UNIVERSITY OF CAMBRIDGE
INSTITUTE OF CONTINUING EDUCATION

Mini-project 5.3: Detecting the anomalous activity of a ship's engine

Applying critical thinking and ML concepts to design and implement a robust anomaly detection model

Arunima Bhattacharya

Data Science With Machine Learning & AI Career Accelerator

June 2025

This page intentionally left blank.

Contents

List of Figures iii

List of Tables v

1 Introduction 1

1.1 Brief Overview 1

1.2 Problem Statement 1

2 Methods 2

2.1 General Overview 2

2.1.1 Interquartile Range (IQR) Method 2

2.1.2 One-Class Support Vector Machine (OCSVM) 2

2.1.3 Isolation Forest 2

3 Results 4

3.1 Exploratory Data Analysis and Statistical Outlier Detection 4

3.1.1 Visualisation 5

3.2 One-Class Support Vector Machine (OCSVM) 6

3.2.1 Visualisation 6

3.3 Isolation Forest 7

3.3.1 Visualisation 7

4 Conclusion 9

A Appendix 10

This page intentionally left blank.

List of Figures

1.1 Hierarchical classification of data 1

2.1 Data preprocessing and outlier detection workflow 3

3.1 Distribution histograms of engine feature data before and after scaling . . . 5
3.2 Outlier Counts and IQR Lines from normalised feature set 6
3.3 OCSVM Anomaly Rates for Selected Configurations within the 1-5% anomaly range 7
3.4 Anomaly detection comparison: OCSVM forms clear decision boundaries, while Isolation Forest isolates anomalies dispersed along edges. 8

A.1 Violin Plots: Feature Distributions by OCSVM Label for anomaly rate =2.01% 10
A.2 Isolation Forest: Distribution of anomalies across each feature for anomaly rate =2.01% 11

This page intentionally left blank.

List of Tables

3.1 Outlier severity and skewness per feature 4
3.2 Outlier Counts and Proportions per Feature 5
3.3 95th Percentile Thresholds and Outlier Ranges 5

4.1 Comparison of anomaly detection methods 9

This page intentionally left blank.

1
Introduction

1.1 Brief Overview

Data is the empirical foundation for insights, decision-making, and advancement across scientific and industrial domains in data science. To better understand the structure and categorisation of data, Figure 1.1 illustrates a hierarchical breakdown.

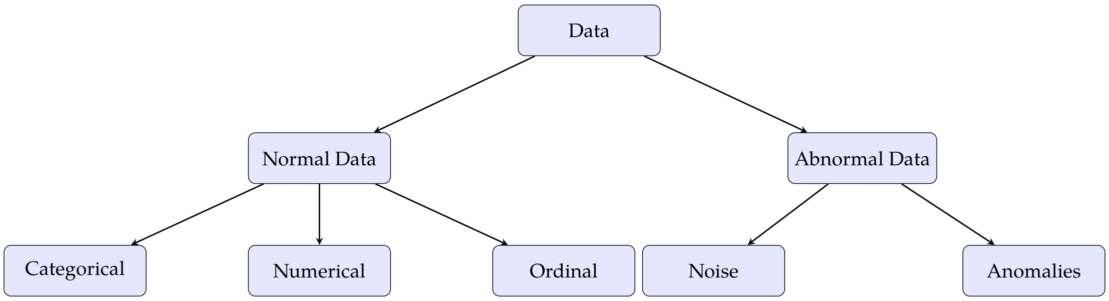


Figure 1.1: *Hierarchical classification of data*

This report focuses on detecting **anomalies**, which, unlike noise, carry meaning, such as indicators of faults, failures, or rare events. Identifying these deviations is essential to maintain the integrity and efficiency of the systems. Detecting anomalies requires a combination of techniques:

1. **Statistical methods**, such as the Interquartile Range
2. **Machine learning** (**ML**) **methods** such as
    - *One-Class Support Vector Machine*
    - *Isolation Forest*

1.2 Problem Statement

Poorly maintained ship engines increase the risk of inefficiency, fuel waste, malfunctions, and safety hazards. This project uses a real-world dataset (Devabrat, 2022) to design a robust anomaly detection model by exploring six key features using the techniques discussed in Chapter 2.

