Predictive Maintenance of Urban Metro Transportation Systems

Dawson Burgess
Computer Science Department
University of Idaho, Moscow, ID
Moscow, ID, United States
burg1648@vandals.uidaho.edu dawsonhburgess@gmail.com

I. Introduction

Urban metro systems are an integral part of public transportation in numerous cities (east cost of America, Europe, etc.). Ensuring their reliability is fundamental for efficient and continuous operation. Unexpected equipment failures can lead to significant service delays, safety risks, and increased operational costs [1]. As cities continue to grow and expand, the demand for reliable public transportation increases, and so does the amount and complexity of the data generated by sensors and monitoring equipment. The demand for advanced maintenance strategies rises in conjunction with all of this growth.

As outlined by Susto et al. [2], maintenance management strategies in industrial and transportation sectors can be categorized into three main approaches: R2F, PvM, and PdM. Each of these have varying levels of efficiency and complexity:

1) Run-to-failure (R2F): This is where maintenance interventions are performed after the occurrence of failures. This is simplest and most frequently used method for handling maintenance. The cost of interventions and associated downtime after failure are more substantial than those associated with planned corrective actions taken in advance.

2) Preventitive Maintenance (PvM): This is when maintenance actions are carried out according to a planned schedule based on time or process iterations. Also referred to as scheduled maintenance - this approach prevents failures. This comes at the cost of often unnecessary and costly corrective actions.

3) Predictive Maintenance (PdM): This is when maintenance is performed based on an estimate of the health status of a piece of equipment. Other authors and research also refer to this class of maintenance as condition-based maintenance. PdM systems allow advance detection of pending failures. This method is built around prediction tools based on historical data, statistical inference methods, and engineering approaches.

Advancements in machine and deep learning technologies have further enhanced the capabilities of PdM systems. Studies by Nair et al. [1] and Davari et al. [4] have demonstrated the effectiveness of employing models such as Long Short-Term Memory (LSTM) networks and Random Forest classifiers in predicting equipment failures. These models can process large amounts of data and identify complex patterns and correlations that may indicate future failures. Research by Veloso et al. [3] has introduced the MetroPT-3 dataset – an established benchmark collection of sensor data from urban metro compressors. This dataset serves as a foundation for developing and evaluating predictive maintenance models.

Challenges still exist in implementing effective PdM systems in metro transportation. A significant hurdle is the class imbalance in failure prediction datasets (a problem faced in this project), where failure events are somewhat rare compared to normal operational data. This imbalance can hurt the performance of ML models, leading to biased predictions and reduced accuracy in failure detection.

This study aims to address these challenges by leveraging the MetroPT-3 dataset to develop and evaluate multiple machine learning models for predicting equipment failures in metro systems. It will explore the effectiveness of using Random Forest classifiers, Support Vector Machines (SVMs), Neural Networks, and LSTM networks on this dataset. This study will employ techniques such as data normalization, noise reduction, resampling, and feature importance analysis to improve the predictive accuracy and reliability of these models. This research will also aim to identify the critical indicators for effective maintenance planning. The findings are expected to inform maintenance strategies that minimize operational disruptions, reduce costs, and ensure the safety and reliability of metro transportation services.

II. Materials and Methods

This section details the dataset utilized, the data pre-processing steps, the visualization techniques used, and the machine learning models implemented. Each subsection outlines the methodologies applied.

