**A**

**PROJECT SCHOOL REPORT**

**ON**

**ANOMALY TRANSFORMER:** **Time series anomaly detection with association discrepancy**

**Submitted By**

**D. Raghuchandra 245322733011**

**N. Hemanth reddy 245322733042**

**T. Abhiram reddy 245322733061**

**G. Shashi preetham 245322733014**

**M. Dinesh 245322733036**

**Under the guidance** **of**

**Ramakrishna kuppa**

**Designation, Department**



# NEIL GOGTE INSTITUTE OF TECHNOLOGY

Kachavanisingaram Village, Hyderabad, Telangana 500058.

**January, 2025**



# NEIL GOGTE INSTITUTE OF TECHNOLOGY

A Unit of Keshav Memorial Technical Education (KMTES)

Approved by AICTE, New Delhi & Affiliated to Osmania University, Hyderabad

# CERTIFICATE

*This is to certify that the project work entitled* “**ANOMALY TRANSFORMER”** *is a bonafide work carried out by* **“D. Raghuchandra , N. Hemanth reddy , T. Abhiram reddy , G. Shashi preetham, M. Dinesh”** of III year V semester **Bachelor of Engineering** *in* **CSE** *during the academic year* **2024-2025 and** *is a record of bonafide work carried out by them*.

## Project Mentor

Ramakrishna kuppa

Professor

**ABSTRACT**

Anomaly detection in time series data is a vital task in various fields, including finance, healthcare, and industrial systems. Traditional methods often fail to capture the complex dependencies and contextual relationships within data, resulting in suboptimal anomaly detection. This project leverages the Anomaly Transformer, a novel approach that introduces the concept of association discrepancy to address these limitations.

Association discrepancy measures the divergence between temporal associations, enabling the model to distinguish normal patterns from anomalies effectively. By focusing on contextual learning and association modeling, the Anomaly Transformer demonstrates superior accuracy in identifying anomalies compared to traditional methods. This work contributes to enhancing anomaly detection techniques and lays the foundation for future research in scalable and efficient solutions for time series analysis.

**CONTENTS**

| **S. NO.** | **TITLE** | | **PAGE NO.** |
| --- | --- | --- | --- |
|  | **ABSTRACT**  **TABLE OF CONTENTS**  **LIST OF FIGURES**  **LIST OF TABLES** | | i  ii  iii  iv |
| **1** | **Introduction**  1.1 Problem statement  1.2 Project objectives  1.3 Applications | | 7  7  8 |
| **2** | **Literature Survey** | |  |
|  | 2.1 | Existing methods | 8 |
|  | 2.2 | Research gap | 10 |
|  |  |  |  |
| **3** | **Proposed Work, Architecture, Technology Stack & Implementation Details** | |  |
|  | 3.1  3.2  3.3  3.4 | Proposed Work  Approach  Technology stack  Implementation details | 11  11  15  16 |
| **4** | **Results & Discussions** | |  |
|  | 4.1 |  |  |
|  | 4.2 |  |  |
| **5** | **Conclusion & Future Scope** | |  |
| **6** | **References** | |  |

**LIST OF FIGURES**

| **Fig. No.** | **Figure Name** | **Page No.** |
| --- | --- | --- |
|  |  |  |
| 1 | Association discrepancy | 12 |
| 2 | Min max strategy | 13 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**LIST OF TABLES**

| **Table No.** | **Table Name** | **Page No.** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. **INTRODUCTION**

**1.1 problem statement**

In many fields, data is collected continuously over time, creating what is known as time series data. Examples include stock prices in the financial sector, sensor readings in industrial systems, patient health monitoring in hospitals, and user activity logs in computer systems. Within this data, certain unusual patterns, known as anomalies, may indicate critical events such as fraud, equipment failure, health emergencies, or security breaches.

Detecting these anomalies is essential because they often require immediate attention. However, traditional methods for anomaly detection rely on simple statistical models or predefined rules, which often work poorly for complex datasets. These methods struggle to capture the relationships between data points over time, especially in cases where multiple variables are involved. For example, in monitoring a machine's performance, factors like temperature, pressure, and vibration are interconnected, and analyzing them in isolation can miss important insights.

As a result, traditional methods are prone to errors like false positives (identifying normal patterns as anomalies) or false negatives (failing to detect true anomalies). This creates inefficiencies and risks, making it clear that more advanced approaches are needed to analyze time series data effectively and improve anomaly detection accuracy.