2
Methods for Anomaly Detection

2.1 General Overview

This report flags 1-5% anomalies in the ship engine's dataset using the following three methods:

2.1.1 Interquartile Range (IQR) Method

IQR flags anomalies beyond 1.5×IQR from above the third quartile (Q3) or below the first quartile (Q1).

2.1.2 One-Class Support Vector Machine (OCSVM)

OCSVM, an unsupervised learning algorithm, models normal data boundaries, classifying data points as normal or anomalous based on their position relative to this boundary. Principal Component Analysis (PCA) is used to project its output to 2D for visualisation.

2.1.3 Isolation Forest

Isolation Forest, an ensemble-based anomaly detection algorithm, isolates anomalies using random trees, making it fast, scalable, and well-suited for high-dimensional data.

See figure 2.1 below for an outline of all preprocessing steps and outlier detection.

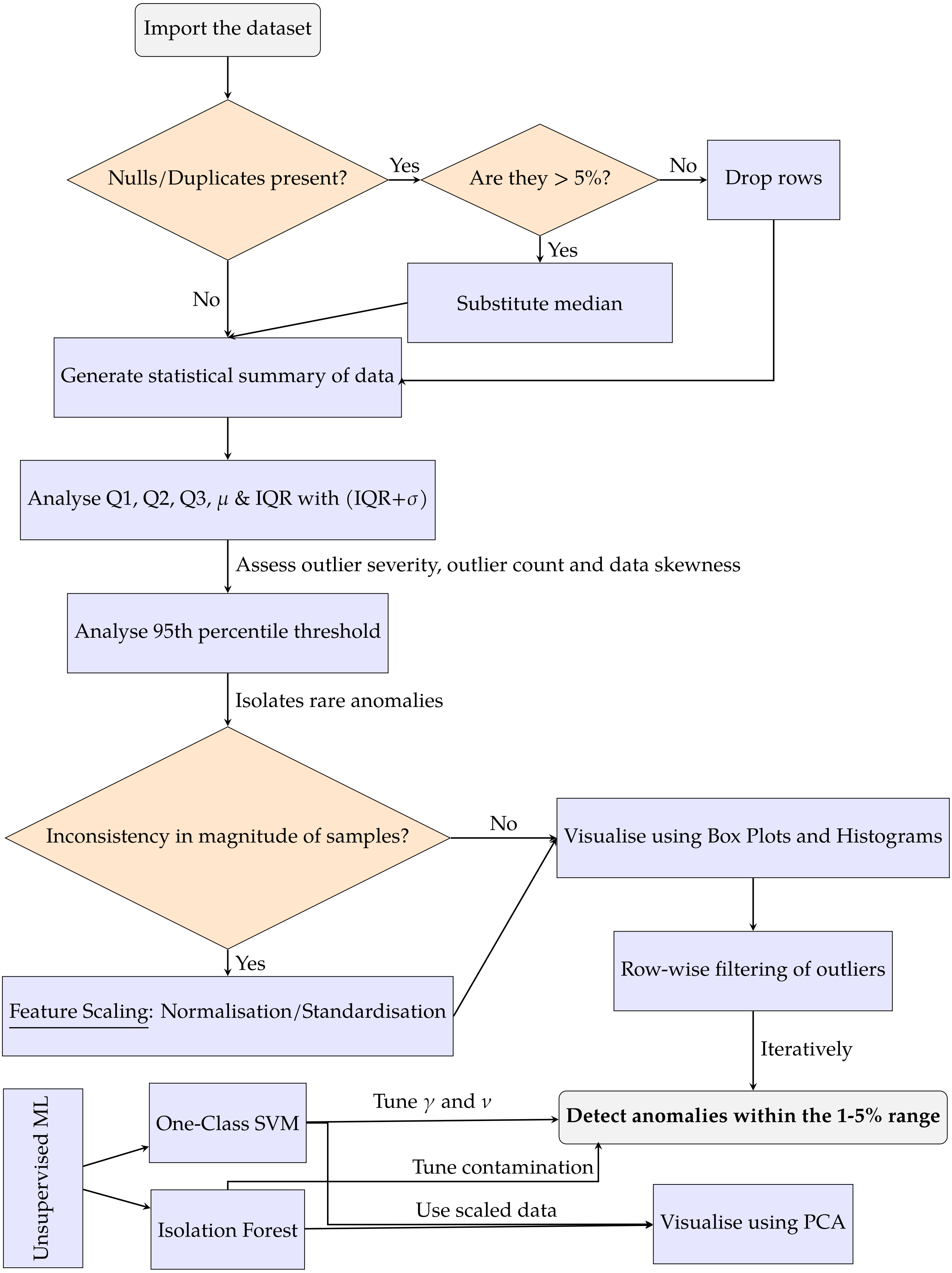

Import the dataset
Nulls/Duplicates present?
Yes
Are they > 5%?
No
Drop rows
Yes
No
Substitute median
Generate statistical summary of data
Analyse Q1, Q2, Q3, μ & IQR with (IQR+σ)
Assess outlier severity, outlier count and data skewness
Analyse 95th percentile threshold
Isolates rare anomalies
Inconsistency in magnitude of samples?
No
Visualise using Box Plots and Histograms
Row-wise filtering of outliers
Yes
Feature Scaling: Normalisation/Standardisation
Iteratively
Unsupervised ML
One-Class SVM
Tune γ and ν
Detect anomalies within the 1-5% range
Tune contamination
Isolation Forest
Use scaled data
Visualise using PCA

Figure 2.1: *Data preprocessing and outlier detection workflow*

3
Results

3.1 Exploratory Data Analysis and Statistical Outlier Detection

The dataset contains no missing, null, or duplicate values, and all entries are numeric. Descriptive statistics and a hybrid IQR–σ method reveal the presence of outliers and skewed distributions across multiple features as summarised in table 3.1.

