In the Heartbeat of Metro Trains: Anomaly Detection and Failure Prediction Unveiled

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

Department of Computer Science & Engineering
AIT 664
April 13, 2024

Geethika Pendyala
George Mason University
Fairfax, Virginia
gpendyal@gmu.edu

Neha Reddy Yenugu
George Mason University
Fairfax, Virginia
nyenugu@gmu.edu

Venkatesh Rakurthi
George Mason University
Fairfax, Virginia
vrakurth@gmu.edu

***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

I. 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.

II. 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.

III. 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.

Attribute Information –

1. **TP2 (bar)**: This attribute records the pressure inside the compressor, providing insights into the operational status and any potential pressure abnormalities that might suggest issues or optimal functioning.
2. **TP3 (bar)**: Similar to TP2, this measures pressure but at the pneumatic panel. It helps in verifying the consistency and reliability of the pressure being relayed through the system, ensuring that it aligns with the compressor's output.
3. **H1 (bar)**: This gauges the pressure drop that happens during the discharge of the cyclonic separator filter. It is essential to comprehend how well the separator maintains pressure stability by removing other impurities and particulates.
4. **DV Pressure (bar)**: This characteristic measures the pressure drop that occurs when air dryer towers release air. Any increase could point to a load imbalance or mechanical inefficiency, while a zero reading shows the compressor is loaded and operating as it should.
5. **Reservoirs (bar)**: Measures the downstream pressure of the reservoirs, typically expected to be close to TP3, thus ensuring that the pressure throughout the system is stable and within safe operating limits.
6. **Motor Current (A)**: This current measurement of a motor phase provides insights into the operational state of the motor, ranging from off and idle states

to under load and start-up conditions, helping diagnose electrical issues or motor efficiency.

7. **Oil Temperature (ºC)**: Monitors the temperature of the oil lubricating the compressor, which is essential for preventing overheating and ensuring the motor and other components operate within safe thermal parameters.
8. **COMP**: This is an electrical signal from the air intake valve, indicating whether the compressor is off or idle. It's a direct indicator of operational status, useful for automated systems to adjust flows and loads.
9. **DV Electric**: Controls the outlet valve of the compressor, becoming active under load conditions and inactive when the compressor is off or idle. This helps in managing the compressor's response to varying operational demands.
10. **TOWERS**: This signal manages the operation of the air-drying towers, indicating which of the two towers is active. This ensures that the air is continuously dried effectively, cycling between the towers for optimal performance.
11. **MPG**: Activates the compressor under load when the air pressure within the production unit falls below a set threshold (8.2 bars). It syncs with the COMP sensor, providing a fail-safe to maintain constant pressure.
12. **LPS**: This sensor activates if the pressure falls below 7 bars, signaling potentially unsafe operating conditions that could compromise system integrity.
13. **Pressure Switch**: This switch signals discharge activities in the air-drying towers, essential for monitoring and controlling the release cycles to maintain air quality and pressure.
14. **Oil Level**: Monitors the oil level within the compressor, becoming active when oil levels are below expected thresholds. It is vital for preventing mechanical failures due to insufficient lubrication.
15. **Caudal Impulse**: This measures the pulse outputs correlating to the volume of air flowing from the production unit to the reservoirs. It provides a quantitative measure of airflow, crucial for assessing system efficiency and performance.