**1.2 Project objectives**

The primary goal of this project is to develop a robust anomaly detection system that overcomes the limitations of traditional methods. By leveraging the advanced capabilities of the Anomaly Transformer model, the project seeks to achieve the following objectives:

Accurate Anomaly Detection: To detect anomalies in time series data with high precision, minimizing false positives and false negatives.

Utilizing Association Discrepancy: To incorporate the concept of association discrepancy, which measures the deviation in relationships between data points across different time contexts. This approach allows the model to distinguish between normal and anomalous patterns effectively.

Handling Complex Datasets: To analyze multivariate time series data, capturing the interdependencies between variables.

Scalability: To ensure that the proposed system can handle large-scale time series data without compromising on performance.

Performance Analysis: To evaluate the model’s accuracy and efficiency by comparing its results with those of traditional anomaly detection techniques.

**1.3 Applications**

The ability to detect anomalies accurately has significant implications across numerous domains, including:

Finance: Monitoring stock markets for unusual price movements, detecting fraudulent activities in transactions, and managing risk in financial systems.

Healthcare: Identifying irregularities in patient monitoring systems, such as sudden changes in heart rate, blood pressure, or oxygen levels, which may indicate critical conditions.

Industry: Predicting equipment failures by analyzing sensor data, such as temperature, pressure, and vibration, to reduce downtime and maintenance costs.

Cybersecurity: Detecting unauthorized access, unusual login attempts, or abnormal data usage patterns in real-time to prevent security breaches.

Weather and Environment: Monitoring weather patterns to detect storms, floods, or other extreme conditions, and analyzing environmental data for changes in air or water quality.

Retail and E-commerce: Analyzing customer behavior to identify suspicious purchase patterns or inventory discrepancies.

The insights gained through improved anomaly detection can lead to better decision-making, more efficient resource allocation, and timely actions in critical scenarios. By addressing these applications, this project demonstrates the broad and impactful utility of advanced anomaly detection techniques.

**2. LITERATURE SURVEY**

**2.1 Existing Methods**

Anomaly detection in time series data has been approached using various methods, including density-estimation, clustering, reconstruction-based models, and autoregression-based techniques. Below are some notable methods in each category:

Density-Estimation Methods:

The Local Outlier Factor (LOF) (Breunig et al., 2000) and Connectivity Outlier Factor (COF) (Tang et al., 2002) calculate the local density and connectivity of data points to identify outliers. These methods work well for detecting anomalies in datasets where the outliers have different densities compared to the majority of data points.

DAGMM (Zong et al., 2018) and MPPCACD (Yairi et al., 2017) integrate the Gaussian Mixture Model (GMM) to estimate the density of data representations, which helps in identifying anomalies based on how likely a data point is in relation to the overall data distribution.

Clustering-Based Methods:

In clustering-based methods, anomaly scores are typically determined by the distance to the cluster center.

Support Vector Data Description (SVDD) (Tax & Duin, 2004) and Deep SVDD (Ruff et al., 2018) focus on gathering normal data representations into compact clusters, where points that deviate from these clusters are classified as anomalies.

THOC (Shen et al., 2020) employs hierarchical clustering to fuse multi-scale temporal features and detects anomalies based on multi-layer distances.

ITAD (Shin et al., 2020) performs clustering on decomposed tensors, making it suitable for detecting anomalies in multi-dimensional time series data.

Reconstruction-Based Models:

Reconstruction-based approaches detect anomalies based on the error between the original and reconstructed data.

LSTM-VAE (Park et al., 2018) combines Long Short-Term Memory (LSTM) networks for temporal modeling with Variational Autoencoders (VAE) for data reconstruction.

OmniAnomaly (Su et al., 2019) extends the LSTM-VAE model by incorporating normalizing flow for better modeling and uses reconstruction probabilities to detect anomalies.

InterFusion (Li et al., 2021) improves upon previous models by using a hierarchical VAE, allowing for the modeling of inter- and intra-dependencies among multiple time series simultaneously.

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are also used in reconstruction-based anomaly detection (Schlegl et al., 2019; Li et al., 2019a; Zhou et al., 2019), where the adversarial loss helps regularize the model and detect outliers.

Autoregression-Based Models:

Vector AutoRegression (VAR) (Anderson & Kendall, 1976) extends ARIMA by modeling the lag-dependent covariance between time series and predicting future values.