Feature	Mean (μ)	Median	Upper Bound	Outlier Severity	Skewness
Engine rpm	791.24	746.00	$\mu + 5\sigma$	Very strong	Moderate Positive
Fuel pressure	6.66	6.20	$\mu + 5\sigma$	Very strong	Strong Positive
Coolant temperature	78.43	78.35	$\mu + 19\sigma$	Extreme	Strong Positive
Coolant pressure	2.34	2.17	$\mu + 5\sigma$	Strong	Moderate Positive
Lubricant oil pressure	3.30	3.16	$\mu + 4\sigma$	Strong	None/ Symmetric
Lubricant oil temperature	77.64	76.82	$\mu + 4\sigma$	Mild	Moderate Positive

Table 3.1: *Outlier severity and skewness per feature*

Standard IQR-based outlier detection

This method quantifies feature-wise outliers using binary flags, as shown in table 3.2. Lubricant oil temperature has the highest outlier rate, followed by Fuel pressure and Coolant pressure.

95th Percentile Analysis

The 95th percentile threshold identifies the top 5% samples for each feature. Most features have the maximum values close to the 95th percentile value, except for Engine rpm and Coolant temperature, which deviate substantially, as shown in table 3.3.

Feature	Outlier Count	Percentage of Total Samples
Engine rpm	464	2.38%
Lub oil pressure	66	0.34%
Fuel pressure	1135	5.81%
Coolant pressure	785	4.02%
Lub oil temperature	2617	13.40%
Coolant temperature	2	0.01%

Table 3.2: *Outlier Counts and Proportions per Feature*

Feature	95th Percentile	Min Above 95th	Max Above 95th
Engine rpm	1324.00	1325.00	2239.00
Lubricant oil pressure	5.06	5.06	7.27
Fuel pressure	12.21	12.21	21.14
Coolant pressure	4.44	4.44	7.48
Lubricant oil temperature	84.94	84.94	89.58
Coolant temperature	88.61	88.61	195.53

Table 3.3: *95th Percentile Thresholds and Outlier Ranges*

3.1.1 Visualisation

Figure 3.1 presents histograms of all features using raw, normalised, and standardised datasets.