A. Description of Dataset

The foundation of this study is the MetroPT-3 dataset, a comprehensive multivariate time series dataset collected from urban metro compressor units in Porto, Portugal, in 2022 [3]. The dataset encompasses over 15 million instances, recorded at a 1Hz sampling rate (every 10 seconds or so), capturing both analog and digital sensor signals for monitoring the health and performance of metro equipment. The dataset contains 15 features, categorized as follows:

1) 7 Analog Sensor Signals:

- *TP2 (bar):* The measure of the pressure on the compressor.
- *TP3 (bar):* The measure of the pressure generated at the pneumatic panel.
- *H1 (bar):* The measure of the pressure generated due to pressure drop when the discharge of the cyclonic separator filter occurs.
- *DV_Pressure (bar):* Measure of the pressure drop generated when the towers discharge air dryers; a zero reading indicates the compressor is operating under load.
- *Reservoirs (bar):* Measure of the downstream pressure of the reservoirs, which should be close to the pneumatic panel pressure (TP3).
- *Oil_Temperature (°C):* Measure of the oil temperature on the compressor.
- *Motor_Current (Amps):* Measure of the current of one phase of the three-phase motor; it presents values close to 0A when it turns off, 4A when working offloaded, 7A when working under load, and 9A when it starts working.

2) 8 Digital Sensor Signals:

- *COMP*: Electrical signal of the air intake valve on the compressor; it is active when there is no air intake, indicating that the compressor is either turned off or operating in an offloaded state.
- *DV_Electric*: Electrical signal that controls the compressor outlet valve; it's active when the compressor is functioning under load and inactive when the compressor is either off or operating in an offloaded state.
- *TOWERS*: Electrical signal that defines the tower responsible for drying the air and the tower responsible for draining the humidity removed from the air; when not active, it indicates that tower one is functioning; when active, it indicates that tower two is in operation.
- *MPG*: electrical signal responsible for starting the compressor under load by activating the intake valve when the pressure in the air production unit (APU) falls below 8.2 bar; it activates the COMP sensor, which assumes the same behaviour as the MPG sensor.
- *LPS:* Electrical signal that detects and activates when the pressure drops below 7 bars.
- *Pressure Switch:* Electrical signal that detects the discharge in the air-drying towers.
- *Oil Level:* Electrical signal that detects the oil level on the compressor; it is active when the oil is below the expected values.
- *Caudal_impulses:* Electrical signal that counts the pulse outputs generated by the absolute amount of air flowing from the APU to the reservoirs.

Nr.	Start Time	End Time	Failure	Severity	Report
#1	4/18/2020 0:00	4/18/2020 23:59	**Air leak**	High stress	
#1	5/29/2020 23:30	5/30/2020 6:00	**Air Leak**	High stress	Maintenance on 30Apr at 12:00
#3	6/5/2020 10:00	6/7/2020 14:30	**Air Leak**	High stress	Maintenance on 8Jun at 16:00
#4	7/15/2020 14:30	7/15/2020 19:00	**Air Leak**	High stress	Maintenance on 16Jul at 00:00

Table 1: Failure Descriptions

3) Failure Information

B. Data Preparation

Data preparation is needed for developing accurate and reliable machine learning models. The data preparation process involved these key steps: data cleaning, handling class imbalance, feature selection, normalization, and noise reduction.

1) Data Cleaning

Initial data exploration revealed inconsistencies in timestamp intervals and potential noise in sensor readings. To address these issues:

a) Timestamp Conversion: The 'timestamp' column was converted to datetime objects to ensure chronological integrity and facilitate time-based operations.

b) Failure Labeling: Based on known failure periods provided, a binary 'failure_anomaly' column was created to indicate failure events.

c) Time Fluctuation Handling (noise): To maintain data consistency, records with time differences exceeding 10 seconds between consecutive timestamps were identified and removed, as these indicated potential sensor lag or recording errors. A total of 51,147 data points were removed for exceeding a variance of 10 seconds between readings. Some variance times between readings were 15,743 seconds and higher.

2) Handling Class Imbalance

The dataset exhibited significant class imbalance. The number of failure instances (*failure_anomaly = 1*) heavily outnumbered by non-failure instances (*failure_anomaly = 0*). This imbalance lead the models to favor the majority class and reduced their ability to detect failures accurately. To address this, the program utilized downsampling/resampling the majority class. For all the models, the majority class (*failure_anomaly = 0*) was downsampled in the training data to achieve a 1:1 ratio relative to the minority class. For example, given 1,148,755 non-failure instances and 17,914 failure instances, we randomly selected 17,914 non-failure instances. This ensured the training dataset was balanced, preventing models from becoming biased toward the majority class. Before resampling was introduced, this was the most

challenging part of the project – dealing with the data imbalance.

