managing the flow of material and gas.
- **Cyclone_Inlet_Draft:** Measures the draft pressure at the inlet where the gas or air enters the cyclone. This helps assess the intake conditions and overall airflow into the cyclone.

Why choose transformers?

Transformers are particularly suited for tasks like anomaly detection in time series data due to several key features like:

1. Self-Attention Mechanism

- **Focus on Relevant Information:** The self-attention mechanism allows the model to weigh different parts of the sequence differently, focusing more on the relevant time steps for detecting anomalies. This is crucial for capturing dependencies and patterns over long sequences, which can be indicative of anomalies.
- **Contextual Understanding:** It enables the model to understand the context of each time step in relation to others, improving its ability to distinguish between normal and abnormal behavior based on complex patterns.

2. Handling Long Sequences

- **Global Context:** Unlike traditional RNNs or LSTMs that may struggle with long-range dependencies due to vanishing gradients, Transformers can process entire sequences at once. This is beneficial for time series data where dependencies might span long periods.
- **Efficient Computation:** Transformers use parallel processing for sequences, which speeds up training and inference compared to sequential models.

3. Encoder-Decoder Architecture

- **Effective Representation:** The encoder-decoder structure is effective for sequence-to-sequence tasks. In anomaly detection, the encoder can capture the normal behavior of the time series, while the decoder reconstructs the sequence. Deviations in reconstruction can signal anomalies.
- **Learning Complex Patterns:** This architecture allows the model to learn complex patterns and relationships in the data, making it better at identifying subtle anomalies.

4. Multi-Head Attention

- **Diverse Pattern Recognition:** Multi-head attention allows the model to focus on different aspects of the data simultaneously. This helps in capturing various types of patterns and dependencies, enhancing the model's ability to detect anomalies that might not be apparent from a single perspective.