Autoregressive models, such as LSTMs, can also be applied for time series prediction (Hundman et al., 2018; Tariq et al., 2019), where anomalies are detected by the prediction error.

The traditional methods mentioned above have their limitations, particularly in handling complex patterns and large-scale data. In contrast to random walk or subsequence-based anomaly detection methods (Cheng et al., 2008; Boniol & Palpanas, 2020), the Anomaly Transformer (our reference model) introduces a new association-based criterion. This criterion integrates temporal models to learn more informative time-point associations and effectively captures complex dependencies in time series data.

**Bottom of Form**

**2.2 Research gap**

Despite the success of traditional methods in time series anomaly detection, they often fail to effectively address the challenges posed by large-scale, high-dimensional, and complex datasets. Traditional techniques such as density-estimation methods (e.g., LOF and COF), clustering-based methods (e.g., SVDD and Deep SVDD), and autoregressive models (e.g., VAR and LSTMs) can struggle with the following limitations:

Scalability: As the volume and dimensionality of time series data increase, many traditional methods become inefficient. For instance, clustering-based methods like SVDD often require substantial computational resources, especially when working with large datasets.

Complexity of Dependencies: Many traditional methods, such as LSTMs and VAR models, can capture short-term dependencies but often fail to model long-range and intricate relationships within the data. These models are particularly challenged in identifying subtle anomalies that emerge from long-term temporal dependencies.

Limited Flexibility: Traditional models usually rely on predefined assumptions, such as stationarity in autoregressive models or the normality of data in density-estimation methods. These assumptions can limit their ability to generalize across diverse types of time series data.

The limitations highlighted above present a significant research gap that calls for more flexible, scalable, and powerful anomaly detection techniques capable of handling complex data patterns. Transformers, which excel in modeling long-range dependencies and complex relationships, provide a promising avenue to overcome these challenges. Their self-attention mechanism allows them to capture both local and global temporal dependencies, making them particularly effective for anomaly detection in high-dimensional, multivariate time series data.

Our research aims to fill this gap by applying Transformer models to time series anomaly detection. The Anomaly Transformer, which leverages self-attention and a novel Association Discrepancy approach, has shown the potential to outperform traditional models in this domain

**3. Proposed Work, Approach, Technology Stack & Implementation Details**

**3.1 Proposed work**

The proposed work focuses on leveraging Transformer models for time series anomaly detection. Traditional methods have shown limitations in accurately detecting anomalies in complex, high-dimensional time series data due to their inability to capture long-range dependencies and non-linear relationships. Our approach addresses these limitations by using a Transformer architecture with an innovative Association Discrepancy criterion to effectively detect subtle anomalies in multivariate time series.

Key features of the proposed work:

Self-attention Mechanism: The Transformer model's self-attention mechanism captures dependencies over different time steps, allowing it to model both short-term and long-term relationships in time series data.

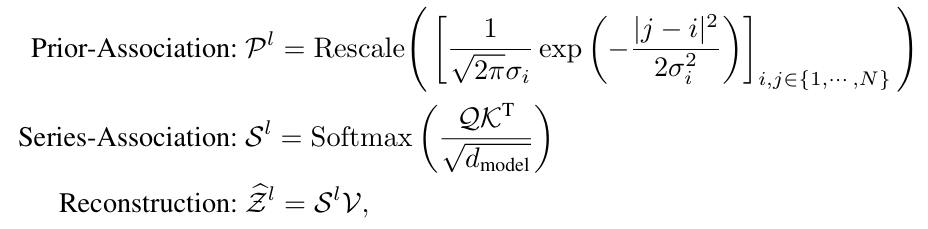
Association Discrepancy: This novel approach enhances the detection of anomalies by focusing on learning associations between different time points, rather than relying solely on pointwise differences.

Scalability: The Transformer model is well-suited for large-scale, high-dimensional datasets, making it scalable for real-world applications.

Our method improves the sensitivity of anomaly detection, reduces false positives, and handles complex temporal relationships more effectively than traditional models.

