Cobblestone Energy
Assignment-Graduate Software Engineer

Efficient Data Stream Anomaly Detection

By

Amardeep (20MI31002)

Indian Institute of Technology Kharagpur

October 2024

Table of Contents

Cobblestone Energy 1

Assignment-Graduate Software Engineer 1

Executive Summary 3

1. Introduction 3

2. Simulating Streaming Data 3

2.1 Data Simulation Approach 3

3. Methods for Anomaly Detection 4

3.1 Z-Score Analysis 4

3.2 Seasonal Smoothing with Holt-Winters 5

Methodology: 5

3.3 Isolation Forest for Outlier Detection 6

3.4 Ensemble Method: Combining Multiple Techniques 7

4. Overview of Anomaly Detection and Annotations 8

5. Performance Evaluation 8

6. Discussion and Recommendations 9

7. Conclusion 9

8. Future Work 10

References : 11

Executive Summary

Detecting anomalies within real-time data streams is essential for maintaining operational integrity, early issue detection, and fraud prevention across sectors such as finance, cybersecurity, and industrial automation. This project investigates a set of methods to identify anomalies within continuous data streams, employing statistical, machine learning, and ensemble techniques. The performance of each method is evaluated using a simulated data stream that represents real-world conditions by incorporating trends, periodic variations, and randomness. The outcomes reveal the strengths and limitations of each approach, helping to establish the most effective strategies for anomaly detection in dynamic data environments.

1. Introduction

In many domains, the ability to promptly identify anomalies in data streams is vital for ensuring timely responses to potential risks and irregularities. Anomalies, or data points that significantly diverge from established patterns, can reveal issues such as equipment malfunctions, security breaches, or unexpected business trends. Addressing these irregularities effectively requires algorithms that can adapt to shifting patterns in streaming data.

2. Simulating Streaming Data

To evaluate various anomaly detection methods, we created a synthetic data stream that mimics the characteristics of real-world time-series data. The generated stream includes gradual trends, cyclic behaviors, and random fluctuations, along with sporadic anomalies inserted to challenge the detection methods.

2.1 Data Simulation Approach

The simulated stream is constructed using:

- Gradual Trend: Represents long-term growth or decline in the data.
- Cyclic Patterns: Modeled through a sine wave, reflecting periodic behaviors such as daily cycles or seasonal changes.
- Random Noise: Introduces minor, unpredictable variations to simulate natural fluctuations.
- Anomalies: Sudden spikes or dips are injected to test the algorithms' sensitivity and detection capabilities.

The simple sine wave pattern serves as a useful baseline, making it easier to evaluate algorithm effectiveness.

3. Methods for Anomaly Detection

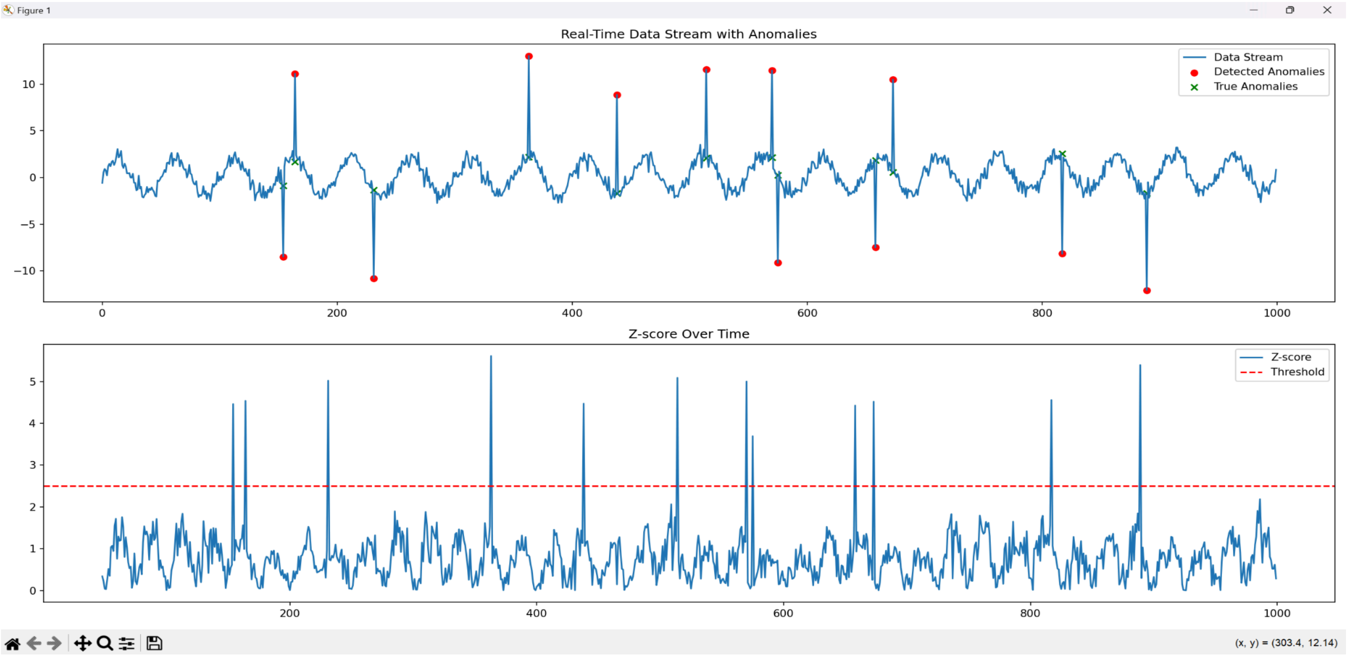
We employed four different techniques for detecting anomalies in real-time data streams, ranging from basic statistical methods to advanced machine learning models. These techniques are described below, highlighting their underlying mechanisms and practical applications.