Class	Before Resampling	After Resampling
Training (0)	861,566	17,914
Training (1)	17,914	17,914

Table 2: Class Distribution Before and After Resampling

3) Feature Selection

Based on preliminary analysis (see figure 1 and 6), five key features were selected for model training:

- Motor_Current
- Oil_Temperature
- DV_Electric
- TP2
- DV_Pressure

4) Normalization Techniques

Normalization is an important preprocessing step for algorithms sensitive to feature scaling. In this studies case it was for the SVM and Neural Network. The StandardScaler was used to standardize the dataset. This method transforms the data to have a mean of zero and a standard deviation of one. For instance, features such as *Motor_current* and *DV_pressure* vary across significantly different ranges, and normalization ensures these features contribute equally to the model's performance.

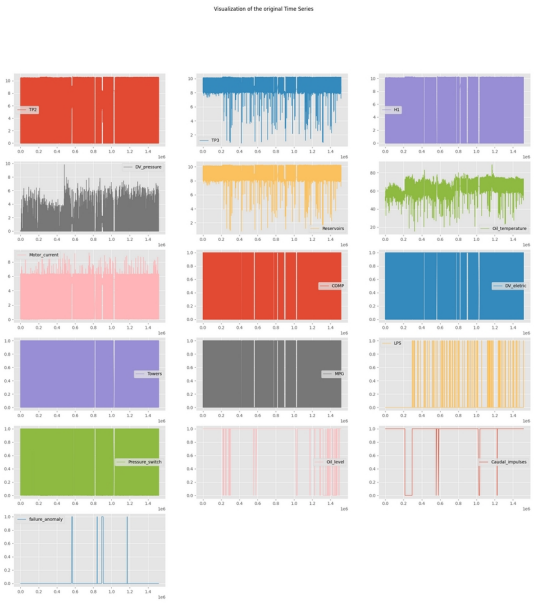

Figure 1: Time-Series Plots of All Features

Without normalization, features with larger magnitudes would dominate smaller-scaled features and skew the model's predictions. After normalization, all features are brought to the same scale. This resulted in a noticeable improvement in both training speed and accuracy for the SVM and Neural Network.

C. Visualization Techniques

Visualization played a pivotal role in understanding data patterns and correlations, guiding feature selection and model evaluation.

1) Time-Series Plots

Time-series plots for every feature were initially generated and examined (see figure 1). After closer inspection of the correlation matrix (see figure 6), time-series plots for each key sensor were generated. These key feature plots overlay failure events to visually inspect potential patterns leading to equipment failures (see figure 2 through 5).

