Fault Detection of Industrial Gas Turbine Engines

Garey Salinas
garey.salinas@colorado.edu
University of Colorado-Boulder
Boulder, Colorado, USA

Abstract

This project develops a fault detection system for industrial gas turbine engines using real-world sensor data from Kaggle. A robust preprocessing pipeline was built to engineer features (e.g., pressure ratios), normalize inputs, reduce dimensionality via PCA, and identify outliers using Isolation Forest and DBSCAN. Classification models—including Logistic Regression, Random Forest, XGBoost, and SVM will be trained and evaluated using k-fold cross-validation. Model explainability will use SHAP to identify features driving early fault onset. The proposed framework supports predictive maintenance in midstream pipeline operations by improving early failure detection, reducing downtime, and enabling targeted interventions.

Keywords

fault detection, gas turbines, feature scaling, feature engineering, dimensionality reduction, Principle Component Analysis, machine learning

1 Introduction

Industrial gas turbine engines are crucial in midstream pipeline operations, powering gas compressors. Unplanned failures can lead to significant operational disruptions, economic losses, and safety hazards. This project, however, takes a proactive approach. It explores sensor-based predictive maintenance using machine learning to detect faults before they escalate into serious issues. The use of data mining for predictive diagnostics not only enhances reliability but also reduces maintenance costs and downtime.

2 Problem Statement

While many machine learning approaches effectively classify known failure types, relatively few focus on identifying early-stage degradation that precedes faults. This is particularly true in scenarios where labeled failure data is limited. In industrial settings, vast amounts of operational sensor data are collected, but most are unlabeled. This creates an opportunity to develop models that leverage labeled and unlabeled data to detect early warning signs.

There is a clear need for scalable, interpretable fault detection systems that combine supervised classification with unsupervised anomaly detection. Such systems must produce reliable predictions while offering transparency into which features drive those predictions. This project aims to support proactive maintenance decision-making, minimize unexpected downtime, and improve operational efficiency in gas turbine-based pipeline systems by building a hybrid and explainable framework.

3 Literature Review

The use of sensor data for predictive maintenance and fault detection has received significant attention in the industrial domain. As noted by Lei et al. [5], machine learning algorithms such as Support Vector Machines (SVM), Random Forests (RF), and deep learning models have been widely applied in the diagnostics of rotating machinery. These models demonstrate strong performance using snapshot sensor data, especially when key features such as pressure and temperature ratios are engineered to capture system behavior.

Outlier detection is also a cornerstone of predictive maintenance. Techniques such as Isolation Forest [6] and DBSCAN [2] have effectively identified anomalous observations that deviate from typical operating conditions. These methods are especially valuable when labeled fault data is scarce or imbalanced, helping uncover operational anomalies that might not be captured by supervised learning alone.

Recent industry insights from IoT Analytics [4] emphasize the growing demand for interpretable and scalable predictive maintenance solutions. Similarly, a comprehensive review published in *Applied Sciences [1] surveys machine learning techniques used in gas turbine diagnostics, reinforcing the applicability of the methods chosen in this project.

Despite advancements in fault classification, many existing models lack transparency and rely heavily on fully labeled datasets. This limits deployment in operational settings where interpretability and human-in-the-loop decision-making are essential. This project addresses these gaps by combining classical machine learning with explainability tools like SHAP to enhance trust and usability in real-world industrial environments.

4 Proposed Work

This project proposes developing and evaluating supervised machine learning models to classify faults in industrial gas turbine engines. The selected models are Logistic Regression, Random Forest, XGBoost, and Support Vector Machines (SVM). These models are well-suited for structured sensor data and strike a balance between interpretability and predictive performance.

We will apply feature normalization using the StandardScaler to ensure consistent input scaling across features. Dimensionality reduction will be performed using Principal Component Analysis (PCA) to remove noise, reduce feature redundancy, and facilitate visualization of class separability. We aim to retain principal components that preserve at least 95%

To address the presence of anomalous data, we will use Isolation Forest and DBSCAN for outlier detection. Rather than removing outliers entirely, we will encode outlier status as new features, allowing models to account for atypical behavior during training and improve robustness.

We will apply SHAP (Shapley Additive explanations) to tree-based and kernel-based models to enhance their interpretability. This will help identify sensor-derived and engineered features, like pressure ratios and temperature ratios. These are some of the most indicative of early fault conditions, providing actionable insights for domain experts.

The dataset used in this project is the Kaggle Gas Turbine Engine Fault Detection Dataset [3], which contains labeled sensor readings (e.g., temperature, pressure, RPM) collected from gas turbine engines. Its combination of real-world signals and labeled fault conditions makes it ideal for training and evaluating supervised learning models.