3.1 Z-Score Analysis

Z-Score is a statistical method that evaluates the distance of each data point from the mean, expressed in terms of standard deviations. It is particularly useful for identifying outliers in datasets where deviations from the norm are easily quantifiable.

Methodology:

- **Rolling Mean and Standard Deviation**: Calculated over a sliding window to ensure continuous adaptation to changes in data distribution.
- **Threshold-Based Detection**: Points are flagged as anomalies if their Z-score surpasses a defined threshold, typically set at ±2.5 or higher.
- **Adjustment for Zero Variance**: A small constant is used to replace zero variance cases, ensuring stability in calculations.



Z-Score Method plot

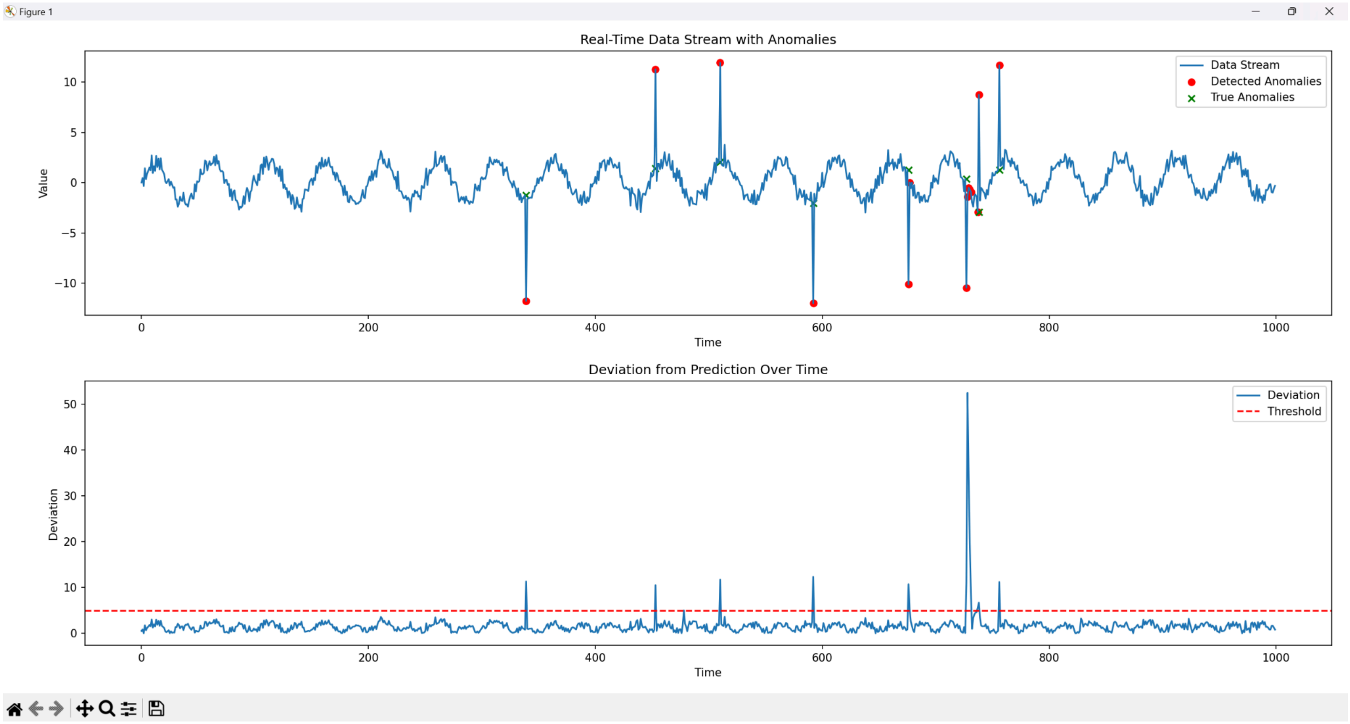
Given the repetitive sine wave pattern in the dataset, Z-Score analysis proved highly effective in distinguishing anomalies, as deviations from the average were straightforward to detect.