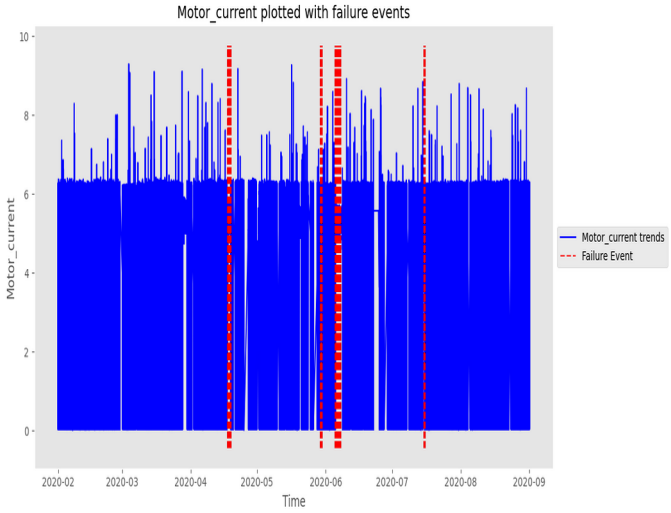

Figure 2: Time-Series Motor_Current With Failure indicators

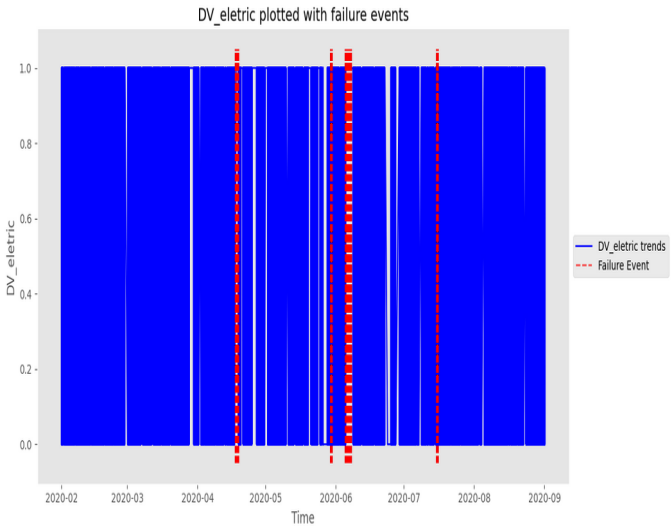

Figure 3: Time-Series DV_Electric With Failure indicators

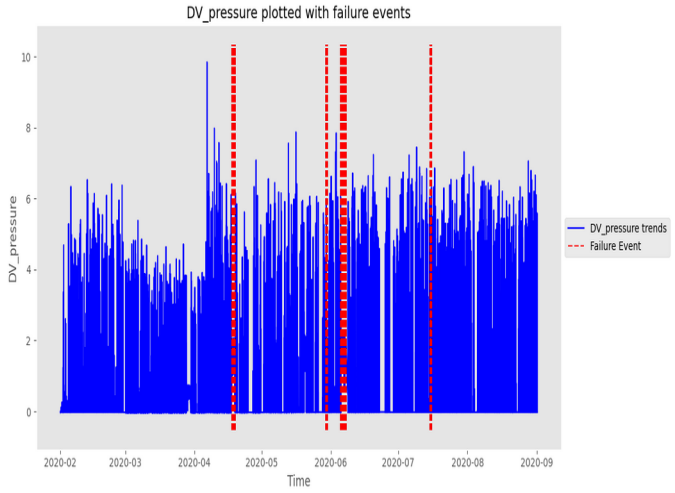

Figure 4: Time-Series DV_Pressure With Failure indicators

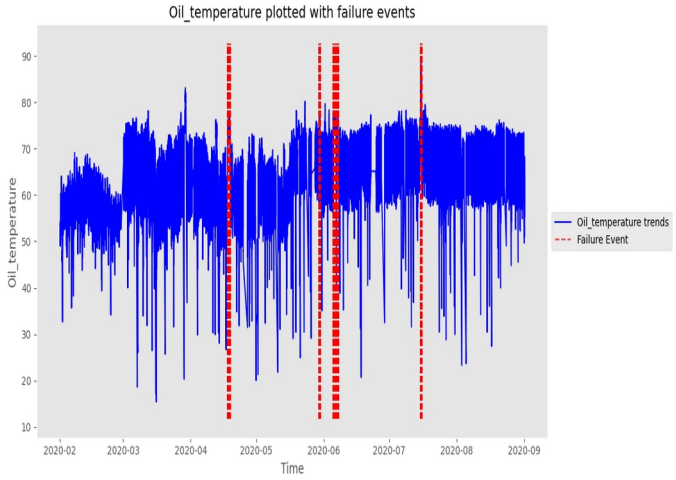

Figure 5: Time-Series Oil_Temperature With Failure indicators

2) Correlation Matrix

A Spearman correlation matrix was created to identify relationships between features, aiding in feature importance analysis.

