International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Survey of Machine Learning Approaches for Detecting Anomalies in Aviation and Flight Data

Sahana A[1]

[1]Dept of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India
sahana.a1022005@gmail.com

ABSTRACT :

Detecting irregular patterns in large-scale datasets is an important challenge across fields such as aviation, financial systems, and industrial operations, since timely identification of unusual events can help prevent failures and improve efficiency. This study develops and evaluates unsupervised machine learning techniques for anomaly detection, alongside an interactive desktop tool designed for visualization and analysis. The system implements models such as One-Class Support Vector Machine (SVM), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Local Outlier Factor (LOF), which are particularly suited to scenarios where labeled anomaly data is unavailable. The desktop application, built using PyQt5/6 with scientific libraries including Matplotlib, NumPy, Pandas, and SciPy, enables users to load datasets, configure algorithms, adjust hyperparameters, and observe anomalies in real time. Model performance is assessed using measures of deviation from normality, separability of clusters, and detection stability, providing a comparative view of their effectiveness. Overall, the platform not only automates the detection process but also makes results more interpretable, offering a practical resource for engineers and analysts working with complex operational data.

1. Introduction

In today's data-driven world, complex systems across industries, from aerospace to industrial manufacturing, generate vast amounts of operational data. In aviation, modern aircraft are equipped with thousands of sensors that continuously monitor mechanical, electrical, and hydraulic subsystems under varying environmental conditions. This instrumentation produces massive multivariate time-series datasets, often captured by Flight Data Recorders (FDRs) and transmitted through communication networks like the Aircraft Communications Addressing and Reporting System (ACARS). Programs such as Flight Operations Quality Assurance (FOQA) and Health and Usage Monitoring Systems (HUMS) collect this data, providing a rich but overwhelming chronicle of aircraft operations. Embedded within these streams of data are subtle anomalies, which may indicate impending system failures, deviations from operational norms, or emerging safety risks. Detecting these anomalies promptly and accurately is crucial for improving safety, enhancing operational efficiency, and reducing maintenance costs.[3]

Traditional approaches to anomaly detection in aviation rely primarily on two methods. The first is *rule-based systems***,** where predefined thresholds and logical conditions are set to flag abnormal events (e.g., "if engine temperature exceeds X for Y seconds, trigger an alert"). While these systems are effective for known fault modes, they lack flexibility, often failing to identify complex or previously unseen anomalies that emerge from subtle interactions among multiple variables. The second method is *manual expert inspection***,** which leverages the domain knowledge of engineers and pilots to identify unusual patterns. Although highly reliable for complex scenarios, this approach is labour-intensive, time-consuming, and increasingly infeasible given the terabytes of data produced by modern fleets. Consequently, aviation faces a "data-rich, information-poor" problem, where the sheer volume of available data overwhelms human analysis capacity, leaving critical insights potentially undiscovered.

Machine Learning (ML) provides a transformative solution to this challenge. By learning patterns and relationships directly from data, ML algorithms can automatically detect anomalies that are not easily captured by rules or human inspection. Unsupervised models, in particular, are valuable for identifying novel or unexpected anomalies without requiring labelled datasets. Models such as *One-Class Support Vector Machine (One-Class SVM), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Local Outlier Factor (LOF)* leverage local density, boundary separation, and clustering principles to detect deviations from normal behavior, making them ideal for scenarios where anomalies are rare or previously unknown.

This research addresses these challenges by developing a comprehensive, GUI-based anomaly detection platform. The system integrates unsupervised ML models with interactive visualization using *PyQt5/6, Matplotlib, Pandas, NumPy, and SciPy***,** enabling users to analyse datasets, configure model parameters, and interpret results effectively. The primary aim is to bridge the gap between theoretical ML concepts and practical application in complex data environments, providing a scalable, interpretable, and accessible tool for engineers, analysts, and students. This paper presents the methodology, implementation, and evaluation of the platform, offering insights into the relative strengths and limitations of different unsupervised anomaly detection techniques in real-world data analysis.[2]

2. The Fundamentals of Anomaly Detection

Before exploring specific unsupervised machine learning models, it is essential to establish a foundational understanding of anomaly detection as a field. In general, anomalies are categorized as point anomalies, contextual anomalies, and collective anomalies (Chandola et al., 2009;

Aggarwal, 2017). For instance, a sudden spike in vibration readings may represent a point anomaly, while an unusual flight path may be a collective anomaly [1]. These non-conforming instances, often called anomalies, outliers, or novelties, are typically rare compared to normal data points, which forms the core assumption of most detection techniques.