- Raw data show severe scale imbalances masking distributions of smaller-scale features.
- Standardisation centres the data but does not resolve skewness.
- Normalisation enables clearer visual comparison.

These visuals support earlier observations on skewness and validate using normalised data for effective visualisation.

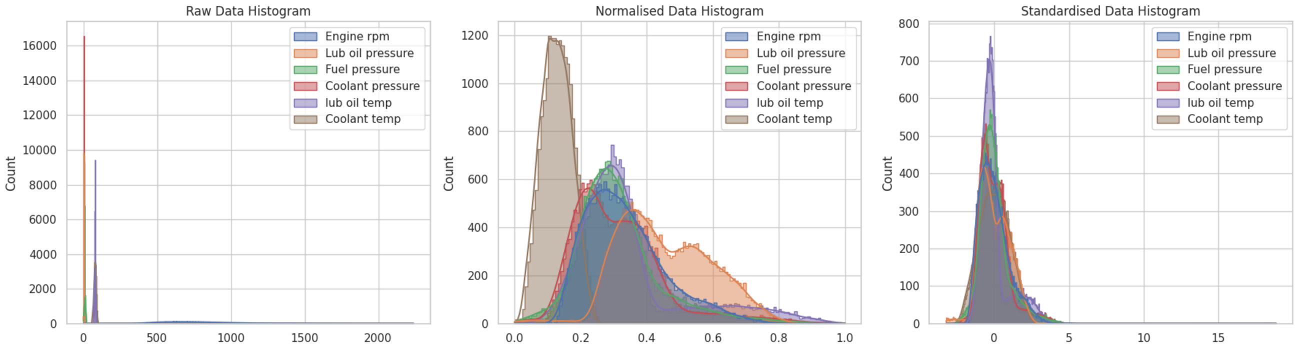


Figure 3.1: *Distribution histograms of engine feature data before and after scaling*

A clearer picture of outliers per feature is illustrated in the figure 3.2.