3.2 Seasonal Smoothing with Holt-Winters

The Holt-Winters technique employs exponential smoothing to forecast future data values by considering level, trend, and seasonal components. This method is well-suited for time-series data with regular, repeating patterns.

Methodology:

- **Three-Part Smoothing**: Adjusts for the current data point, underlying trend, and seasonal pattern simultaneously.
- **Real-Time Updating**: The model continuously revises its predictions as new data arrives.
- **Deviation-Based Anomaly Detection**: Anomalies are detected by comparing observed values with the model's predictions, with larger deviations indicating potential anomalies.



Holt - Winter plot

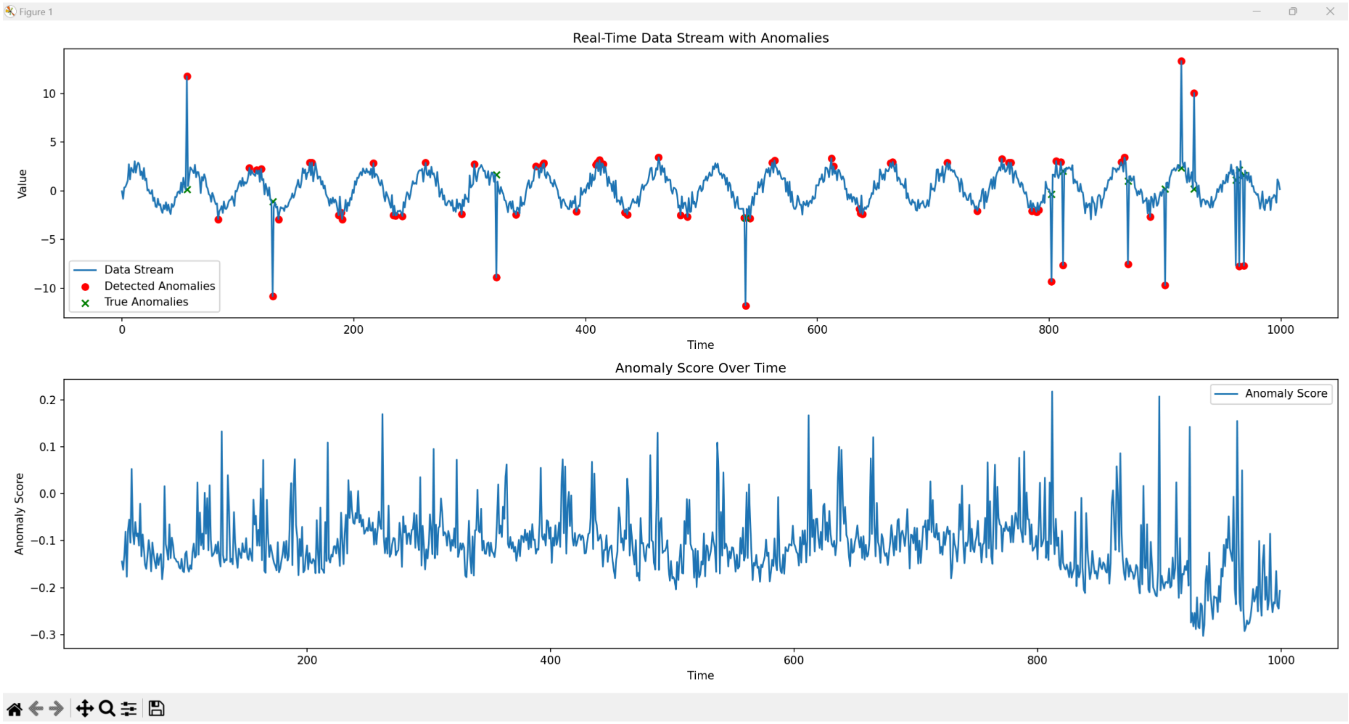
Holt-Winters requires parameter tuning to accommodate different seasonal patterns, which may affect its performance in dynamic environments.