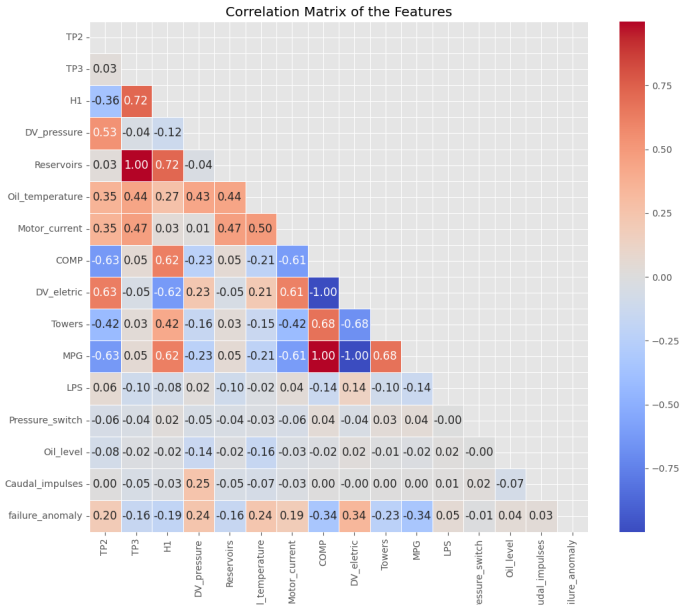

Figure 6: Correlation Matrix of The Features

D. Model Design and Experimental Setup

To address the predictive maintenance objectives, three machine learning models were developed and evaluated: Random Forest Classifier, SVM, and Neural Network (the LSTM model was scrapped from the project – expanded upon in its section). Each model was selected based on its ability for handling large-scale, imbalanced, and multivariate time-series data.

1) Random Forest Classifier

The Random Forest classifier was chosen for its ability to handle high-dimensional data, and inherent feature importance estimation. This makes it perfect for our use case and for failure prediction tasks. Default parameters were used initially, with further tuning to optimize performance. Estimators was set t0 500, a max depth of 20 was utilized, and a balanced class weight. The model was trained on the balanced dataset, utilizing the selected key features. Post-training, feature

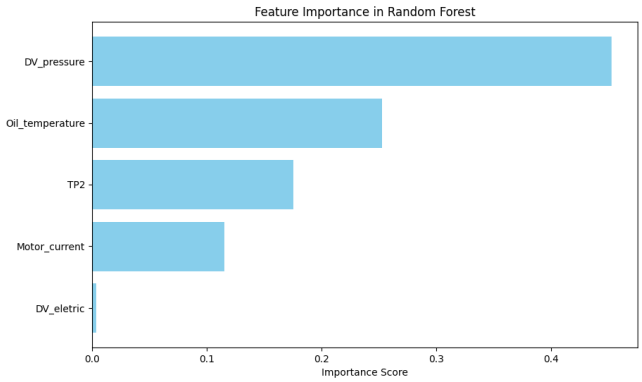

Figure 7: Feature Importance in Random Forest

importance scores were extracted to identify the most influential sensors.

2) Support Vector Machine (SVM)

The SVM was employed for its effectiveness in high-dimensional spaces and its ability to create clear margins of separation between classes, making it a strong candidate for anomaly detection. A radial basis function (RBF) kernel was selected after experimenting with linear and polynomial

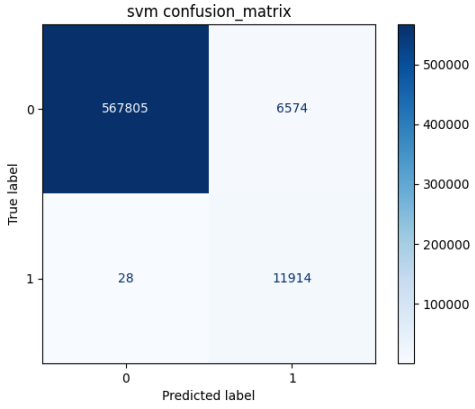

Figure 8: Confusion Matrix for SVM

kernels. The RBF kernel offered better decision boundaries for capturing the non-linear relationships between *Oil_temperature* and *TP2*. The model was trained on the normalized and balanced dataset.