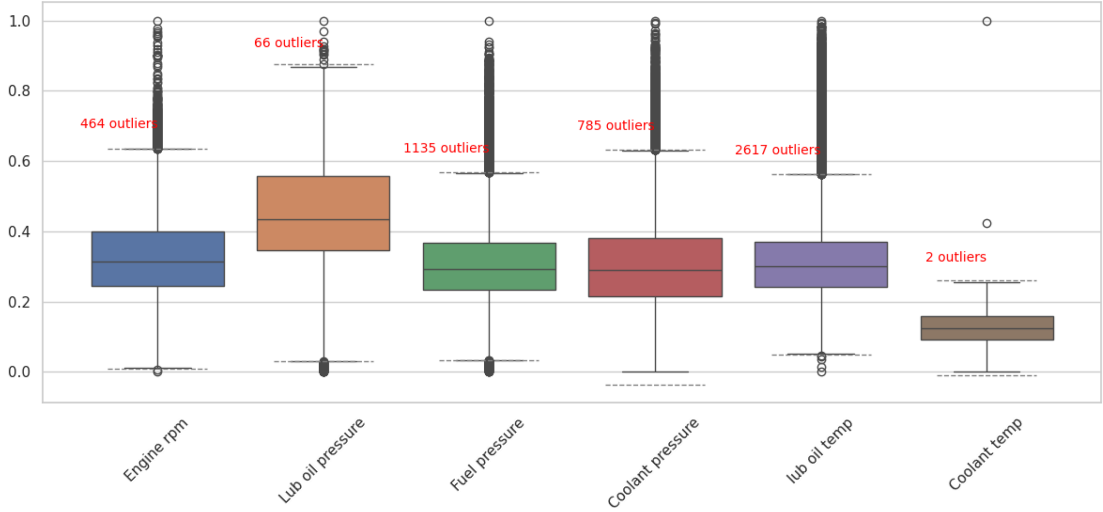


Figure 3.2: *Outlier Counts and IQR Lines from normalised feature set*

Result from simultaneous outlier detection: Applying row-wise IQR filtering, only rows with ≥ 2 feature-level outliers meet the 1–5% anomaly rate, resulting in 422 flagged samples (2.16%).

3.2 One-Class Support Vector Machine (OCSVM)

OCSVM is applied to the normalised and standardised datasets.

- Figure 3.3 illustrates how different $(\gamma - \nu)$ combinations affect the anomaly rates for these scaled datasets.
    - Normalised data more frequently achieve the 1–5% rate than standardised data.
- ν has a more immediate and interpretable effect on the analysis than γ.
- Temperature-related characteristics contribute most significantly to the first principal component (PC1), suggesting that they are among the most informative features for identifying anomalies.
    - This aligns with the statistical results, where *coolant temperature* exhibited the most extreme outliers, and *lubricant oil temperature* showed the highest outlier frequency.

3.2.1 Visualisation

Normalised dataset better achieved the target rate of 1–5% during training, while the standardised data produced cleaner, more interpretable decision boundaries. The configuration $\gamma = 0.1$, $\nu = 0.02$ produced the best result, yielding a cohesive decision boundary without internal fragmentation and detecting 627 outliers (3.21%). This outcome is illustrated in the left panel of Figure 3.4.

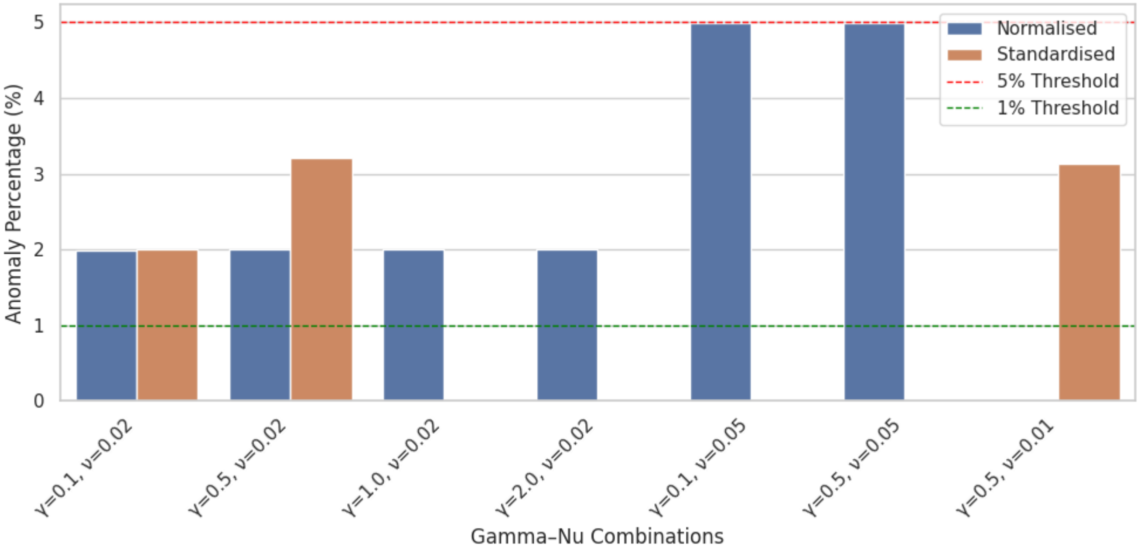


Figure 3.3: *OCSVM Anomaly Rates for Selected Configurations within the 1-5% anomaly range*

3.3 Isolation Forest

Following OCSVM, Isolation Forest offers a robust alternative that is insensitive to feature scaling. Its core parameter, *contamination*, directly controls the anomaly rate (e.g., 0.01 targets 1% anomalies). This makes it ideal for the project as setting contamination to 0.01 and 0.05 yields 196 and 977 anomalies, respectively.