IV. Priliminary Results

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

```
summary(data)
```

```
      ...1                timestamp                       TP2               TP3               H1        
 Min.   :       0   Min.   :2020-02-01 00:00:00.00   Min.   :-0.032   Min.   : 0.730   Min.   :-0.036  
 1st Qu.: 3792368   1st Qu.:2020-03-23 05:05:04.50   1st Qu.:-0.014   1st Qu.: 8.492   1st Qu.: 8.254  
 Median : 7584735   Median :2020-05-17 08:07:06.00   Median :-0.012   Median : 8.960   Median : 8.784  
 Mean   : 7584735   Mean   :2020-05-16 22:58:36.63   Mean   : 1.368   Mean   : 8.985   Mean   : 7.568  
 3rd Qu.:11377102   3rd Qu.:2020-07-10 03:07:27.50   3rd Qu.:-0.010   3rd Qu.: 9.492   3rd Qu.: 9.374  
 Max.   :15169470   Max.   :2020-09-01 03:59:50.00   Max.   :10.676   Max.   :10.302   Max.   :10.288  
  DV_pressure        Reservoirs      Oil_temperature Motor_current      COMP          DV_eletric      
 Min.   :-0.03200   Min.   : 0.712   Min.   :15.40   Min.   :0.020   Min.   :0.000   Min.   :0.0000  
 1st Qu.:-0.02200   1st Qu.: 8.494   1st Qu.:57.77   1st Qu.:0.040   1st Qu.:1.000   1st Qu.:0.0000  
 Median :-0.02000   Median : 8.960   Median :62.70   Median :0.045   Median :1.000   Median :0.0000  
 Mean   : 0.05596   Mean   : 8.985   Mean   :62.64   Mean   :2.050   Mean   :0.837   Mean   :0.1606  
 3rd Qu.:-0.01800   3rd Qu.: 9.492   3rd Qu.:67.25   3rd Qu.:3.808   3rd Qu.:1.000   3rd Qu.:0.0000  
 Max.   : 9.84400   Max.   :10.300   Max.   :89.05   Max.   :9.295   Max.   :1.000   Max.   :1.0000  
     Towers             MPG              LPS         Pressure_switch    Oil_level      Caudal_impulses 
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.0000  
 Median :1.0000   Median :1.0000   Median :0.00000   Median :1.0000   Median :1.0000   Median :1.0000  
 Mean   :0.9198   Mean   :0.8327   Mean   :0.00342   Mean   :0.9914   Mean   :0.9042   Mean   :0.9371  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
```

Figure1 Summary statistics

The statistics from Figure 1 reveal insightful details about the dataset's attributes. The "Towers" attribute, which indicates which air-drying tower is active, has a mean value close to 1. This frequent switching points to active management, likely to optimize tower performance or for maintenance purposes.

The "LPS" attribute, representing flow rate, shows a notably low average of 0.00342. This indicates that generally, the flow rate is minimal, with most measurements showing no flow, suggesting a tightly controlled system that increases flow only as needed, or a design optimized for low flow operation.

In contrast, the "Pressure_switch" and "Oil_level" attributes have high mean values of 0.9914 and 0.9042, respectively, indicating that pressure in the air-drying towers and oil levels are mostly within ideal ranges. Frequent activation of the pressure switch reflects ongoing efforts to maintain optimal pressure for system efficiency and safety. Similarly, consistently good oil levels suggest effective maintenance practices and possibly automated monitoring to prevent operational disruptions.

Overall, these statistics highlight a well-maintained system with critical factors like oil levels and pressure closely monitored, ensuring the system operates efficiently and adapts to various operational needs.

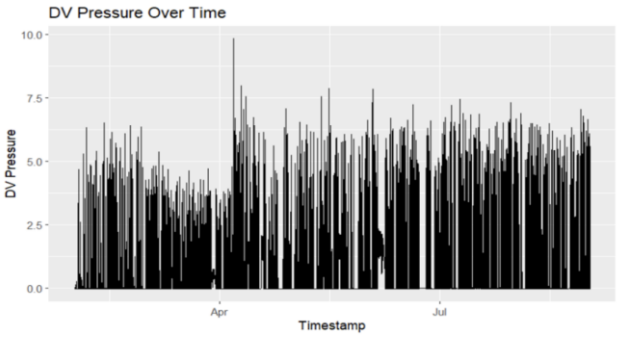

Figure 2 DV_Pressure over time

Figure 2 showcases a time-based graph of the DV_Pressure attribute, highlighting how this pressure varies across different months. The graph reveals that DV_Pressure is lowest in February and peaks in April. DV_Pressure measures the drop in pressure when compressed air, cleared of moisture and impurities by air dryers, is released. A reading of zero suggests no pressure drop and indicates efficient compressor operation under normal loads. Higher readings, however, point to the compressor working harder under heavier loads, which might stem from increased operational demands or system inefficiencies.

The fluctuations in DV_Pressure over the months could be influenced by factors such as higher production requirements or environmental conditions impacting air dryer efficiency. For instance, the lower pressure in February might reflect reduced demand or better operational conditions. Conversely, the spike in April could signal increased system strain.

Understanding these patterns is crucial for effective management of the compressor and air dryer systems. It guides timely adjustments or maintenance to mitigate

inefficiencies or prevent potential breakdowns. Moreover, monitoring these trends helps in planning system upgrades or capacity enhancements for improved resilience and long-term system reliability. This data thus plays a vital role in strategic operational planning and ensuring ongoing system efficiency.

