**In the Heartbeat of Metro Trains: Anomaly**

**Detection and Failure Prediction Unveiled**

Website link: <https://mason.gmu.edu/~vrakurth/>

***Abstract* — In the dynamic landscape of rail transportation, maintaining the reliability and safety of critical components such as compressors is paramount. Compressor failures can lead to costly downtime, operational disruptions, and, most importantly, compromise passenger safety. Leveraging advancements in data analytics, this research aims to identify the key factors contributing to compressor failures in metro trains. By analyzing sensor data and failure reports from the MetroPT-3 dataset, we endeavor to uncover patterns and correlations that can inform predictive maintenance strategies. Utilizing a Decision Tree Classifier, we partitioned our data into training and test sets, with the target variable being "Compressor," and eight sensors influencing its failure. Through thorough analysis, we identified "MPG," "Tower," and "Caudal impulses" as the most significant predictors, resulting in the highest predictive accuracy. To ensure the reliability and generalization of our model, a five-fold cross-validation technique was employed on the training set, mitigating overfitting, and enhancing the model's robustness. This research contributes to advancing predictive maintenance strategies within metro train systems, providing actionable insights to stakeholders to mitigate risks and uphold passenger safety. By leveraging data analytics techniques on the MetroPT-3 dataset, our study sheds light on the complex dynamics of compressor failures, empowering rail transportation stakeholders with knowledge to enhance system reliability and operational efficiency.**

Keywords—compressors, failures, sensor

# Introduction

Rail transportation is a cornerstone of modern economies, facilitating the movement of goods and people across vast distances with efficiency and reliability. As countries invest in expanding and modernizing their rail networks, ensuring the safety and functionality of critical components like compressors in metro trains becomes paramount. Compressors play a vital role in maintaining the proper operation of trains, making their failure a significant concern for both safety and operational efficiency.

Analyzing sensor data and failure reports can provide invaluable insights into the factors contributing to compressor failures. By understanding these key factors, proactive maintenance strategies can be developed to prevent catastrophic accidents and financial losses. Leveraging advanced data analytics and machine learning techniques, such as predictive maintenance models, becomes essential in this endeavor. This paper aims to address the pressing need for dependable and effective transportation systems, particularly in the railroad sector. By focusing on compressor failures and utilizing datasets like MetroPT-3, which contains sensor data from trains, this research seeks to create accurate predictive models to anticipate failures. By doing so, maintenance costs can be reduced, downtime minimized, and the overall reliability of railway systems enhanced.

Moreover, the findings and methodologies developed in this research have broader implications for predictive maintenance in the railroad sector. Engineers and maintenance staff can benefit from a comprehensive understanding of the elements leading to compressor failures, allowing for more efficient maintenance practices. Furthermore, the insights gained can be extrapolated to other critical railway components, such as bearings and brakes, thereby further improving system efficiency and dependability.

In summary, this paper presents a comprehensive approach to anticipating compressor failures in the railway sector through the application of machine learning models and advanced data analytics techniques. By leveraging available data and predictive maintenance methodologies, this research aims to enhance the safety, reliability, and efficiency of rail transportation systems.

# Liturature review

1) Review on the Traction System Sensor Technology of a Rail Transit Train: The survey highlights the vital role of sensors in collecting essential data like voltage, current, speed, and temperature for ensuring the efficiency and safety of rail transit systems. It explores the complexities involved in signal acquisition and processing, discussing challenges related to diverse signal types and susceptibility to interference. Additionally, the survey emphasizes strategies to overcome these challenges, including analog signal processing intricacies, sensor data sampling, digital filtering technologies, and sensor fault diagnosis methods, advocating for robust sensor systems to maintain the integrity and functionality of rail transit systems. [2].

2)Recent applications of big data analytics in railway transportation systems: A recent literature review examines the application of Big Data Analytics in the railway sector, summarizing its use across operations, maintenance, and safety applications. While the focus primarily centers on assessing infrastructure health, significant attention is also given to challenges related to train components, weather conditions, geographical positioning, and other variables. Employing Mayring's content analysis methodology on 115 articles, the review establishes a classification framework across specific Railway Transportation Systems domains, depth of analytics, types of models utilized, and BDA techniques applied, identifying research gaps and future directions. This analysis highlights the extensive utilization of BDA within RTS, emphasizing its role in improving operational efficiency, maintenance practices, and safety protocols in the railway industry. [3].