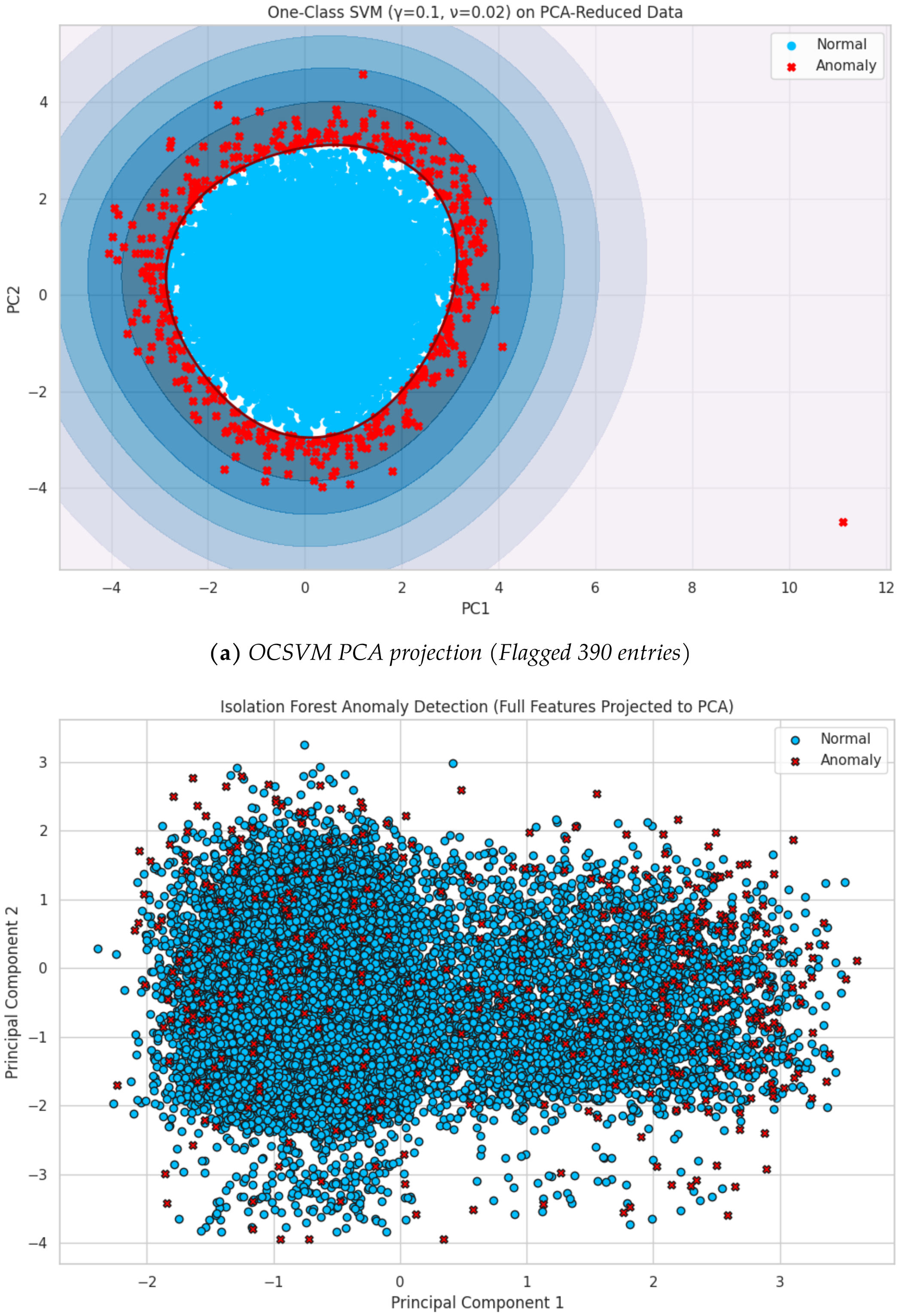
3.3.1 Visualisation

The bottom panel of figure 3.4 illustrates the anomaly separation pattern of Isolation Forest in PCA-reduced space for a 2% anomaly rate.