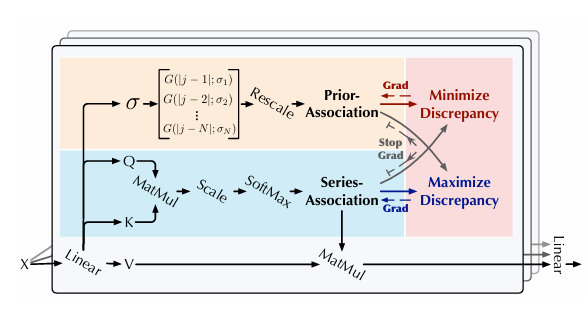
**3.2 Approach**

**Anomaly-Attention Mechanism**

The Anomaly-Attention mechanism is a novel component of the Anomaly Transformer designed to effectively model both local and global temporal dependencies in time-series data. Unlike traditional self-attention mechanisms, which focus solely on global patterns, Anomaly-Attention employs a two-branch structure to capture distinct aspects of temporal relationships. The Prior-Association branch uses a learnable Gaussian kernel to emphasize adjacent time points, reflecting the continuity inherent in normal time-series patterns. On the other hand, the Series-Association branch leverages the self-attention mechanism to dynamically identify long-range dependencies across the entire time series. By comparing these two associations, the mechanism calculates an Association Discrepancy, which quantifies the divergence between local and global temporal patterns. This discrepancy serves as a key indicator for distinguishing normal points from anomalies, as anomalies tend to deviate significantly from typical association patterns. Through this dual perspective, the Anomaly-Attention mechanism provides a robust framework ****for detecting anomalies in complex temporal datasets.

**Association Discrepancy**

The Association Discrepancy is a core concept in the Anomaly Transformer model, introduced to quantify the difference between two key forms of temporal associations: Prior-Association and Series-Association. These two associations represent different perspectives of temporal dependencies within a time series. The Prior-Association models local temporal relationships using Gaussian kernels, while the Series-Association leverages global temporal patterns through the self-attention mechanism.



The calculation of Association Discrepancy is mathematically defined as:

AssDis(P, S; X) = [KL(P(i,:)L || S(i,:)L) + KL(S(i,:)L || P(i,:)L)](i=1,...,N)

where KL is the KL divergence computed between two discrete distributions corresponding to every row of P and S.

**Minimax Strategy**

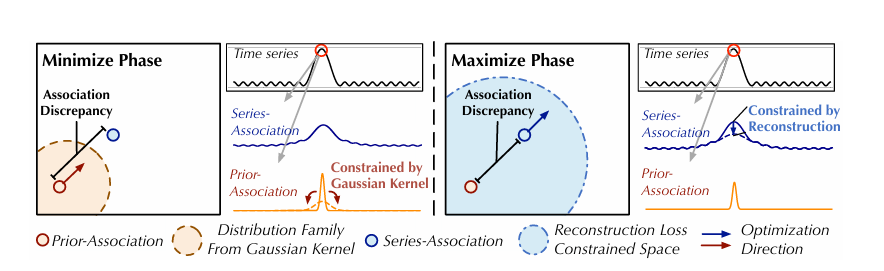
The minimax strategy is employed to enhance the difference between normal and abnormal points by operating on the association discrepancy between prior-association and series-association. The primary goal is to optimize the anomaly detection process by adjusting the association discrepancy to make anomalies more detectable.

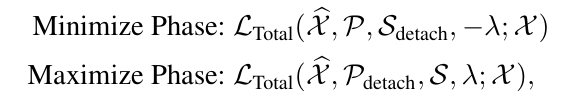
**Phases of Minimax Strategy**

The minimax strategy operates in two phases:

**Minimization Phase:** During this phase, the prior-association is forced to approximate the series-association. This ensures that the prior-association—which captures adjacent time points and local information—adapts to the varying temporal patterns in the data. The model minimizes the difference between the prior-association and series-association for normal points, helping the model learn a reliable representation of normal behavior in the data.

**Maximization Phase:** In this phase, the series-association is optimized to maximize the association discrepancy. This forces the series-association to emphasize non-adjacent time points, effectively making it harder for anomalies to reconstruct these associations. Since anomalies tend to have weaker connections with distant points in the series, this maximization process increases the detectability of abnormal points by exaggerating their differences from normal patterns.



The minimax strategy is reflected in the objective function, where the total loss combines a reconstruction loss and a discrepancy loss. The reconstruction loss guides the series-association to reconstruct the input data, while the discrepancy loss maximizes the gap between prior- and series-associations for abnormal time points. This dual-objective approach enhances the model's ability to separate normal from abnormal points. The loss function can be expressed as:

Sdetach: During the minimization phase, the gradient of the series-association (S) is detached to prevent it from updating. This allows only the prior-association (P) to be optimized and learn from the series-association.

Pdetach: During the maximization phase, the gradient of the prior-association (P) is detached so it remains fixed. This allows only the series-association (S) to be optimized, ensuring that S focuses on increasing the association discrepancy.