3)Wireless Sensor Networks for Condition Monitoring in the Railway Industry: The Survey This study examines the application of Wireless Sensor Networks (WSNs) in railway condition monitoring, emphasizing practical engineering solutions for both stationary and mobile monitoring systems. Sensors play a crucial role in facilitating unbiased and comprehensive data collection to evaluate the health and longevity of railway infrastructure and rolling stock. Challenges highlighted encompass sensor choice, optimization of network topology, and energy consumption efficiency. Various techniques, including ambient energy harvesting and event detection, are investigated to address these challenges. Findings demonstrate improved reliability and precision in condition monitoring, suggesting potential applications in fault detection and long-term assessment of structural health in railway systems [4].

4) A Global Manufacturing Big Data Ecosystem for Fault Detection in Predictive Maintenance Wenjin Yu et al. employed anomaly detection for predictive maintenance, presenting a four-layer architecture. This framework comprised big data intake, management, analytics, and visualization levels, encompassing functionalities from IoT data acquisition to real-time system condition monitoring. The significance of visual analytics was underscored, particularly in the monitoring stage, where engineers monitored compressor conditions post-anomaly detection. A deterministic mechanism was implemented, where timestamps transitioned from '0' to '1' if over 15 anomalies were detected out of 300 observations within five-minute windows. The engineer's involvement helped mitigate the risk of false alarms, enhancing the reliability of anomaly detection systems. This approach highlights the integration of data analytics and human oversight to optimize predictive maintenance strategies, ensuring timely and accurate identification of potential faults in critical components [5].

5) Metro Rail — Predictive Maintenance Based on Anomaly Detection This paper investigates the efficacy of autoencoders for anomaly detection in both digital and analog signals, assessing sparse autoencoders and variational autoencoders on industrial machinery data. Results reveal that sparse autoencoders excel in accuracy and recall with digital signals, while variational autoencoders exhibit slightly lower precision due to misclassification of normal data. Analog models, however, demonstrate notably inferior performance in precision and recall. Additionally, the paper explores advanced anomaly detection techniques like LSTM neural networks, transformers, and causal inference methods, suggesting their potential to enhance detection accuracy and offer deeper insights into anomaly causes [6].

6) The MetroPT dataset for predictive maintenance This paper offers a comprehensive analysis of the MetroPT dataset, which records failure events within a metro train system, detailing failure types, involved components, and times. It introduces an evaluation protocol based on True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) metrics, aiming to minimize false alarms and missed failures. Additionally, it discusses two recent studies utilizing the dataset, one employing a rule-based system for compressor state alerts and the other exploring deep learning autoencoders for failure prediction, both indicating potential for accuracy and explanation enhancement [7].

7) A Survey on Data-Driven Predictive Maintenance for the Railway Industry This paper provides an overview of recent research on machine learning (ML) and deep learning (DL) algorithms for prognostics and health management (PdM) in the railway industry, focusing specifically on vehicles like trucks and railcar wheels. It discusses challenges encountered by both academia and industry in implementing ML/DL algorithms for PdM in this sector. The paper reviews various methodologies, including Random Forest, Support Vector Machine, Long Short-Term Memory classifier, and mapping function using Random Forest regression model, with performance evaluated using metrics like Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) [8].

8) MetroPT-3 Anomaly Detection using Machine Learning and Deep Learning The paper focuses on leveraging the MetroPT-3 dataset for predictive maintenance and anomaly detection, particularly in compressor systems. This dataset is tailored to enable the development of models that can predict the remaining useful life (RUL), detect anomalies, and facilitate predictive maintenance practices. The study employs a diverse set of machine learning algorithms to address the challenges associated with compressor health monitoring systems. These algorithms range from traditional approaches such as Linear Regression, KNN, Random Forest, and Support Vector Machine, to more advanced techniques like Naive-Bayes classification and XGBoost. Additionally, the study explores the efficacy of Extreme Machine Learning Models, Deep Learning, and an Ensemble model comprising the top three performing algorithms. The goal of this comprehensive approach is to enhance predictive maintenance practices and contribute to the evolution of compressor health monitoring systems using various machine learning paradigms [9].