- A clustered structure with no correlation between PC1 and PC2 is observed.
- Anomalies (red) are dispersed, clearly visible, and scattered on the periphery and in less dense areas.

Conclusion: While Isolation Forest more easily flags a defined range of outliers, OCSVM is visually more interpretable due to its smooth boundary and clearer anomaly separation[1].

[1] See Appendix A for a clearer understanding of anomalies at the 2.01% rate across features.



(**a**) *OCSVM PCA projection (Flagged 390 entries)*

(**b**) *Isolation Forest PCA projection (Flagged 391 entries)*

Figure 3.4: *Anomaly detection comparison: OCSVM forms clear decision boundaries, while Isolation Forest isolates anomalies dispersed along edges.*

4
Conclusion

Temperature features contribute significantly to the analysis.

This project develops a robust anomaly detection workflow using statistical and unsupervised ML methods, with key findings summarised in Table 4.1.

Method	**Dataset**	**Key Parameters**	**Number of Anomalies**	**Percent Outliers**	**Notes**
IQR	Normalised	IQR across features	**422** with ≥ 2 per row	**2.16%**	Column-wise outlier flags; binary indicators per feature.
OCSVM	Normalised (training), Standardised (PCA and Visualisation)	Kernel coefficient (γ), Anomaly fraction parameter (ν)	389–975	1.99–4.99%	Full-feature model; PCA for visualisation. **At 2% anomaly rate,** $\sim$ 390 **entries are flagged.**
Isolation Forest	Standardised	contamination	196–977	1–5%	Scalable, distribution-agnostic; PCA for interpretation. **At 2% anomaly rate, 391 entries are flagged.**

Table 4.1: *Comparison of anomaly detection methods*

- IQR provided a simple statistical baseline, revealing skewed distributions and structured outliers, but lacked flexibility.
- OCSVM allowed decision boundary control and benefited from standardised data for clearer projections, though it remained challenging to tune.
- Isolation Forest scaled efficiently, detected edge-case anomalies without feature scaling, and offered direct control over the anomaly rate.

A
Appendix

Feature-wise anomalous behaviour from unsupervised ML models

1. Low lubricant oil temperature and low fuel pressure consistently indicate anomalies across both ML models, highlighting them as key parameters.
2. Isolation Forest detects rare but critical outliers in coolant temperature clearly, unlike in the OCSVM violin plot.
3. For the remaining features, Isolation Forest distributions highlight anomaly ranges more clearly than OCSVM violin plots.