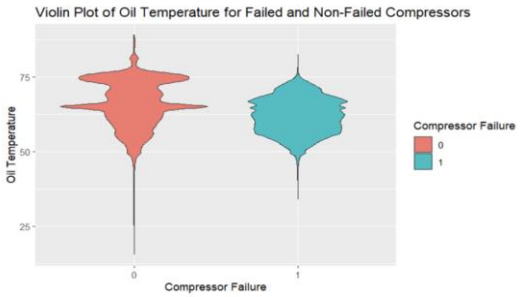

Figure 3 Violin graph of Oil Temperature with Compressors.

Figure 3 presents a violin plot that illustrates oil temperature distributions under two conditions of compressor operation: failure (0) and success (1). This plot merges features of density and box plots, providing a clear view of the data's distribution, central trend, and variability.

The plot shows that compressor failure is associated with a greater number of outliers in oil temperature, suggesting significant deviations from typical values. These extremes in temperature, either too high or too low, may contribute to or indicate compressor failures, highlighting potential operational issues.

In contrast, the plot for successful compressor operation (1) has fewer outliers, indicating a more stable and consistent oil temperature. This stability suggests that the compressor is maintaining oil temperature within a desirable range, which is crucial for effective operation.

The difference in outlier presence between failures and successes is important for predictive maintenance. Monitoring oil temperature for significant changes can help identify and address potential issues before they result in compressor failure. This proactive approach not only enhances compressor efficiency but also extends its lifespan.

Overall, the insights from this violin plot are invaluable for understanding how oil temperature impacts compressor performance and underline the importance of maintaining stable temperatures for reliable compressor operation.

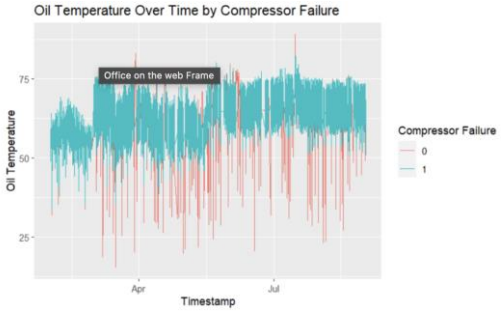

Figure 4 Line graph between Oil Temperature and Compressor

Figure 4 features a graph plotting Oil Temperature against Compressor Failure over time, revealing critical insights into compressor reliability. This graph shows a clear trend where the rate of compressor failures far exceeds the rate of successful operations throughout the observed period. It pinpoints times when high or low oil temperatures coincide with increased failures, suggesting that extreme temperatures are key contributors to compressor malfunctions.

The graph indicates periods of heightened risk, helping maintenance teams identify when and why compressors are most likely to fail. The higher frequency of failures may point to systemic issues such as inadequate maintenance, poor quality oil, or inappropriate operational settings that do not maintain oil temperature within safe limits.

This data is crucial for developing predictive maintenance strategies. By monitoring oil temperature closely and adjusting operations proactively before critical temperatures are reached, it's possible to greatly reduce compressor failures. Such a proactive maintenance approach not only boosts operational efficiency but also prolongs equipment life by preventing excessive wear and damage.

Overall, the insights from this graph are invaluable for enhancing operational management, highlighting the direct link between oil temperature and compressor failures and guiding strategic improvements in maintenance and operational practices to enhance system reliability.

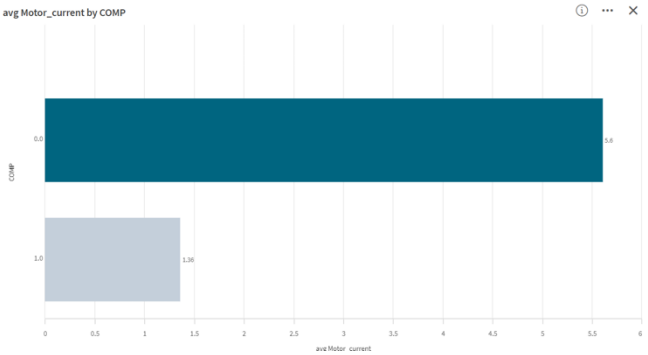

Figure 5 Compressor failure against average motor current

In the graph shown in Figure 5, which compares compressor failure to average motor current, an important trend is observed: the success rate of operations where the motor current remains within typical levels is about three times higher than where it deviates, indicating a likely link between motor current and compressor performance. This finding suggests that maintaining motor current within a standard range generally leads to successful compressor operations, while deviations from this range are often precursors to failures.