9) MetroPT: A Benchmark dataset for predictive maintenance The MetroPT dataset is a product of the explainable Predictive Maintenance (XPM) project conducted in Porto, Portugal, involving an urban metro public transportation service. Collected in 2022, this dataset serves to assess machine learning techniques for online anomaly detection and failure prediction. It encompasses several types of data, including analog sensor signals (such as pressure, temperature, and current consumption), digital signals (including control and discrete signals), and GPS information (latitude, longitude, and speed). This dataset's unique features make it suitable for evaluating online machine learning methods, and it is considered a valuable benchmark for developing predictive maintenance models [10].

While much existing literature addresses predictive maintenance, anomaly detection, and condition monitoring in railway systems, our research focuses specifically on understanding the key factors behind compressor failures. Previous studies have explored sensor technology, big data analytics, wireless sensor networks, and machine learning in railways, but our work delves uniquely into the root causes of compressor issues. Through detailed analysis of sensor data and failure reports, we aim to identify critical variables like voltage fluctuations, temperature abnormalities, and mechanical stress that led to compressor malfunctions. Our study offers targeted insights into the challenges specific to compressor systems, providing actionable information to improve predictive maintenance strategies and optimize compressor health monitoring in rail transit systems.

# Dataset

The MetroPT-3 dataset represents a valuable resource for researchers and practitioners in the field of predictive maintenance and anomaly detection within metro train systems. The dataset provides insights into the broader operational context of metro train operations. It captures the intricate interplay between various system components and environmental factors, shedding light on the complex dynamics involved in ensuring the reliability and safety of metro train operations.

In addition to its utility in model development for predictive maintenance and anomaly detection, the dataset offers opportunities for exploring the broader implications of data-driven approaches in railway industry applications. By encompassing a diverse array of signals from both analog and digital sensors, the dataset provides a comprehensive view of the operational status and performance metrics of the compressor system. Researchers can leverage this rich source of information to uncover hidden patterns, identify optimization opportunities, and enhance overall system efficiency.

Furthermore, the dataset's temporal coverage from February to August 2020 allows for the analysis of seasonal variations, long-term trends, and event occurrences within the metro train system. This temporal dimension enables researchers to investigate the dynamic nature of system behavior and its evolution over time, offering valuable insights into the underlying mechanisms driving performance degradation, fault propagation, and maintenance requirements.

The dataset was gathered to assist in creating models for predictive maintenance, anomaly detection, and predicting the remaining useful life (RUL) of compressors using deep learning and machine learning techniques. It contains multivariate time series data from both analog and digital sensors installed on a train's compressor. This data covers the period from February to August 2020 and comprises 15 signals, including pressures, motor current, oil temperature, and electrical signals from air intake valves. The dataset also includes records of industrial equipment events such as temporal behavior and fault events, which were logged by the sensors. Data logging was done at a frequency of 1Hz using an onboard embedded device.[1]

# Priliminary Results

After loading the dataset into RStudio, we proceeded to summarize its contents.

A screenshot of a computer

Description automatically generated

Figure1 summary statistics

From the summary statistics provided in Figure1, it is evident that the attribute "Towers," which indicates the activation status of a tower, shows the presence of towers across the dataset, as inferred from the mean value close to 1. Regarding the flow rate represented by "LPS", the mean value of 0.00342 suggests a low average flow rate, with most observations showing no flow (indicated by quartiles). Conversely, attributes such as "Pressure\_switch" and "Oil\_level" show high mean values of 0.9914 and 0.9042, respectively, indicating their frequent activation or sufficiency in most observations. This summary underscores the prevalence of certain conditions or statuses within the dataset, providing insights into the operational dynamics of the system under study.