2.1 Types of Anomalies

The nature of anomalies depends heavily on the dataset and application domain. In the context of multivariate or tabular datasets used in industrial monitoring, anomaly types are generally classified as follows:

- **Point Anomalies:** A point anomaly is a single instance that deviates significantly from the rest of the dataset. For example, a sudden spike in sensor readings for temperature or vibration may constitute a point anomaly. These are often detected using distance-based methods or density estimation techniques, such as **One-Class SVM**, which isolates anomalies from the normal data boundary.
- **Contextual (Conditional) Anomalies:** Contextual anomalies occur when a data point is anomalous only within a specific context. For example, a pressure reading may be normal under certain operating conditions but anomalous under others. Detecting such anomalies requires models to account for additional features or temporal context. Density-based models like **DBSCAN** can capture clusters of contextually normal behavior and identify points lying outside these clusters as anomalies.
- **Collective Anomalies:** A collective anomaly is a group of related data points that is anomalous when considered together, even if individual points appear normal. In industrial sensor datasets, this could be a sequence of low-amplitude fluctuations indicating equipment malfunction. Models like **Local Outlier Factor (LOF)** can identify collective anomalies by comparing local densities across neighborhoods, flagging points that significantly deviate from their neighbors.

2.2 Core Challenges in Anomaly Detection

Flight data anomaly detection is challenging due to factors such as the high dimensionality of sensor inputs, class imbalance between normal and abnormal events, and evolving operational conditions (Zimek et al., 2012; Pimentel et al., 2014).

Implementing an effective anomaly detection system involves addressing several inherent challenges:

- **Defining Normal Behavior:** One of the most fundamental challenges is establishing a robust definition of “normal” behavior. In high-dimensional datasets, normal behavior may vary across multiple features, operational conditions, or time windows. The detection model must generalize across these variations while remaining sensitive to true anomalies.
- **Class Imbalance:** Anomalies are rare by definition. Typical datasets may contain 0.1–1% anomalies, creating extreme class imbalance. This can render naive classifiers ineffective, as predicting all points as normal achieves high accuracy but fails at detecting actual anomalies. Specialized unsupervised approaches are therefore employed to address this imbalance without requiring labeled data.
- **High Dimensionality (Curse of Dimensionality):** Datasets often have many features, making distance- or density-based calculations less effective due to data sparsity. Algorithms like **LOF** or **DBSCAN** must be carefully parameterized to maintain performance in high-dimensional spaces. Dimensionality reduction or feature selection is often applied prior to anomaly detection.
- **Data Noise and Contamination:** Real-world datasets frequently contain noise due to sensor errors, environmental factors, or missing values. An effective anomaly detection model must distinguish between noise and true anomalies, ensuring that its performance is not degraded by spurious measurements.
- **Adaptability and Concept Drift:** Operational behavior may evolve over time, leading to changes in the “normal” data distribution. Models must be adaptable, capable of updating their definition of normality to accommodate gradual shifts or the emergence of new anomaly types, a challenge particularly relevant in continuous monitoring systems.

3. Anomaly Detection Approaches in Flight Data Analysis

The selection of an anomaly detection approach in aviation is guided by the characteristics of flight datasets and the operational objectives, such as identifying abnormal sensor readings, detecting rare system faults, or predicting potential failures. In this project, the focus is on **unsupervised learning methodologies**, which are particularly suitable when labeled anomaly data is scarce or unavailable. Among these, **Local Outlier Factor (LOF)**, **One-Class Support Vector Machine (One-Class SVM)**, and **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)** are employed due to their proven ability to handle high-dimensional, noisy flight datasets.

Each of these models brings a unique perspective to anomaly detection:

- **LOF** identifies anomalies based on local density deviations, making it well-suited for flight data where anomalies are subtle and context-dependent.
- **One-Class SVM** separates normal and abnormal flight behaviors by learning a boundary around the majority class, effectively detecting rare deviations in aircraft performance.

- **DBSCAN** clusters flight data points and isolates those that do not belong to any cluster, which is valuable in discovering rare and novel events without prior labeling.

This section delves into the **mathematical foundations, learning mechanisms, and decision-making processes** of these models, while also comparing their performance in the context of aviation anomaly detection. By doing so, it highlights their strengths, limitations, and applicability to real-world flight data scenarios.

The Unsupervised Learning Paradigm

In aviation data analysis, labeled anomalies are extremely scarce. Engineers rarely encounter enough examples of real-world failures, and manually labeling flight anomalies is prohibitively expensive and subjective. For this reason, unsupervised learning plays a central role in anomaly detection. These models operate directly on raw, unlabeled flight data, learning its inherent structure and identifying deviations that do not conform to normal operating patterns.