3) Neural Network

A Neural Network was implemented to capture nonlinear relationships within the data, using its flexibility and capacity to model complex patterns. The model had two hidden layers with 64 and 32 neurons using ReLU activation functions. The output layer had a single neuron with a sigmoid activation function for binary classification. It had the following training parameters:

- Adam optimizer with a learning rate of 0.001. Higher learning rates (0.01) caused the loss to oscillate, and lower rates (0.0001) resulted in slower convergence without significant improvements.
- Binary cross-entropy loss function.
- 10 epochs with a batch size of 64.

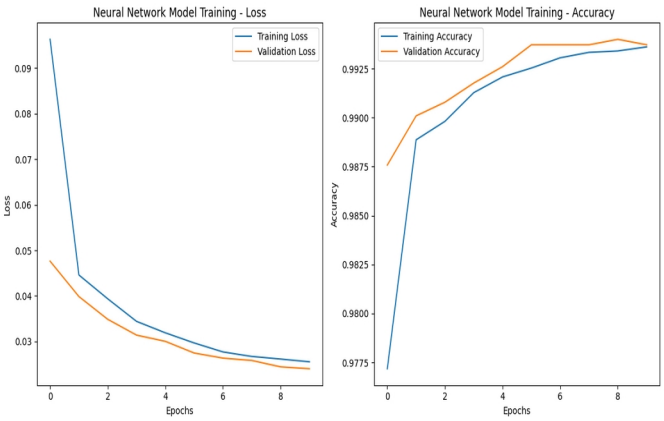

Figure 9: Training History of Neural Network

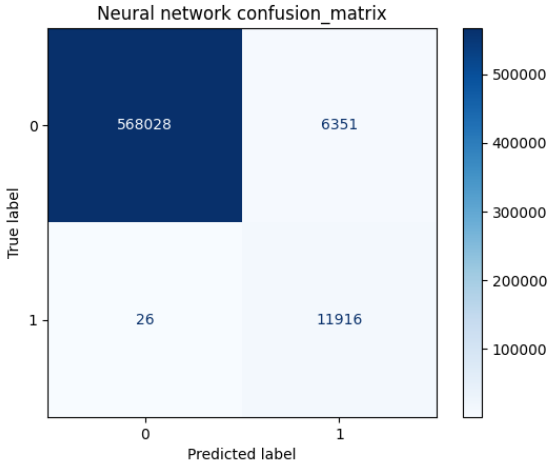

Figure 10: Confusion Matrix for Neural Network

The model's performance was assessed using accuracy and loss metrics on the test set.

4) Exploratory Sequential Modeling with LSTM

An exploratory attempt was made to use a LSTM network to model the temporal dependencies in the dataset. LSTMs are well-suited for time-series data due to their ability to capture sequential patterns over long time horizons. For this study extensive tuning was perfromed and lots of resampling of data, but the LSTM model failed to deliver satisfactory results. Below the challenges faced and the insights derived are outlined.

a) Data Limitations for Temporal Modeling: The features selected lacked strong temporal correlations. Even after creating up to 50-step sequences, the model struggled to identify meaningful patterns for minority class (failure) instances. Most failure events in the dataset were abrupt rather than gradual, and this limits the LSTM's ability to leverage sequential information.

b) Imbalance Amplification in Sequences: Class imbalance posed a significant challenge when generating sequences. Resampling techniques like SMOTE and others were applied to balance the data, but the resulting sequences often introduced artificial correlations that did not generalize well to unseen data.

c) Overfitting and Generalization Issues: The LSTM consistently overfitted to the training data despite regularization techniques like dropout and early stopping. The model achievied over 99% accuracy on the training set, and the model's performance on the test set was poor for the minority class. In particular, recall for the failure class (*failure_anomaly = 1*) remained at 0%.