The significant disparity in success and failure rates based on motor current levels highlights the importance of continuously monitoring motor current as a predictive tool for assessing compressor health. By keeping track of these readings, maintenance teams can identify unusual changes in current that may indicate potential problems with the compressor. Early detection of such anomalies allows for prompt maintenance actions, which can prevent compressor failures and enhance the overall reliability of the system.

This insight not only aids in optimizing operational efficiency but also supports the development of effective

preventive maintenance strategies. Monitoring motor current can therefore be a key factor in extending the lifespan of compressors and ensuring their consistent and reliable performance.

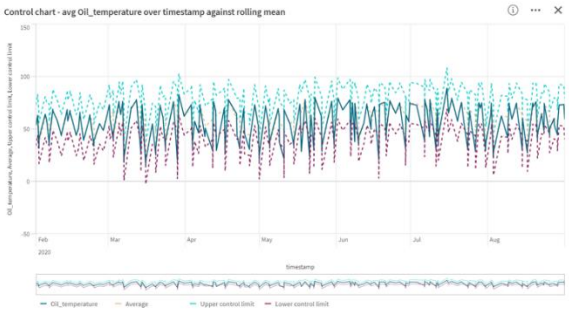

Figure 6 Control chart

The detailed analysis of Oil Temperature readings against specific timestamps, as shown in Figure 6, offers a deep dive into the timing and effects of temperature fluctuations on compressor failures. Utilizing visualization tools like Qlik Sense enhances this analysis by breaking down data to very precise levels—hours, minutes, and seconds. Such detailed temporal insights allow for the identification of exact moments when changes in oil temperature align with compressor failures.

This fine-grained temporal analysis is critical in understanding how and when compressor issues are most likely to occur. By mapping out the exact timing of temperature anomalies—whether spikes or drops—relative to compressor failures, maintenance teams can pinpoint periods of heightened risk with great accuracy. Recognizing these critical times enables operators to take specific, timely actions to adjust or repair the system before failures occur.

The ability to monitor oil temperature changes in real-time and correlate them directly with operational outcomes helps in creating more effective and targeted maintenance strategies. This proactive approach not only prevents downtime but also optimizes compressor performance, ensuring the system runs smoothly and efficiently while reducing the likelihood of unexpected malfunctions.

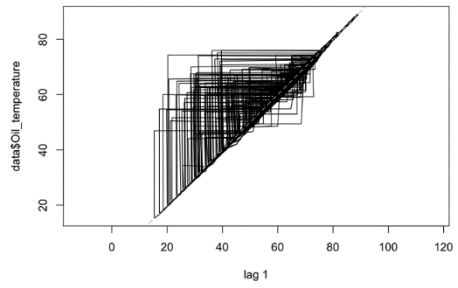

Figure 7 Lag plot

The lag plot shown in Figure 7 is a tool used to visualize and analyze time series data. On this plot, the x-axis displays the lag value, which is the amount by which the data has been shifted in time. The y-axis, on the other hand, represents the data values at each specific time step. Each point on the plot is a comparison between a data value at a certain time and its value at a previous time, dictated by the lag value.

The appearance of a diagonal pattern in this lag plot is significant. It indicates that the data exhibits strong autocorrelation, meaning there is a high degree of correlation between values in the series at one time and their previous values at set lags. This pattern suggests that the future values of the data can often be predicted with some accuracy based on its past values, due to this temporal dependence.

Understanding this autocorrelation is crucial for analyzing patterns and behaviors within time series data. It helps in forecasting and modeling, as the past values provide substantial information about future trends. This insight is especially valuable in fields like financial forecasting, weather prediction, and inventory management, where predicting future trends based on past data is essential.

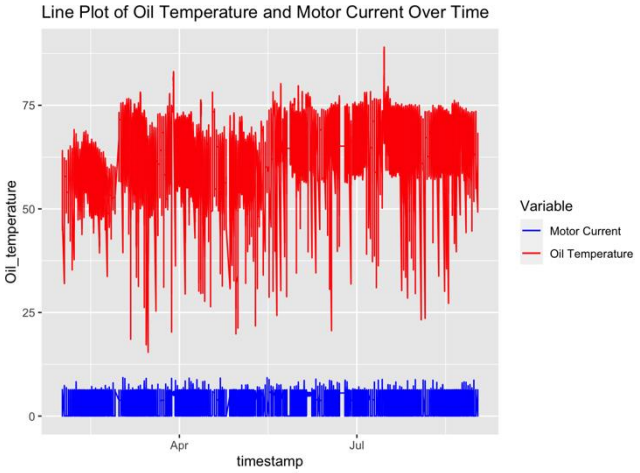

Figure 8 Oil temperature and motor current

In Figure 8, the graph illustrates two different aspects of a machine's operation: oil temperature and motor current, each represented by its own line on the plot. The line representing oil temperature shows considerable variation, fluctuating dramatically over time. This indicates that the temperature of the oil changes frequently and significantly, which could be due to various operational activities or external conditions affecting the machine.