The goal is to build a robust classification framework capable of detecting gas compressor failures. This is particularly relevant for midstream pipeline operations, where turbines serve as the primary driving mechanism for compression systems.

Dataset

- **Source:** Kaggle Gas Turbine Engine Fault Detection Dataset
- https://www.kaggle.com/datasets/ziya07/gas-turbine-engine-fault-detection-dataset

Data Preprocessing

- Missing value handling
- Feature Scaling using StandardScaler
- Feature engineering (pressure, speed, and temperature ratios etc.)

Modeling Techniques

- **Supervised Models:** Logistic Regression, Random Forest, XGBoost, Support Vector Machines
- **Unsupervised Models:** Isolation Forest, DBSCAN

Tools & Libraries

- Python (Pandas, Scikit-learn, XGBoost, Seaborn)
- Jupyter Notebook
- Optional: Streamlit for interactive results dashboard

5 Evaluation Plan

- **Classification Metrics**
  - Accuracy
  - Precision
  - Recall
  - F1-Score
- **Evaluation Metrics**
  - Confusion Matrix
  - Visualization of Anomalies
- **Validation Strategy**
  - K-Fold Cross-Validation to ensure robust model performance across different subsets of the data
- **Interpretability**
  - Use SHAP values or feature importances to explain and interpret model predictions

6 Timeline

This timeline outlines the key phases of the project, beginning with literature review and data preparation, followed by model development and anomaly detection. The final days are allocated to compiling results and preparing the final report and presentation. Table 1 summarizes the major project milestones. The project is scheduled for completion by April 28, 2025, aligning with the course submission deadline.

7 Conclusion

In this project, we will propose a predictive maintenance approach for industrial gas turbines that leverages real-world sensor data with machine-learning techniques. The framework will recognize patterns to improve fault detection. The findings from this work highlight the potential for significant impact on industrial monitoring and maintenance optimization. A complete list of cited works can be found in the References section.

8 Citations and Bibliographies

This project builds on a range of research studies and industry data sources. The foundational background includes a comprehensive survey on anomaly detection by Chandola et al. [2] and a systematic review of machinery health prognostics by Lei et al. [5]. Isolation Forest, introduced by Liu et al. [6], is one of the primary algorithms utilized in this project for unsupervised anomaly detection.

Recent market insights from IoT Analytics [4] underscore the increasing demand for predictive maintenance systems across industrial applications. Additionally, a review published in *Applied Sciences* [1] provides a current overview of machine learning models specifically applied to gas turbine anomaly detection, reinforcing the relevance of the selected techniques.

The dataset used for experimentation is the Gas Turbine Engine Fault Detection Dataset hosted on Kaggle [3], which provides real-world sensor data and labeled fault types suitable for supervised classification tasks.

References

[1] Various Authors. 2024. Review of Machine Learning Models for Gas Turbine Anomaly Detection. *Applied Sciences* 14, 11 (2024), 4551. doi:10.3390/app14114551

[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Computing Surveys (CSUR)* 41, 3 (2009), 1–58. doi:10.1145/1541880.1541882

[3] Ziyao Chen. 2020. Gas Turbine Engine Fault Detection Dataset. https://www.kaggle.com/datasets/ziya07/gas-turbine-engine-fault-detection-dataset. Accessed: 2025-04-16.

[4] IoT Analytics. 2023. Predictive Maintenance Market Report 2023. https://iot-analytics.com/predictive-maintenance-market. Accessed: 2025-04-16.

[5] Yaguo Lei, Naipeng Li, Lin Guo, Ning Li, Tao Yan, and Jing Lin. 2020. Machinery Health Prognostics: A Systematic Review from Data Acquisition to RUL Prediction. *Mechanical Systems and Signal Processing* 138 (2020), 106761. doi:10.1016/j.ymssp.2019.106761

[6] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*. IEEE, 413–422. doi:10.1109/ICDM.2008.17

Table 1: Proposed Project Timeline

Phase	Dates	Tasks
Project Planning & Literature Review	Apr 16 – Apr 18	Finalize proposal, conduct literature review, and outline the modeling approach
Data Preprocessing	Apr 18 – Apr 21	Clean dataset, perform exploratory data analysis (EDA), scale features, engineer new variables, apply PCA, and detect outliers
Model Building, Evaluation, Feature Importance	Apr 21 – Apr 24	Train and evaluate classification models (Logistic Regression, Random Forest, XGBoost, SVM), assess performance metrics, and analyze feature importance using SHAP
Final Report & Presentation	Apr 25 – Apr 28	Compile findings, create visualizations, write the final report, and prepare presentation slides