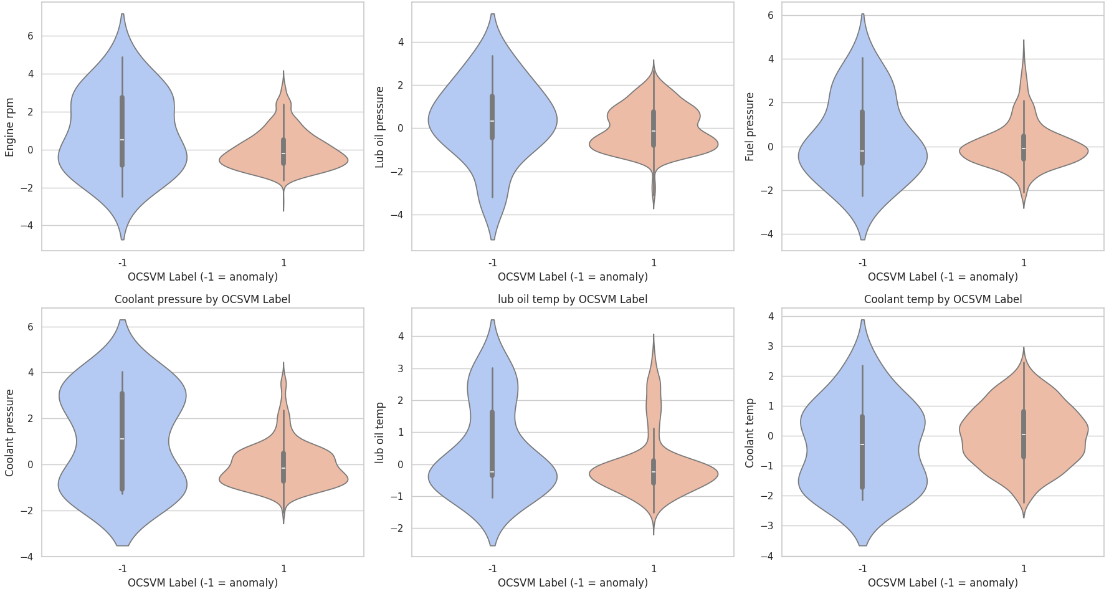


Figure A.1: *Violin Plots: Feature Distributions by OCSVM Label for anomaly rate =2.01%*

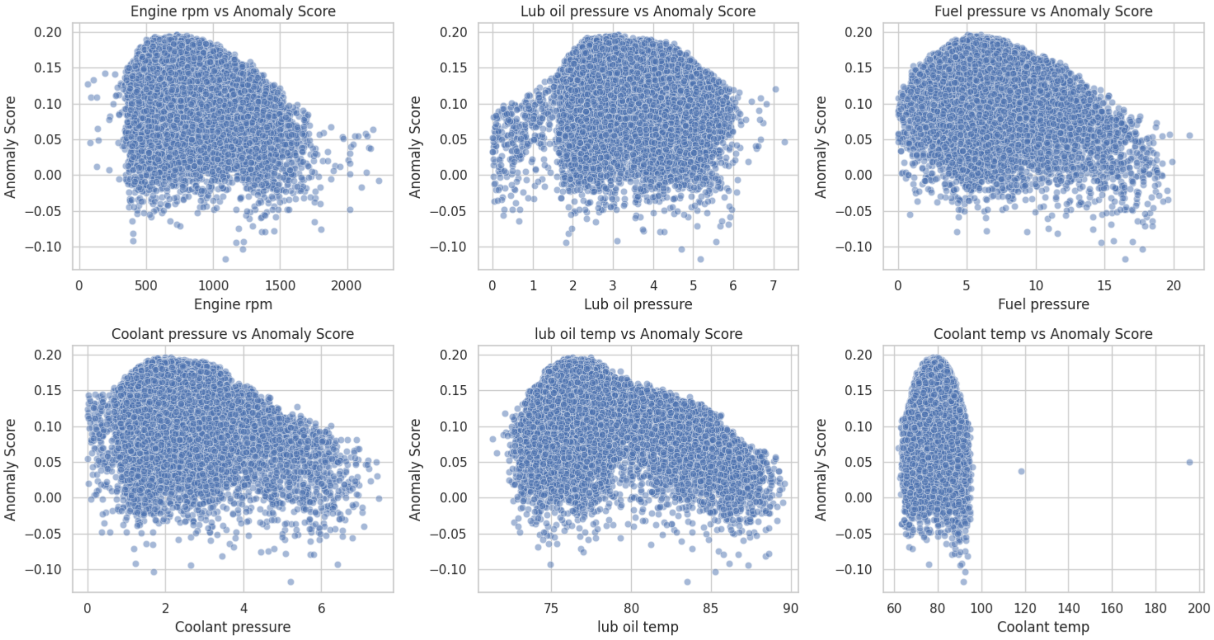


Figure A.2: *Isolation Forest: Distribution of anomalies across each feature for anomaly rate =2.01%*

This page intentionally left blank.