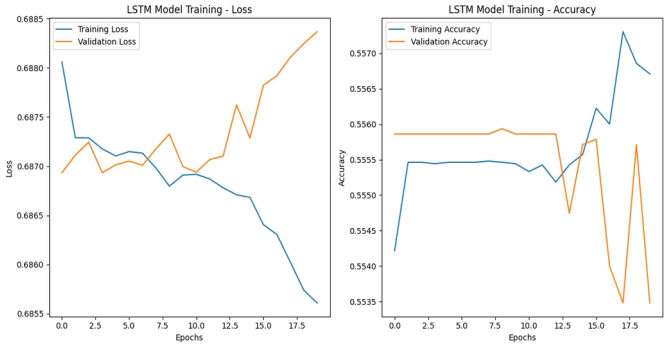

Figure 11: Training History of LSTM Model

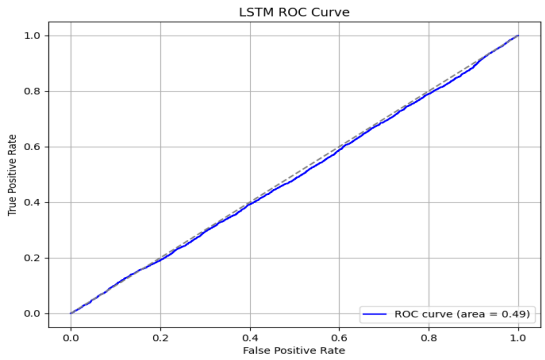

Figure 12: ROC Curve for LSTM Model

III. Results

This section presents the outcomes of the experiments conducted using various machine learning models for predictive maintenance in urban metro systems. Most of the figures in section 2 cover the visual results of the models, so this section will be focused on summerizing results and placing them into tables.

A. Overview of Model Performance

Three primary models were evaluated—random dorest, SVM, and a neural network. This models were chosen to address the challenges of class imbalance, noisy data, and the non-linear relationships in the dataset. The random forest model emerged as the most successful across all metrics, and the neural network demonstrated strong capability in capturing non-linear interactions. The LSTM model was ultimately scrapped from the study despite its theoretical advantages for sequential data. It failed to deliver meaningful results due to insufficient temporal context and dataset characteristics.

B. Random Forest Classifier

The Random Forest model achieved the highest overall accuracy (99.5%) and demonstrated the best performance in classifying the minority class (*failure_anomaly = 1*), with a recall of 100%. The model effectively leveraged the resampled training data to ensure failure predictions were not overwhelmed by the majority class. The feature importance analysis revealed that *DV_pressure* (44.3%) and *Oil_temperature* (27.2%) were the most predictive variables.

Class	Precision	Recall	F1	Support
0	1	0.99	1	574379
1	0.80	1	0.89	11942
Macro avg	0.90	1	0.94	586321

Table 3: Random Forest Classification Report

Figure 7 presents a bar chart demonstrating the full set of feature importance scores.

C. SVM

The SVM model achieved an accuracy of 98.90% , a recall of 100% for the failure class, and a slightly lower precision (64%). This can be expected due to its sensitivity to imbalanced data. The kernel-based approach of the SVM permitted it to capture linear and a few non-linear patterns in the data. Tt struggled with the high dimensionality and noise present in features like *DV_electric*.

Class	Precision	Recall	F1	Support
0	1	0.99	0.98	574379
1	0.64	1	0.78	11942
Macro avg	0.82	0.99	0.89	586321

Table 4: SVM Classification Report

The confusion matrix for the SVM model is shown in Figure 8, showing the distribution of correct and incorrect predictions.