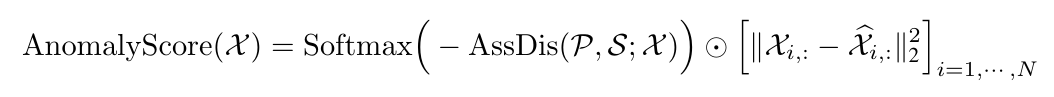
**Association-based Anomaly Criterion**

The association-based anomaly criterion is a key element in the Anomaly Transformer, designed to leverage both the association discrepancy and reconstruction error for anomaly detection. The idea is to use these two measures together to more effectively differentiate between normal and abnormal time points.

Association Discrepancy: The association discrepancy is computed as the difference between the prior-association (local, adjacent points) and the series-association (global, full series). Normal points are expected to have larger association discrepancies because they exhibit strong connections with distant time points, while anomalies typically have lower discrepancies due to weaker global associations.

Reconstruction Error: Alongside the association discrepancy, the model also computes the reconstruction error, which is the difference between the original input time series and its reconstructed version. Anomalies are harder to reconstruct accurately, so they tend to have higher reconstruction errors compared to normal points.

**Anomaly Score**

The final anomaly score is computed by combining both the association discrepancy and reconstruction error. The criterion integrates these two signals to improve the model's ability to identify anomalies more effectively. The formula for the anomaly score is:

**Interpretation of Anomaly Score**

Normal Points: For normal time points, the association discrepancy tends to be larger, but the reconstruction error is lower. The softmax term effectively reduces the impact of the association discrepancy on the anomaly score for these points.

Anomalous Points: For anomalies, both the association discrepancy is small (due to weak global connections), and the reconstruction error is higher. This leads to a larger anomaly score, making it easier to detect these points.

**Collaborative Detection**

This approach allows the model to exploit two distinct signals—temporal associations and reconstruction accuracy—making it more robust in detecting a wide range of anomalies in time series data. The integration of both the association-based and reconstruction-based criteria ensures that the model is sensitive to different kinds of abnormal behaviors, from subtle deviations in associations to outright failures in reconstruction.

**3.3 Technology stack**

Frontend: Streamlit

Backend: Express.js

Programming Language: JavaScript, Python

Frameworks/Libraries: PyTorch (for model implementation), Matplotlib (for visualizations) ,scikit learn (for data preprocessing and evaluation metrics) , numpy (for numerical computations), seaborn (for creating statistical visualizations), pandas (for data manipulation and analysis)

**3.4 Implementation details**

Datasets:

The datasets used in this project are **Synthetic weather testing dataset designed to simulate real-world environmental conditions** and **Bitcoin** dataset which consists of open, close, volume pf the bitcoin. It contains attributes such as date, temperature, humidity, wind speed, rainfall, solar radiation, atmospheric pressure, and visibility. Additionally, the dataset includes a binary "Is Anomaly" column, indicating whether each observation is anomalous. The data is complete, with no missing values, and provides a comprehensive view of weather patterns, making it suitable for evaluating anomaly detection models. By incorporating both normal and anomalous data points, this dataset enables a thorough assessment of the model's performance in identifying deviations within time series data.

Data Preprocessing:

Irrelevant Column Removal: The dataset was cleaned by removing columns that were not relevant to the anomaly detection task, ensuring that only meaningful features were used for training and evaluation.

The data is processed to convert it into a suitable format for model training and inference. This involves padding the feature data to a consistent length and converting it into embeddings with 512 dimensions. Specifically, we pad the input data to ensure each feature vector has 512 dimensions, which serves as the input to the transformer model for anomaly detection.

Handling Missing Values: Rows with missing values were dropped to maintain the integrity of the data without introducing biases.

Training Process:

Loss Function: The model uses a reconstruction-based loss function that measures the difference between the input time-series data and its reconstruction by the model. This helps the model learn to capture the underlying patterns in the data, making it proficient at identifying anomalies. Additionally, a regularization term based on association discrepancy is incorporated into the loss function. This term penalizes the model for failing to properly identify discrepancies between normal and anomalous time points, further enhancing its ability to detect anomalies effectively. Together, these components ensure that the model not only learns to reconstruct the time-series data but also focuses on the crucial features that distinguish anomalous behavior.