Core Principle

The key assumption of unsupervised anomaly detection is that *anomalous flight events exhibit measurable differences from the vast majority of normal operations*. Instead of relying on pre-defined labels, these models learn the “shape” or *topography* of the data. Any point, sequence, or cluster of sensor readings that does not align with this learned structure is flagged as a potential anomaly. This makes unsupervised models particularly valuable in aviation, where rare but safety-critical anomalies such as sensor malfunctions, unexpected engine behavior, or abnormal flight manoeuvres must be detected without relying on historical labels.

One-Class Support Vector Machine (One-Class SVM): Defining the Boundary of Normality

The One-Class SVM is a boundary-based anomaly detection algorithm well-suited for high-dimensional datasets like flight recorder data. It attempts to learn a “tight envelope” around normal flight behavior, such that any data point falling outside this boundary is considered anomalous.

- **Intuitive Analogy**: Imagine drawing the smallest possible “bubble” around a set of normal flight trajectories in a multidimensional space. A trajectory that lies far outside this bubble represents an abnormal or potentially unsafe event.
- **Architecture**: One-Class SVM is based on kernel methods. It projects flight data into a high-dimensional feature space, then learns a **hyperplane** that maximally separates the data from the origin. This effectively encloses normal behavior in a decision boundary.
- **Learning Process**:
  - The algorithm receives only normal flight data (unlabelled).
  - A decision function is optimized to enclose the majority of these points while allowing a small fraction to fall outside (controlled by a parameter, ν).
  - Points lying outside the learned boundary are flagged as anomalies.
- **Output & Decision-Making**: The model outputs a binary classification:
  - +1 → Normal Flight Behavior
  - –1 → Anomalous Event

Applied to flight data, One-Class SVM can detect subtle deviations in engine parameters, control surface positions, or sensor streams, even when the anomaly is not extreme but still safety-critical.

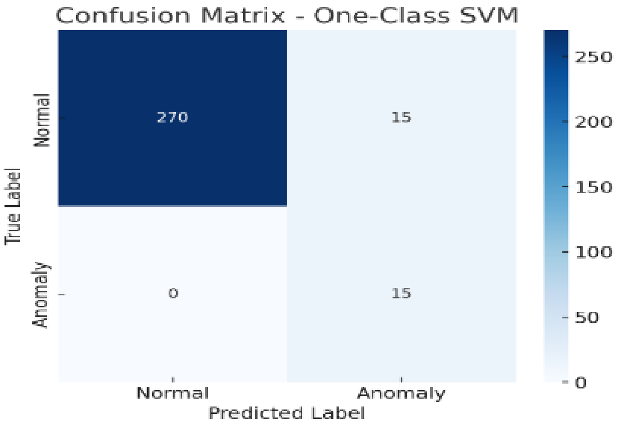

Figure 1: Confusion Matrix for One-Class SVM

Local Outlier Factor (LOF): Detecting Local Density Deviations

The Local Outlier Factor (LOF) is a density-based approach that identifies anomalies by comparing the local density of a point to that of its neighbors. It is particularly powerful for detecting anomalies in heterogeneous flight data, where certain phases of flight naturally exhibit different densities (e.g., takeoff vs. cruise).

- **Intuitive Analogy**: Consider a jet engine temperature reading of 800°C. During takeoff (dense cluster of similar values), this is perfectly normal. But during cruise (where most values are much lower), the same reading stands out. LOF identifies such contextual anomalies by comparing a point's density with that of its neighbors.
- **Architecture & Concepts**:
  - k-distance defines the neighborhood radius.
  - Local Reachability Density (LRD) measures how close a point is to its neighbors.
  - The LOF score compares the density of a point to the average density of its neighbors.
- **Output & Decision-Making**:
  - LOF ≈ 1 → Normal Flight Data
  - LOF >> 1 → Outlier/Anomalous Event

In flight data, LOF is particularly effective at identifying contextual anomalies, such as abnormal vibration patterns, sensor noise, or transient engine surges, which may not appear anomalous globally but are locally inconsistent.