3.3 Isolation Forest for Outlier Detection

Isolation Forest, a tree-based machine learning algorithm, detects anomalies by isolating data points through random partitions. It is especially useful for datasets with complex, high-dimensional structures.

Methodology:

- **Incremental Learning**: Periodically retrains the model on the latest data to accommodate changes in the underlying distribution.
- **Anomaly Scoring**: Data points receive anomaly scores based on how easily they are isolated, with higher scores indicating potential outliers.
- **Concept Drift Handling**: The model adapts to evolving patterns by updating itself with the most recent data.



Isolation Forest Plot

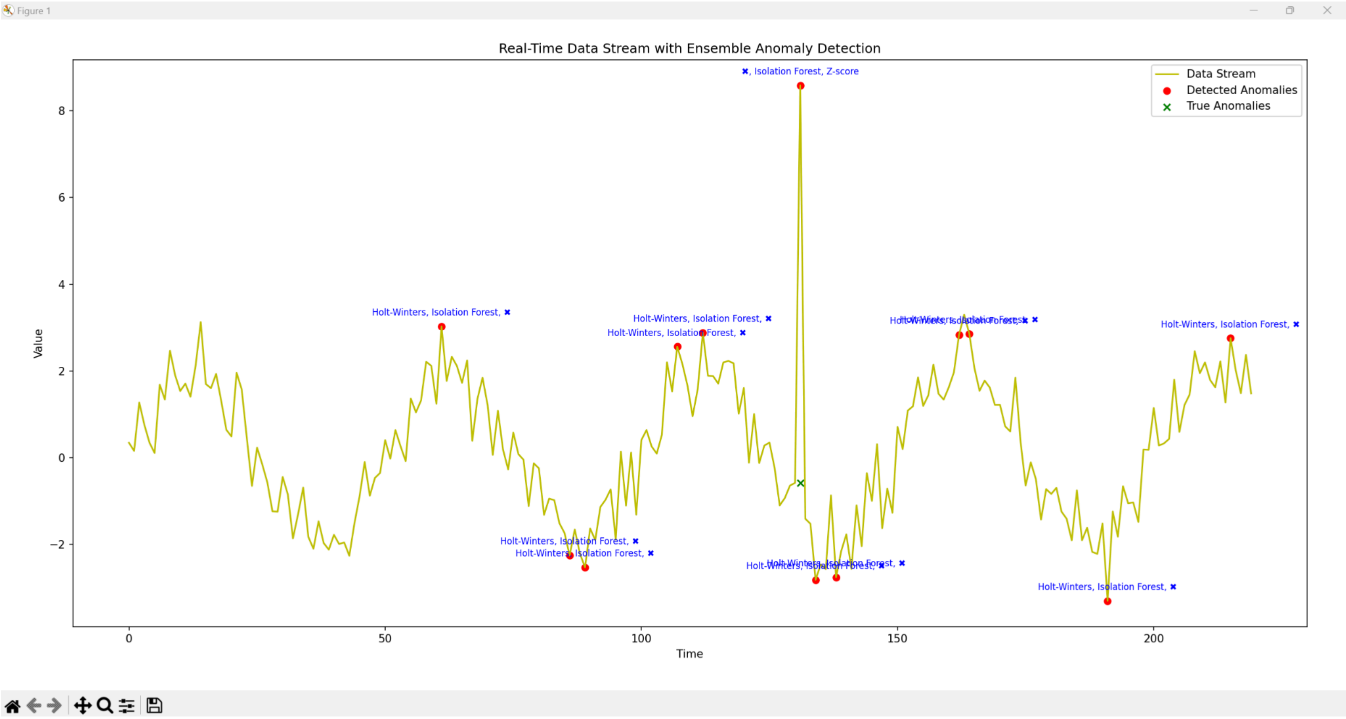
While Isolation Forest is robust in various contexts, it requires frequent updates to maintain accuracy in streaming data applications.

3.4 Ensemble Method: Combining Multiple Techniques

The ensemble approach integrates predictions from Z-Score, Holt-Winters, and Isolation Forest, using a majority voting scheme to determine the final anomaly decision. This strategy combines the strengths of each method to achieve a more balanced and reliable detection outcome.