Conversely, the line depicting motor current is notably more stable, with very little variation over the same period. This stability suggests that the motor current remains consistent regardless of other operational changes. Such steadiness in motor current could indicate a well-maintained motor system or one that is not heavily impacted by fluctuations in operational conditions.

Different dynamics within the machine's operation are highlighted by the contrasting behaviors of these two lines: significant fluctuations in oil temperature versus stability in motor current. Even though the oil temperature may be susceptible to several variables and needs to be closely watched to avoid problems like overheating, the motor's current stability indicates system dependability in that area. Maintaining optimal operation and avoiding

potential failures brought on by temperature extremes or other problems require an understanding of these patterns.

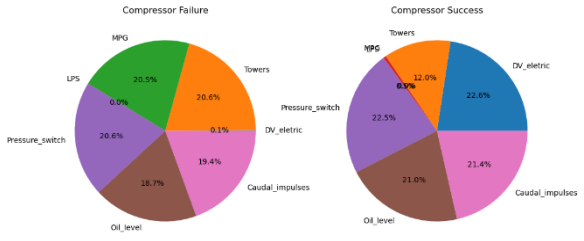

Figure 9 Compressor failure and Compressor success

The analysis of factors leading to compressor failure and success reveals distinct patterns in their contribution levels. For compressor failure, four main factors stand out: Towers, Caudal Impulses, Oil Level, and Pressure Switch, with each contributing approximately 20% to failures. MPG also plays a role, accounting for slightly over 20%, while DV Electric shows a minimal impact at just 0.1%. LPS, according to the data, does not influence compressor failure.

Conversely, the scenario for compressor success involves a different set of contributions from these sensors. DV Electric, Pressure Switch, and Towers are the leading factors, each contributing about 22% to successful operations. Caudal Impulses and Oil Level also play significant roles, contributing 21.4% and 21%, respectively. Interestingly, MPG and LPS do not seem to impact compressor success, as per the data.

This contrasting distribution underscores that while certain factors like Towers, DV Electric, Caudal Impulses, Oil Level, and Pressure Switch are crucial in both scenarios, their effects vary significantly between compressor failure and success. These findings highlight the need for careful monitoring and management of these sensors to optimize compressor performance and prevent failures, adapting strategies based on their different impacts on operational outcomes.

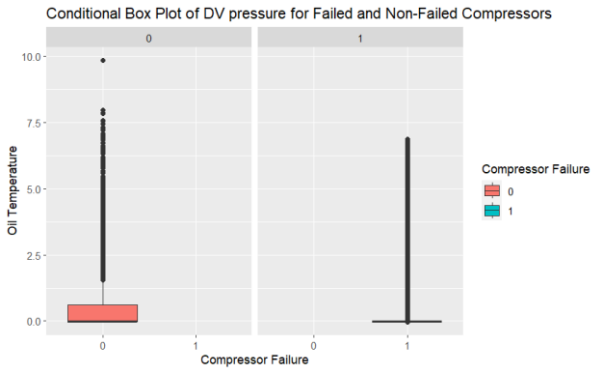

Figure 10 Compressor Failure and Oil Temperature

The graph presented is a conditional box plot that illustrates the distribution of oil temperature (OT) in relation to the status of compressor failures, based on DV pressure readings for failed and non-failed compressors. The x-axis categorizes compressors into two groups based on whether a failure occurred (0 for no failure, 1 for failure), while the y-axis represents the oil temperature in degrees Celsius.

For compressors that did not fail (indicated by 0 on the x-axis), the box plot shows a considerably lower range and variability in oil temperature. The main body of the box, which indicates the interquartile range (IQR), is very narrow and situated at the lower end of the scale, suggesting that most oil temperature readings for these compressors are low and tightly grouped. There are a few outliers indicated by the dots above the box, showing that occasionally the oil temperature does rise significantly, but these are exceptions rather than the norm.