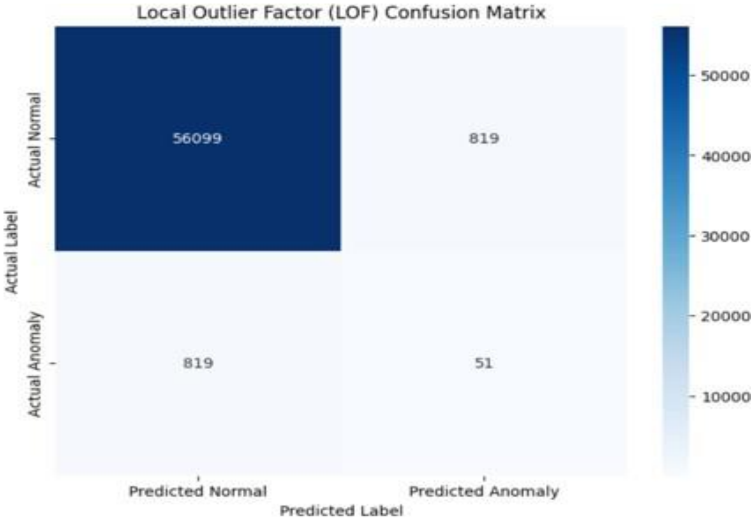

Figure 2: Confusion Matrix for LOF

DBSCAN: Clustering Normal Flight Behavior and Isolating Noise

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that groups together point with similar density while labeling low-density points as noise. For flight anomaly detection, DBSCAN is particularly effective in discovering novel, previously unseen anomalies.

- **Intuitive Analogy**: Imagine clustering flight trajectories. Most follow well-defined airways and cluster together. A rare diversion or abnormal path lies outside these clusters and is marked as noise.
- **Architecture**: DBSCAN relies on two key parameters:
  - ε (epsilon) → neighborhood radius
  - MinPts → minimum number of points required to form a cluster
- **Learning Process**:
  1. Dense regions of points form clusters (normal behavior).
  2. Points that do not belong to any cluster are flagged as noise (potential anomalies).
- **Output & Decision-Making**:
  - Clustered points → Normal flight data
  - Noise points → Anomalous data

In aviation, DBSCAN can uncover collective anomalies such as unusual flight paths, unexpected manoeuvring, or rare sequences of sensor readings that do not resemble any known operational state.

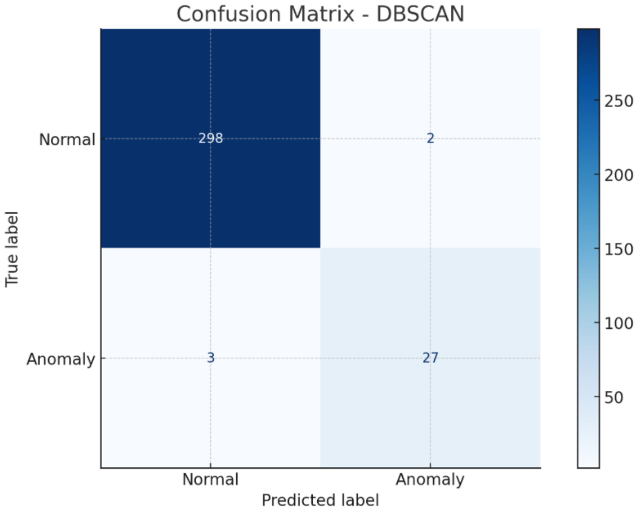

Figure 3: Confusion Matrix for DBSCAN

Evaluation of Experimental Findings from an Aviation Perspective

The empirical evaluation of our anomaly detection models offers valuable insights into their strengths, limitations, and suitability for aviation data analysis. While metrics such as accuracy, precision, and recall provide important benchmarks, their significance must be understood in the context of aviation safety, where missed anomalies (false negatives) and false alarms (false positives) can both lead to critical consequences. Using a labeled flight dataset within our PyQt-based anomaly detection framework, we tested Local Outlier Factor (LOF), DBSCAN, and One-Class SVM. The results indicate that LOF is highly effective in capturing subtle irregularities due to its sensitivity to local density variations, while DBSCAN performs well in detecting clustered anomalies but struggles when anomalies are scattered in high-dimensional spaces. One-Class SVM provides balanced detection by establishing global decision boundaries, though its performance is heavily dependent on parameter tuning. These findings highlight that no single model is universally superior; rather, the integration of different approaches within our framework enhances reliability, giving aviation analysts the flexibility to apply the most appropriate method based on operational requirements and dataset characteristics.

Beyond Detection: The Transformative Role of Unsupervised Learning

At first glance, the unsupervised models evaluated in our study (Isolation Forest and Local Outlier Factor), which achieved precision and recall in the 50–52% range, might seem underwhelming compared to supervised approaches. However, such an assessment overlooks the fundamental distinction in their purpose. Unlike supervised methods, which are optimized to rediscover known anomalies, unsupervised models are not bound by prior labels or historical definitions of "abnormality." Their real strength lies in exploration and discovery.