A graph showing a number of timestamp

Description automatically generated

Figure 2 DV\_Pressure over time

When we plot (Figure 2) DV\_Pressure over time, it shows at which month the DV\_Pressure is highest (April) and lowest (February). DV pressure represents the drop in pressure when compressed air is released from air dryers. These dryers remove moisture and impurities from compressed air, ensuring it's clean and dry. A zero DV pressure means no drop when air is released, indicating the compressor is working well. Higher readings suggest the compressor is under heavy load.

A graph of oil temperature

Description automatically generated

Figure 3 violin graph of Oil Temperature with Compressors.

When we plot (Figure 3) a violin graph of Oil Temperature with Compressors. Compressor indicating 0 is a failure and 1 is success. We can see that with compressor failure, there are more outliers, and with Success, there are fewer outliers when compared to other ones.

A graph with blue and red lines

Description automatically generated

Figure 4 line graph between Oil Temperature and Compressor

When plotting a graph (Figure 4) between Oil Temperature and Compressor Failure over Time, we can observe that Success rate is much less when compared to the failure rate. And at what frequent times does the compressor fail due to Oil Temperature.

A graph with a bar and a number

Description automatically generated with medium confidence

Figure 5 compressor failure against average motor current

When graphing (figure 5) compressor failure against average motor current, a striking observation emerges: the success rate of motor current readings is approximately three times higher than the failure rate. This significant difference underscores the potential correlation between motor current levels and compressor performance. The higher success rate suggests that typical motor current levels are more commonly associated with operational success, while deviations from these norms may signal impending compressor failures. This insight highlights the importance of monitoring the motor current as a predictive indicator for compressor health, enabling proactive maintenance interventions to prevent failures and ensure system reliability.

A graph with lines and dots

Description automatically generated

Figure 6 Control chart

Examining the precise Oil Temperature readings against timestamps (Figure 6) provides granular insights into the temporal patterns of temperature fluctuations and their impact on compressor failures. Leveraging visualization tools like Qlik Sense allows for detailed examination down to the hour, minute, and second levels. This level of temporal granularity enables pinpointing specific instances when oil temperature variations coincide with compressor failures, facilitating a deeper understanding of the temporal dynamics driving compressor malfunctions. By visualizing the exact timing of temperature spikes or drops to failure events, operators can identify critical periods of vulnerability and implement targeted interventions to optimize compressor performance and mitigate failure risks effectively.

A graph of a line graph

Description automatically generated with medium confidence

Figure 7 lag plot

This graph (Figure 7) is called a lag plot, and it's used to visualize and analyze the behavior of time series data. The x-axis represents the lag value, which is the number of time steps or periods the data is shifted compared to itself. The y-axis shows the corresponding data values at each time step. Each point on the plot compares the value of the data series at a given time (y-axis) against its value at a previous time step determined by the lag value on the x-axis. The diagonal pattern we see here indicates that the data exhibits strong autocorrelation or temporal dependence. In other words, the values at different time steps are highly correlated with their lagged counterparts.

A graph showing a graph of oil temperature and motor current

Description automatically generated

Figure 8 oil temperature and motor current

From Figure 8 the oil temperature is represented by a line that fluctuates significantly over time, indicating that the temperature of the oil changes quite a bit. On the other hand, the motor current is represented by a line that remains relatively stable, suggesting that the current doesn’t change much.

A screenshot of a computer

Description automatically generated

A close-up of a pie chart

Description automatically generatedFigure 9 Compressor failure and Compressor success

For Compressor Failure: The most significant factors contributing to compressor failure appear to be Towers, Caudal impulses, Oil level, and Pressure switch, each accounting for around 20% of the failures.MPG and DV electric have relatively lower contributions, with MPG contributing 20.5% and DV electric only contributing 0.1%. LPS does not contribute to compressor failure based on the provided data.

For Compressor Success: In contrast, the distribution of sensors for compressor success is different. DV electric, Pressure switch, and Towers are the most significant factors contributing to compressor success, each accounting for around 22% of the successes. Caudal impulses and Oil level also have notable contributions, with Caudal impulses contributing 21.4% and Oil level contributing 21%. MPG and LPS do not contribute to compressor success based on the provided data.