In contrast, the box plot for compressors that failed (indicated by 1 on the x-axis) shows a dramatically different pattern. This plot does not actually display a 'box', implying there is no significant IQR to report; instead, it features a line that extends across a broad range of the y-axis. This suggests that oil temperature readings for failed compressors vary widely, from low to very high, indicating less stability in the system's thermal management.

This visual comparison effectively highlights how oil temperature behaves differently in compressors based on their failure status, with non-failed compressors generally maintaining a stable, low temperature and failed ones exhibiting a wide range of temperatures. This could suggest that monitoring and controlling oil temperature might be crucial in preventing compressor failures.

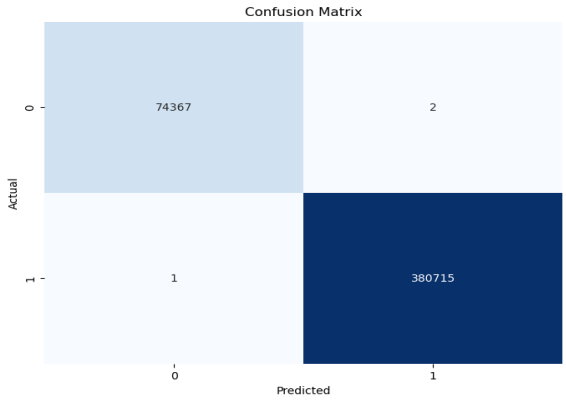

Figure 11 Confusion Matrix

The provided confusion matrix offers detailed insight into the model's performance in predicting compressor failures. The matrix shows the following:

- True Positives (TP): 380,715 – This high number suggests that the model is highly effective at correctly identifying actual compressor failures.
- True Negatives (TN): 74,367 – This indicates that the model also accurately recognizes when compressor failures do not occur.
- False Positives (FP): 2 – The extremely low number of false positives means that the model very rarely mislabels successful operations as failures, which is crucial for avoiding unnecessary maintenance actions.
- False Negatives (FN): 1 – Similarly, the model almost never misses actual compressor failures, as indicated by the very low false negative rate.

These statistics demonstrate that the model excels in accurately detecting both occurrences and non-occurrences of compressor failures. The high true positive rate combined with exceptionally low rates of false positives and false negatives points to a model that is both reliable and precise, minimizing the risk of both over-maintenance and unnoticed failures. Overall, the model's performance in predicting compressor failures is commendable, characterized by high accuracy and minimal misclassification, making it a valuable tool for operational management and preventive maintenance strategies.

```
Accuracy: 0.995704099234209
Cross-validation Scores: [0.99590814 0.99575276 0.99566329 0.99553143 0.99568681]
Mean CV Accuracy: 0.9957084857474724
```

Figure 12 Accuracy obtained

In our analysis, we used a Decision Tree Classifier to model and predict compressor failures. Our dataset was divided into two parts: a training set and a testing set. The primary variable of interest, or target variable, in our study is "Compressor", and it is influenced by readings from eight different sensors.

Among these sensors, three in particular—"MPG", "Tower", and "Caudal_impulses"—proved to be most significant in achieving the highest accuracy in our predictions. This suggests that these sensors are crucial indicators of potential compressor failures.

To enhance the reliability and generalization of our model, we implemented cross-validation using a five-fold split on the training data. This method involves dividing the training set into five smaller sets. The model is trained on four of these sets and validated on the fifth, and this process is repeated five times with each of the subsets used exactly once as the validation data. This technique helps in minimizing overfitting and ensuring that our model performs well not just on our training data but also on new, unseen data.

Employing this rigorous approach allows us to rely on the model's predictive power more confidently regarding compressor performance, providing a robust tool for preemptive maintenance and operational planning.

V. SYSTEM ARCHITECTURE

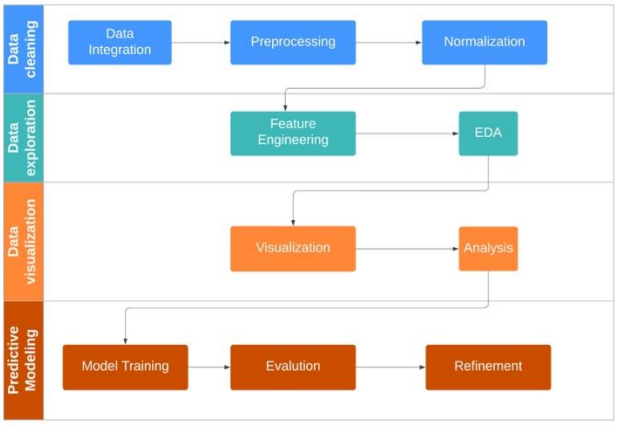

Figure 9 System architecture [12]