Methodology:

- **Voting Mechanism**: If at least two out of three algorithms identify a data point as an anomaly, it is classified as such.
- **Integration of Diverse Perspectives**: Combines insights from statistical analysis, time-series forecasting, and machine learning.
- **Continuous Evaluation**: The system remains adaptable by evaluating new data as it arrives, ensuring real-time detection capability.



Ensemble Learning Plot

The ensemble approach demonstrated superior accuracy in detecting anomalies compared to individual methods, owing to its ability to leverage different detection principles.

4. Overview of Anomaly Detection and Annotations

- Tracking Anomalies:

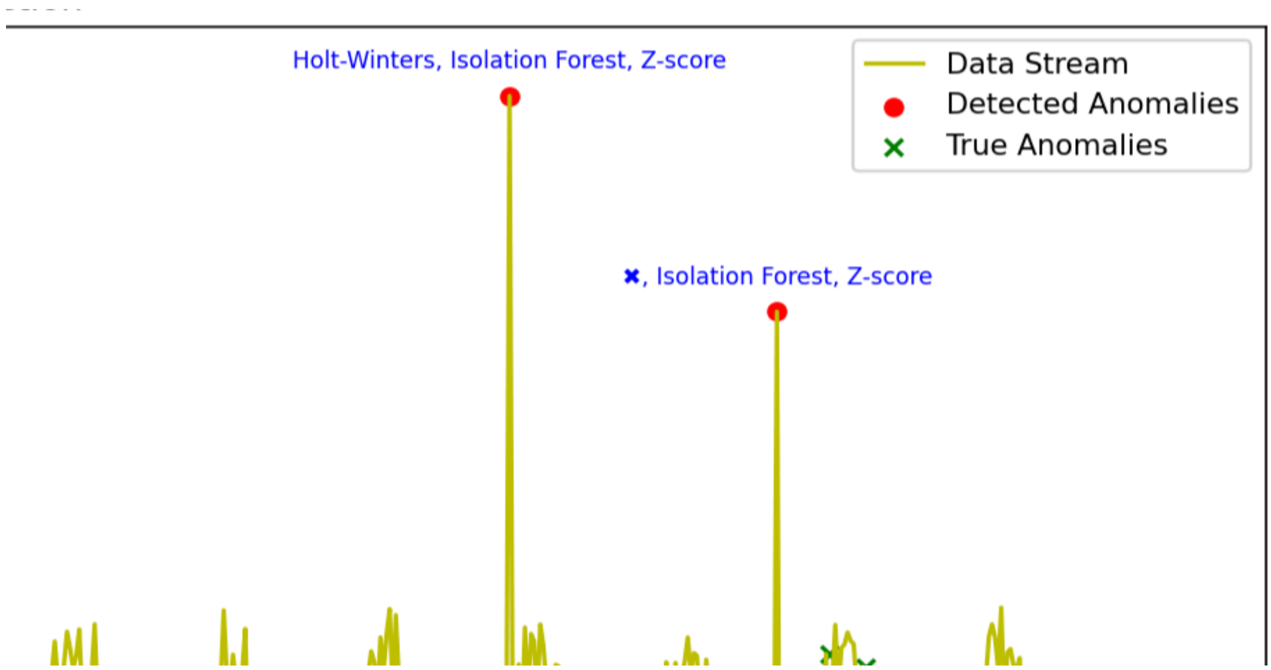
The code checks if a point is identified as an anomaly by the ensemble.If true, it records the index and notes which algorithms (Holt-Winters, Isolation Forest, Z-score) flagged it.

- Recording Details:

Algorithms that detected the anomaly are listed, or marked with '✖' if they didn't. The relevant details for each anomaly are captured.

- Plotting Annotations:

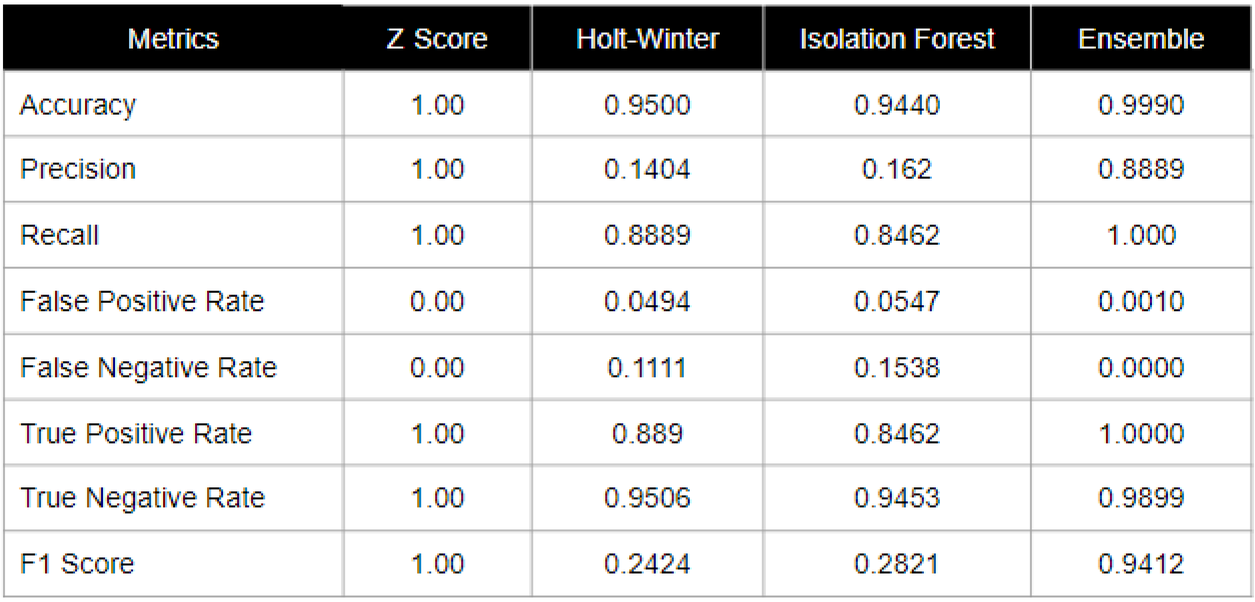
Annotations are added near each detected anomaly on the plot. The algorithms involved are shown slightly above the anomaly point in blue text.



Right most anomaly point - Holt winter predicted false , Isolation Forest predicted true, Z-score predicted true so as per voting ensemble predicted true.

5. Performance Evaluation

Each method was assessed using metrics such as precision, recall, accuracy, and F1-score. The evaluation revealed that the methods performed differently under varying data conditions.



Metrics	Z Score	Holt-Winter	Isolation Forest	Ensemble
Accuracy	1.00	0.9500	0.9440	0.9990
Precision	1.00	0.1404	0.162	0.8889
Recall	1.00	0.8889	0.8462	1.000
False Positive Rate	0.00	0.0494	0.0547	0.0010
False Negative Rate	0.00	0.1111	0.1538	0.0000
True Positive Rate	1.00	0.889	0.8462	1.0000
True Negative Rate	1.00	0.9506	0.9453	0.9899
F1 Score	1.00	0.2424	0.2821	0.9412

Evaluation Metrics

6. Discussion and Recommendations

The results from this study suggest that:

- **Z-Score Analysis** is a viable option for datasets with simple patterns but struggles with complex seasonal behaviors.
- **Holt-Winters** can accurately detect anomalies in data with repeating cycles but may need manual parameter adjustments.
- **Isolation Forest** offers flexibility in handling diverse datasets but requires regular updates.
- **Ensemble Methods** provide a comprehensive solution by combining the benefits of various algorithms, making them suitable for general use cases.

7. Conclusion

This project explored several approaches to detecting anomalies in real-time streaming data. The study highlighted the strengths and weaknesses of each technique, with the ensemble approach emerging as the most effective solution. This research lays a foundation for more sophisticated anomaly detection systems that could incorporate advanced techniques such as deep learning or adaptive models.

8. Future Work

Future research could involve:

- **Incorporating Neural Networks**: Such as LSTM models for handling complex time-series data.
- **Adaptive Parameter Optimization**: Using machine learning to dynamically adjust parameters.
- **Exploring Additional Ensemble Techniques**: Including more diverse models to improve robustness.

.

References :

[1] https://builtin.com/machine-learning/anomaly-detection-algorithms

[2] https://www.datacamp.com/tutorial/introduction-to-anomaly-detection

[3]https://www.ibm.com/topics/anomaly-detection#:~:text=In%20business%20data%2C%20three%20main,contextual%20anomalies%20and%20collective%20anomalies.

[4]https://corporatefinanceinstitute.com/resources/data-science/ensemble-methods/#:~:text=Ensemble%20methods%20are%20techniques%20that,accuracy%20of%20the%20results%20significantly.

[5]https://medium.com/@abhishekjainindore24/different-types-of-ensemble-techniques-bagging-boosting-stacking-voting-blending-b04355a03c93