Overall, Towers, DV electric, Caudal impulses, Oil level, and Pressure switch seem to play crucial roles in both compressor failure and success, albeit with different magnitudes.

A blue and white squares with numbers

Description automatically generated

Figure 10 Confusion Matrix

Based on the provided confusion matrix:

True Positives (TP): 380,715 True Negatives (TN): 74,367 False Positives (FP): 2 False Negatives (FN): 1 From this, we can analyze the performance of the model:

High True Positive (TP) rate indicates that the model is effective at correctly predicting compressor failures. Low False Positive (FP) rate suggests that the model rarely misclassifies successful compressor operations as failures, which is desirable. The False Negative (FN) rate is very low, indicating that the model rarely fails to identify actual compressor failures. Overall, the model seems to perform well in predicting compressor failures, with a high level of accuracy and a low rate of misclassifications.

A number on a white background

Description automatically generated

Figure 11 Algorithm Used

We used Decision tree Classifier and split our data into the Train and Test model.

Here, our target variable is “Compressor”, and we have 8 sensors that are affecting its failure.

By considering “MPG, Tower, Caudal\_impulses”, we achieved the highest accuracy.

Cross-validation with a five-fold split is employed on the training set to ensure the model's generalization and reliability.

# Proposed approach

# 

Figure 9 System architecture[12]

## Data Integration:

   - Integrate the time-stamped sensor data with corresponding failure reports to create a unified dataset linking sensor readings with instances of compressor failures.

## Data Preprocessing and Normalization:

   - Cleanse the dataset to rectify anomalies and handle missing values.

   - Normalize the sensor readings to ensure uniform scaling across all features, facilitating effective analysis.

## Feature Engineering:

   - Extract relevant features from the dataset that may contribute to compressor failures.

   - Engineer new features to capture complex relationships and patterns in the data.

   - Techniques such as trend analysis, Fourier transformations for cyclic patterns, and anomaly detection will be employed to enhance feature representation.

## Exploratory Data Analysis (EDA) & Visualization:

   - Conduct EDA to gain insights into the distribution and characteristics of the data.

   - Utilize visualizations such as histograms, box plots, and heatmaps to identify outliers, understand feature correlations, and detect any underlying patterns that may influence compressor failures.

## Machine Learning Model Selection:

   - Decision trees or logistic regression are considered due to their simplicity and interpretability, making them ideal for initial analysis.

## Model Training and Evaluation:

   - Split the dataset into training and testing sets for model training and evaluation.

   - Train the selected machine learning model on the training data.

   - Evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, and F1-score.

   - Employ cross-validation techniques to ensure robustness of the model's performance.

## Model Refinement and Optimization:

   - Iteratively refine the model based on feedback from evaluation results.

   - Fine-tune hyperparameters to optimize the model's performance.

   - Implement techniques such as feature selection and regularization to enhance model generalization.

# Timeline

|  |  |  |
| --- | --- | --- |
| Done | Brainstorm Project Ideas | Week3 -Week4 |
| Done | Project Proposal | Feb 18, 2024 |
| Done | Data Integration | Week 5 |
| Done | Preprocessing | Week 6 |
| Done | Normalization | Week7 |
| Done | Feature Engineering | Week 8 – Week9 |
| Done | Milestone 1 | Mar 31, 2024 |
| Done | Exploratory Data Analysis | Week10 -Week11 |
| Done | Visualization | Week 12 |
| Done | Milestone 1 | Apr 14, 2024 |
| Done | Statistical Analysis and Hypothesis Testing | Week 13 |
| Done | Refinement | Week 14 |
| Done | Project submission and  presentation | Week 15 |

# Conclusion

The conclusion of the provided information is that investing in the maintenance and modernization of rail networks is essential for economic growth and safety. Neglecting these aspects can lead to serve consequences, including accidents and financial losses. The historical advancements in manufacturing, driven by innovations like AI and machine learning, highlight the importance of continuous improvement and adaptation in the rail industry. Utilizing smart machines and networked systems can enhance maintenance practices and decision-making, ultimately improving efficiency and safety across rail networks.