The process of developing a robust predictive model for compressor failures involves several critical steps, each designed to ensure the accuracy and reliability of the predictions. This comprehensive approach includes data integration, preprocessing, feature engineering, exploratory analysis, model selection, training, evaluation, and refinement. Here's an expanded explanation of each step:

A. *Data Integration:*

- To build a predictive model for compressor failures, the first step involves integrating sensor data, which is time-stamped, with corresponding failure reports. This creates a cohesive dataset where each instance of sensor data is directly linked to whether a failure occurred. This linkage is crucial as it provides the foundational data structure from which predictive insights are derived.

B. *Data Preprocessing and Normalization:*

- Data preprocessing is a critical step to prepare the raw data for effective analysis. This includes cleaning the data to fix anomalies and filling in missing values, which are common issues in real-world data. Normalization is also performed to scale the sensor readings uniformly across all features. This step is important because it ensures that no single feature dominates the model due to scale differences, which could skew the model's predictions.

C. *Feature Engineering:*

- Feature engineering is the process of using domain knowledge to extract new features from the raw data that might be more predictive of the target variable. This could include decomposing time-series data into trend and cyclic components, using Fourier transformations to capture periodic patterns, or creating flags based on anomaly detection. These engineered features can often uncover complex patterns in the data that simple models might miss.

D. *Exploratory Data Analysis (EDA) & Visualization:*

- EDA is employed to explore and better understand the data's underlying structure as well as to unearth any inherent patterns or anomalies that might influence predictions. Visualization tools such as histograms for distribution analysis, box plots for outlier detection, and heatmaps for correlation analysis are particularly useful. These tools help in visualizing complex relationships and trends in the data, providing insights that guide further analysis.

E. Machine Learning Model Selection

- Choosing the right model is crucial. For predicting compressor failures, decision trees and logistic regression are considered good starting points due to their simplicity and high interpretability. These models provide a clear indication of how input features relate to the target outcome, making them especially useful in settings where understanding the decision-making process is as important as the accuracy of the predictions.

F. Model Selection, Training, and Assessment:

- The dataset is divided into training and testing sets after the model is chosen. With the training set, the model gains the ability to forecast failures by utilizing the features. The model is tested on a testing set to assess its performance after training. An objective way to assess the performance of the model is to use evaluation metrics like F1-score, accuracy, precision, and recall. The purpose of cross-validation is to make sure that the model performs consistently across various subsets of data. Depending on the size of the dataset, this is often done five or ten times.

G. Model Optimization and Refinement:

- It is rare for the initial model to be flawless. Iteratively improving and optimizing the model is done based on the evaluation. This could entail fine-tuning the model's hyperparameters, which are its performance-enhancing settings. The model's capacity to generalize to new data is improved by methods like regularization, which keeps the model from overfitting, and feature selection, which retains only the most significant features.

By following these procedures, the predictive model is painstakingly developed to offer maintenance teams useful insights in addition to accurately forecasting compressor failures. By using a methodical approach, it is ensured that all available data is utilized and that the resulting model is strong, dependable, and ready to work in real-world scenarios.

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 2	Apr 14, 2024
Done	Statistical Analysis and Hypothesis Testing	Week 13
Done	Refinement	Week 14
Done	Project submission and presentation	Week 15

VI. Conclusion

The data emphasizes how crucial it is to continuously invest in the upkeep and modernization of rail networks as these are important factors that promote both increased safety and economic growth. Neglecting these systems' maintenance and technological advancements can have serious consequences, such as accidents and significant financial losses. The manufacturing sector has demonstrated historically how technological advancements can drive industries forward, especially when they integrate machine learning and artificial intelligence (AI). This is especially important for the rail sector, which will gain a lot from implementing these technologies.

The introduction of smart technologies and networked systems in the rail sector can revolutionize how maintenance is conducted. For example, using AI to analyze data collected from sensors across the rail network can help predict and prevent potential faults before they lead to failures. Machine learning algorithms can process vast amounts of operational data to optimize routes and schedules, reducing delays and increasing efficiency.

Furthermore, these technologies facilitate improved decision-making by offering thorough analytics that support maintenance plans, forecast system degradations, and recommend prompt interventions. By reducing the possibility of mishaps brought on by malfunctioning equipment, this proactive maintenance strategy not only guarantees more efficient operations but also improves safety.