Optimizer: The Adam optimizer was chosen for training due to its adaptive learning rate capabilities and efficiency in handling sparse gradients. Adam combines the benefits of both Momentum and RMSprop, making it suitable for complex neural network architectures like the Anomaly Transformer. The optimizer adjusts the learning rate during training, ensuring faster convergence while preventing issues like vanishing or exploding gradients, which can occur in deep learning models. This helped stabilize the training process and enabled the model to reach optimal performance without requiring extensive manual tuning of the learning rate.

Hyperparameters:

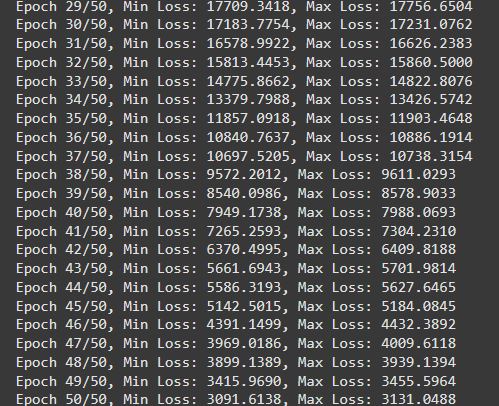
Learning Rate: A learning rate of 0.01 was used. This value strikes a balance between allowing the model to learn quickly and preventing it from overshooting optimal weights. A learning rate that is too high could cause instability, while one that is too low could slow down convergence.

Epochs: The model was trained for 50 epochs. This choice of epochs ensures that the model has enough time to learn from the data, iterating through the training set multiple times to fine-tune its weights. The number of epochs was chosen to ensure convergence without overfitting. Early stopping was implemented (if applicable) to halt training once the model stopped improving on the validation set.

Model parameters:

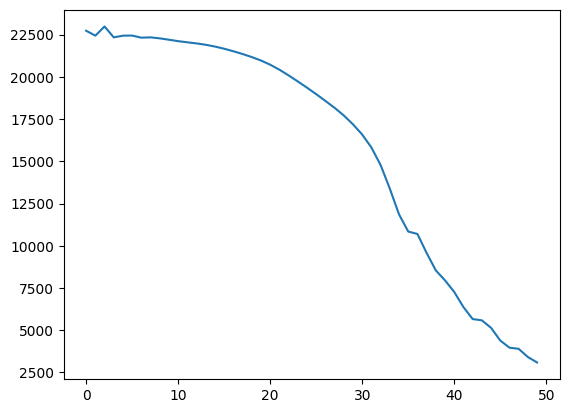
**4. Results & Discussions**

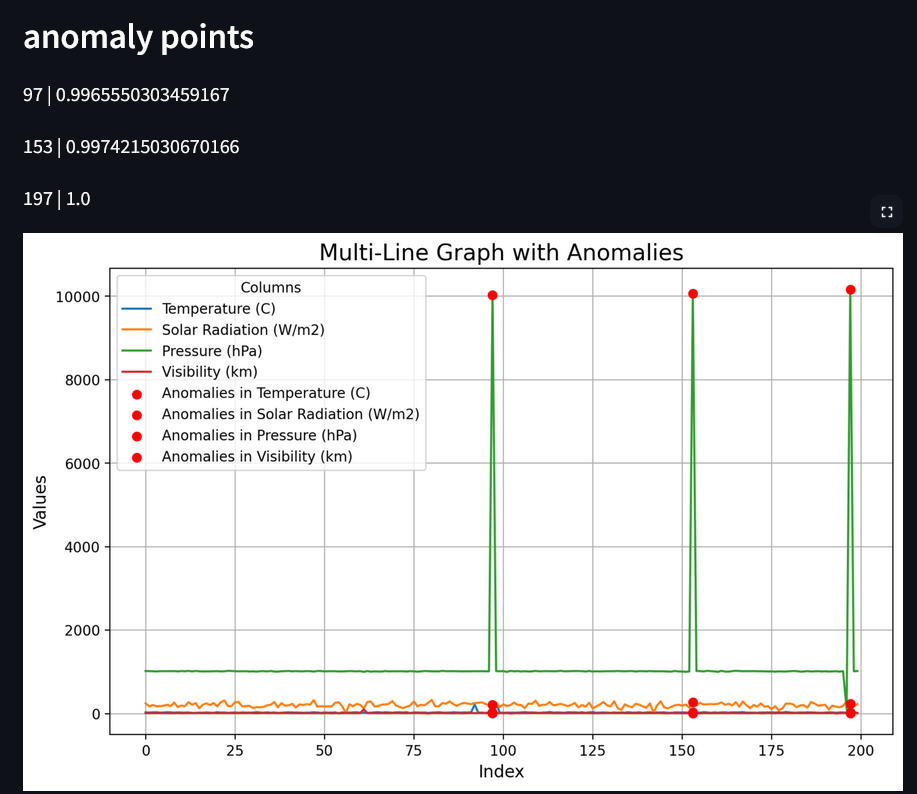
Training is done with the synthetic weather dataset, which was preprocessed and used for model training. The model’s performance is evaluated by tracking the training loss over epochs. The following images present the loss and the model's progression over time:

****

The graph below shows how the training loss evolves over the course of 50 epochs. Initially, the loss is relatively high, indicating that the model is still learning the patterns in the data. As the training progresses, the loss gradually decreases in a smooth curve, reflecting the model’s improvement in detecting anomalies. This continuous decrease suggests that the model is effectively optimizing its parameters and refining its predictions with each epoch.

The curve indicates consistent progress, as the model adapts and becomes better at anomaly detection over time. The training loss reduction signifies that the model is successfully learning from the synthetic weather dataset and improving its ability to identify anomalies as it moves through the epochs.



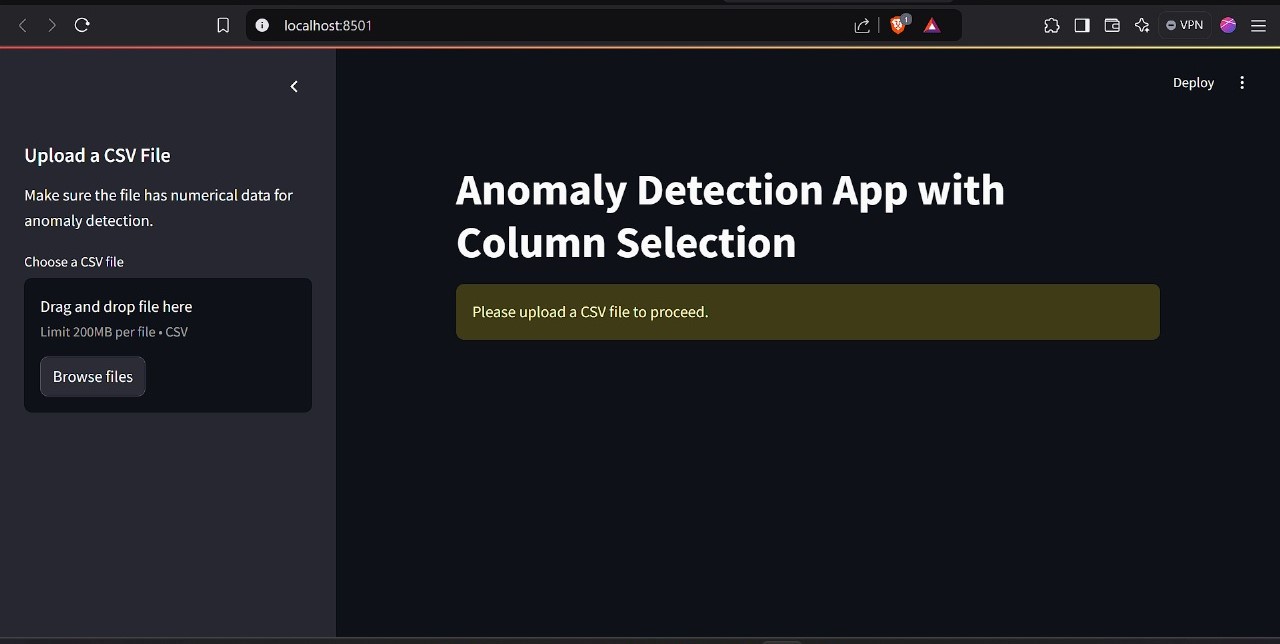


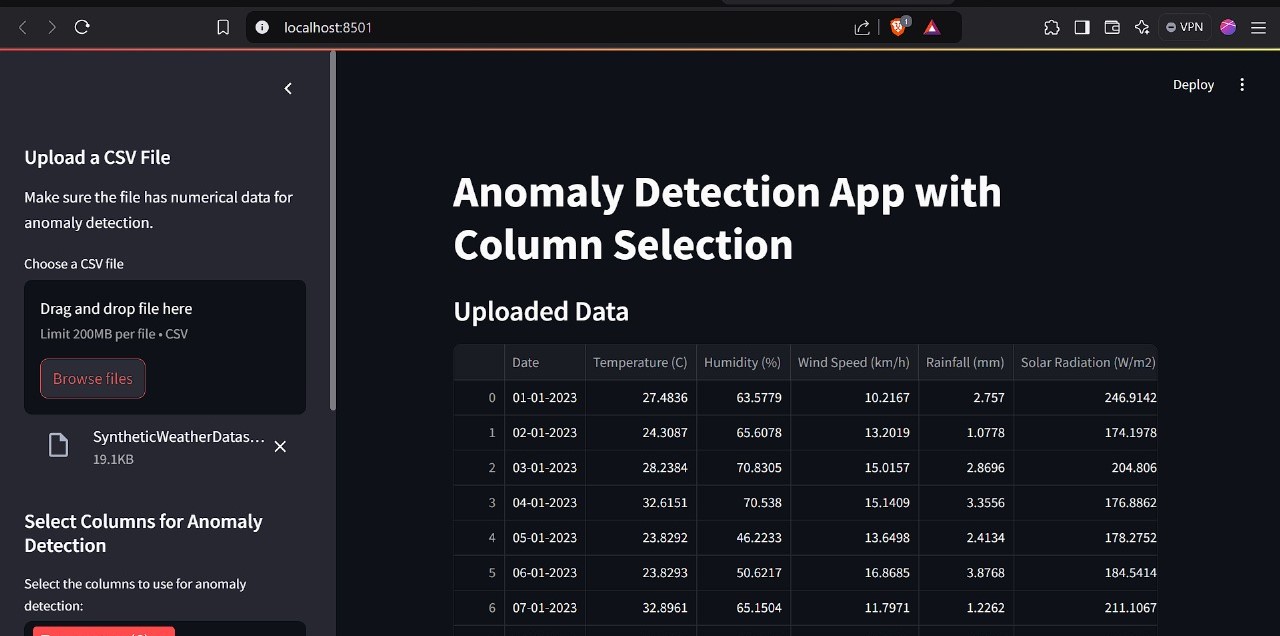
The figure presents a multi-line graph depicting data for Temperature (°C), Solar Radiation (W/m²), Pressure (hPa), and Visibility (km), with each variable represented by different colored lines. Anomalies are marked by red dots at specific points, indicating outliers in the data. Anomalies were detected at indices 97, 153, and 197, with corresponding scores of 0.99655, 0.99742, and 1.0. The y-axis shows the variable values, while the x-axis displays the index from 0 to 200. The Pressure variable, in particular, exhibits significant spikes at these points, indicating unusual behavior, while the other variables remain relatively stable. These detected anomalies are crucial for identifying abnormal patterns within the dataset.

The following images showcase the front-end user interface of the anomaly detection system. The interface is designed to provide a clean and intuitive experience for users. The key elements of the front-end include:

1. **Anomaly Detection Dashboard**: This image displays the main dashboard where users can upload datasets, view anomaly detection results, and interact with the system.
2. **Upload and Visualize Buttons**: The interface includes buttons such as "Upload File" for uploading datasets, "Get Results" for processing the data, and "Visualize Results" for displaying the detected anomalies in graphical form.

These images provide a visual overview of the system’s user interface, demonstrating its simplicity and functionality.





**5.Conclusion**

This paper explores the challenge of unsupervised anomaly detection in time-series data by introducing a methodology centered on learning time-point associations using Transformers. Departing from traditional approaches, we emphasize the concept of association discrepancy to better differentiate between normal and anomalous data points. To this end, we propose the Anomaly Transformer, which features a two-branch Anomaly Attention mechanism designed to effectively model and amplify these discrepancies. Additionally, a minimax strategy is employed to further enhance the distinction between normal and abnormal points. By integrating association discrepancy into a unified criterion, the model achieves seamless collaboration between reconstruction accuracy and association analysis.

**6.References**

Anomaly Transformer : Time series anomaly detection with association discrepancy

Jiehui Xu∗, Haixu Wu∗, Jianmin Wang, Mingsheng Long

Vaswani, Ashish, et al. "Attention Is All You Need." Proceedings of the Advances in Neural

Information Processing Systems (NeurIPS), 2017

Optimal Ratio for Data Splitting V. Roshan Joseph H. Milton Stewart School of Industrial and Systems Engineering Georgia Institute of Technology, Atlanta, GA 30332, USA