# Future work

Continued innovation, such as AI and machine learning, will enhance rail safety, efficiency, and predictive maintenance, ensuring smoother operations and reduced downtime.

Countries will invest in expanding and modernizing their rail networks to meet growing demand, improving connectivity, and reducing congestion.

The focus will be on developing sustainable rail systems, including electrification, reducing emissions, and increasing energy efficiency.

# Acknowledgement

To professor Dr. Ebrima N. Ceesay: Your guidance was a beacon, your lectures a revelation, and your support unwavering.

To our TA Grace Boddu: Your dedication and timely feedback were the cornerstone of our technical success.

To our colleagues: Your collaborative spirit and insightful feedback transformed our learning journey.

To the research authors: Your work illuminated our path and shaped our approach.

Together, you have all left an indelible mark on our project, for which we are profoundly grateful.

##### References

1. Dataset: MetroPT-3 Dataset UCI Machine Learning Repository. (n.d.). <https://archive.ics.uci.edu/dataset/791/metropt+3+dataset>
2. Feng J, Xu J, Liao W, Liu Y. Review on the Traction System Sensor Technology of a Rail Transit Train. Sensors (Basel). 2017 Jun 11;17(6):1356. doi: 10.3390/s17061356. PMID: 28604615; PMCID: PMC5492406.
3. Ghofrani, F., He, Q., Goverde, R. M. P., & Liu, X. (2018). Recent applications of big data analytics in railway transportation systems: A survey. Transportation Research Part C: Emerging Technologies, 90, 226-246. <https://doi.org/10.1016/j.trc.2018.03.010>
4. Wireless Sensor Networks for Condition Monitoring in the Railway Industry: The Survey IEEE Xplore Full-Text PDF: (n.d.). <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6963375>
5. W. Yu, T. Dillon, F. Mostafa, W. Rahayu, and Y. Liu, "A Global Manufacturing Big Data Ecosystem for Fault Detection in Predictive Maintenance," in IEEE Transactions on Industrial Informatics, vol. 16, no. 1, pp. 183-192, Jan. 2020, doi: 10.1109/TII.2019.2915846.
6. Gupta, V. (2021, May 11). Metro Rail Predictive Maintenance Based on Anomaly Detection. Medium. <https://medium.com/@vgupta701/metro-rail-predictive-maintenance-based-on-anomaly-detection-0008ffa7a5b7>
7. Kaur, S., & Kaur, S. (2021). The MetroPT dataset for predictive maintenance. ResearchGate. <https://www.researchgate.net/publication/366219552_The_MetroPT_dataset_for_predictive_maintenance>
8. "Narjes Davari, Bruno Veloso, Gustavo de Assis Costa, Pedro Mota Pereira, Rita P. Ribeiro, and João Gama. 'A Survey on Data-Driven Predictive Maintenance for the Railway Industry.' Sensors 21, no. 17 (2021): 5739. <https://doi.org/10.3390/s21175739>"
9. harveyphm. (n.d.). MetroPT-3-Anomaly-Detection. GitHub. <https://github.com/harveyphm/MetroPT-3-Anomaly-Detection>
10. Zenodo. (2022). MetroPT: A Benchmark dataset for predictive maintenance [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.6854240>
11. Angelopoulos, A., Michailidis, E. T., Νομικός, Ν., Trakadas, P., Hatziefremidis, A., Voliotis, S., & Zahariadis, T. (2019). Tackling Faults in the Industry 4.0 Era—A survey of Machine-Learning Solutions and Key aspects. Sensors, 20(1), 109. https://doi.org/10.3390/s20010109
12. *Figure Generated using Luciid chart*. Lucid visual collaboration suite: Log in. (n.d.). https://lucid.app/lucidchart/1790ee8b-8a61-4830-bb4c- 34d1a6e026dc/edit?invitationId=inv\_5147e080-6a60-46b1-b0ac- e4ab1eca4d26&page=m-5o7ONTd-nK#