In conclusion, it is imperative that cutting-edge technologies like artificial intelligence (AI) and machine learning be incorporated into the rail sector to future-proof the networks. More dependable, effective, and safe rail systems are the result of these modernization efforts, and these systems are necessary to maintain both public trust in rail transportation and economic growth. Being at the forefront of technological innovation will be essential to the industry's continued evolution and preservation of its relevance and operational excellence.

VII. LIMITATIONS

To completely grasp the project's scope and possible difficulties, it is necessary to consider the inherent limitations of the structured and thorough approach for creating a predictive model for compressor failures. In practical

applications, these drawbacks may affect the model's overall efficacy and dependability.

1. **Data Availability and Quality:** Ensuring the quality and completeness of the data is a major challenge in any predictive modeling project. The accuracy, completeness, or bias of the sensor data will directly impact how reliable the model's predictions are. Inconsistent data collection, incorrect readings, and missing values can all distort analysis and produce false conclusions.
2. **Complexity of Feature Engineering**: While feature engineering can significantly enhance model performance by introducing more relevant variables, it also increases the complexity of the model. This can make the model more difficult to interpret and can lead to overfitting if not carefully managed. Moreover, the process of feature engineering requires deep domain knowledge and creativity, which might be limited by the expertise available.
3. **Model Interpretability**: Decision trees and logistic regression are chosen for their interpretability, which is crucial for operational settings where understanding the decision-making process is important. However, these models may sometimes sacrifice accuracy for simplicity, especially in complex systems with nonlinear relationships and interactions among variables. More sophisticated models might capture these complexities better but at the cost of being less interpretable.
4. **Generalization of the Model**: While cross-validation helps in assessing the model's ability to generalize across different data subsets, there is always a risk that the model may not perform well on completely new, unseen data. This is particularly true if the model has been tuned to perform optimally on the specific characteristics of the training data. External factors not represented in the training data can also affect the model's performance.
5. **Adaptability to New Conditions**: As a result of environmental changes, wear and tear, or new operating procedures, compressor systems may experience modifications over time. If a model is not regularly updated with new data, which necessitates ongoing data collection and model retraining, it may not be able to adjust well to these changes.
6. **Resource Intensive**: Developing and maintaining a predictive model requires significant resources, including skilled personnel, computational power, and time. The need for continuous monitoring, updating, and refining the model to adapt to new data or conditions can be resource-intensive and costly.
7. **Dependency on External Factors**: The performance of compressor systems can be influenced by a range of external factors such as environmental conditions, operational demands, and maintenance practices. These factors might not be fully captured by the model, especially if the data does not include variables that adequately represent these influences.
8. **Ethical and Privacy Concerns**: Collecting, storing, and analyzing sensor data from compressor systems can raise privacy and security concerns, particularly if the data includes information that could be considered sensitive. Ensuring the security of the data and complying with applicable privacy laws and regulations is crucial.

VIII. Future work

The rail sector is expected to see substantial improvements if new technologies like artificial intelligence (AI) and machine learning are developed and implemented. By enabling more accurate and predictive maintenance, these developments will improve rail safety by enabling railway companies to address issues before they cause disruptions, guaranteeing smoother operations and minimizing downtime. Large-scale data from sensors and other sources can be analyzed by AI to forecast equipment failures or hazardous situations, resulting in more proactive maintenance plans.

Moreover, as the demand for more efficient and extensive rail services grows, countries are increasingly investing in the expansion and modernization of their rail networks. This expansion is not just about adding more tracks or trains; it involves integrating advanced technology to improve connectivity and alleviate congestion. Modernized rail systems can handle more trains at higher speeds, significantly enhancing passenger and freight movement across regions and reducing the strain on existing infrastructure.

The development of sustainable rail systems is also receiving more attention. This involves initiatives such as electrifying railroads to switch from diesel to electric engines, which are quieter and cleaner. As part of the worldwide effort to fight climate change, reducing emissions from trains is increasingly being prioritized. Moreover, enhancing the energy efficiency of trains and rail operations contributes to lowering the transportation industry's total carbon footprint.

These improvements aim not only to enhance the functionality and capacity of rail systems but also to make rail travel a more attractive option compared to road transport, which is often less energy efficient and more polluting. By focusing on sustainability, safety, and efficiency, the rail industry is positioning itself as a critical component of future transportation networks that prioritize both performance and environmental responsibility.

IX. 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.' 2qSensors 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, N., 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#