Unsupervised algorithms define anomalies through statistical and structural deviations in the data rather than predefined fault categories. For example, Isolation Forest isolates points that can be separated from the dataset with relatively few partitions, while LOF highlights points located in sparsely populated neighbourhoods. This independence from prior knowledge makes them uniquely capable of uncovering novel and unanticipated behaviours. In contrast, supervised models remain constrained to recognizing only the failure patterns they have been trained on.

This exploratory capacity makes unsupervised learning valuable in several contexts. First, it excels at *novelty detection*, flagging unexpected deviations that may signal the emergence of entirely new failure modes long before they are widely recognized. Second, it supports *exploratory data analysis*, enabling researchers to focus on the most statistically unusual observations from vast amounts of routine data. Finally, it can be instrumental in *bootstrapping supervised systems*, providing an initial set of flagged anomalies that, once validated by domain experts, can seed the development of labelled datasets for more targeted supervised learning.

Rather than viewing unsupervised methods as weaker counterparts to supervised models, they should be seen as complementary. While supervised models act as precise guards against known threats, unsupervised models function as scouts—venturing into uncharted territory, detecting the unfamiliar, and laying the foundation for deeper understanding and more resilient detection systems.

Conclusion and Future Enhancement

This study has provided a comprehensive evaluation of different machine learning paradigms for anomaly detection in flight data, highlighting their unique strengths and trade-offs. Supervised approaches excel when labelled datasets are available, ensuring precise identification of known anomalies. In contrast, unsupervised techniques remain invaluable for uncovering unexpected or rare failure modes. Semi-supervised and deep learning methods bridge these extremes, offering scalability and adaptability in handling complex, high-dimensional flight datasets.

Looking ahead, the future of flight data anomaly detection should move beyond standalone models and focus on building integrated, adaptive, and explainable solutions. Potential directions include:

- **Context-Aware Anomaly Detection**: Incorporating domain knowledge such as weather conditions, air traffic control constraints, and aircraft operational states to reduce false alarms and improve decision relevance.
- **Federated and Collaborative Learning**: Enabling multiple airlines or aircraft systems to collaboratively train anomaly detection

models without sharing raw data, thereby preserving confidentiality while improving generalization.

- **Energy-Efficient Edge AI**: Optimizing models for on-board deployment with low-power hardware accelerators to enable continuous, real-time anomaly monitoring without compromising aircraft resources.
- **Adaptive Feedback Loops**: Allowing models to evolve during operation by leveraging pilot reports, maintenance logs, and post-flight reviews to dynamically refine anomaly classification.
- **Transparent and Trustworthy AI**: Embedding explainability mechanisms into models so that pilots, engineers, and regulators can clearly understand why a particular anomaly was flagged, building trust in automated systems.

REFERENCES

1. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104.
2. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation, 13*(7), 1443–1471.
3. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231.
4. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR), 41*(3), 1–58.
5. Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE, 11*(4), e0152173.
6. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications, 60*, 19–31.
7. Aggarwal, C. C. (2017). *Outlier Analysis*. Springer.
8. Ma, J., & Perkins, S. (2003). Time-series novelty detection using one-class support vector machines. *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 1741–1745.
9. Angiulli, F., & Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. *European Conference on Principles of Data Mining and Knowledge Discovery*, 15–27. Springer.
10. Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing, 99*, 215–249.
11. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422.
12. Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., ... & Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery, 30*(4), 891–927.
13. Lee, W., & Stolfo, S. J. (1998). Data mining approaches for intrusion detection. *Proceedings of the 7th USENIX Security Symposium*, 6, 79–94.
14. Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., & Srivastava, J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. *Proceedings of the SIAM International Conference on Data Mining*, 25–36.
15. He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters, 24*(9-10), 1641–1650.
16. Schubert, E., Zimek, A., & Kriegel, H. P. (2014). Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery, 28*(1), 190–237.
17. Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine Learning, 54*(1), 45–66.
18. Kwon, D., Kim, H., & Kim, J. (2017). A survey of deep learning-based network anomaly detection. *Cluster Computing, 22*(1), 949–961.
19. Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 5*(5), 363–387.
20. Xu, X., & Wang, Q. (2005). An adaptive network intrusion detection method based on PCA and support vector machines. *Proceedings of the International Conference on Machine Learning and Cybernetics*, 3, 1425–1429.