D. Neural Network

Overall, the Neural Network performed well. It achievied an accuracy of 98.8% and an ROC AUC score of 0.993. Its ability to capture non-linear relationships was evident in its high precision (64%) and recall (100%) for the failure class. The model required extensive regularization and hyperparameter tuning to prevent overfitting given the class imbalance.

Class	Precision	Recall	F1	Support
0	1	0.99	0.99	574379
1	0.65	1	0.79	11942
Macro Avg	0.99	0.99	0.99	586321

Table 5: Neural Network Classification Report

The training history from figure 9 shows a steady decrease in loss and increase in accuracy over the epochs. This showcases effective learning without significant overfitting. The confusion matrix in figure 10 illustrates the model also has accurate classification with minimal errors.

E. LSTM

The exploratory LSTM model failed to achieve meaningful results, and maintained a recall for the failure class consistently at 0%. This failure can be attributed to two primary factors: the lack of strong temporal patterns in the dataset, and the architectural challenges of training LSTMs on imbalanced data. Abrupt failure events offered insufficient temporal context for the LSTM to leverage.

Class	Precision	Recall	F1	Support
0	0.56	1	0.71	574379
1	0	0	0	11942
Overall	0.56	0.5	0.36	586321

Table 6: LSTM Classification Report

The ROC curve in Figure 12 shows an AUC score of 0.50 – this would be equivalent to random guessing.

Model	Accuracy (0)	Precision (0)	Recall (0)	F1(0)	Precision (1)	Recall (1)	F1(1)
Random Forest	99.50%	1	0.99	1	0.90	1	0.89
SVM	98.90%	1	0.99	0.98	0.64	1	0.78
Neural Network	98.70%	1	0.99	0.99	0.65	0.99	0.99

Table 7: Model Performance Comparison

IV. Conclusion

The expansion of urban metro systems highlight the need for effective maintenance strategies to ensure operational safety, cost-effectiveness, and efficiency. This study focused on developing and evaluating machine learning models for predictive maintenance using the MetroPT-3 dataset. The primary objective was to leverage sensor data to accurately predict equipment failures to in turn allow for proactive maintenance interventions.

A. Key Takeaways

1) Model Performance

Random Forest Classifier and Neural Network models demonstrated near perfect performance, achieving accuracies of 98.91% and 98.68%. These models displayed high precision,

recall, and F1-scores for both failure and non-failure classes (see table 3 and table 5). These models both can reliabily predict equipment failures. The SVM also performed remarkably well with an accuracy of 97.80%. This shos of its ability to handle highly-dimensional datasets effectively.

Resampling techniques significantly improved minority class recall across all models. The Neural Network's recall for failure events improved from 40% to 100% after resampling, showcasing the need to balance the training dataset.

2) Feature Importance

The Random Forest model's feature importance analysis revealed that DV_Pressure and Oil_Temperature were the most significant predictors of equipment failures. This was followed by TP2 and Motor_Current, demonstrating how their roles are still important for failure prediction. DV_Electric showed near zero importance, suggesting its limited influence on failure events within the dataset. These insights could guide future feature selection and sensor deployment strategies to gather more effective data.

B. Implications

The superior performance of the Random Forest and Neural Network models showcases the potential deep learning techniques in predictive maintenance applications. By accurately predicting equipment failures, these models enable metro operators to implement timely maintenance actions. The identification of DV_Pressure and Oil_Temperature as key indicators provides actionable insights for monitoring and maintaining metro components.

C. LSTM Reflection

The challenges encountered with the LSTM model highlighted critical aspects of the dataset and problem domain:

1) Focus on Feature Importance: The failure of the LSTM showed that the dataset's key features were more indicative of immediate conditions (sensor spikes and abrupt changes) rather than temporal patterns. This helped reinforce the decision to focus on models like random forests. This shift in focus could better capture these non-temporal relationships.

2) Impact of Abrupt Failures: The lack of strong temporal trends in failure events helped lead to the conclusion that failures are event-driven versus process-driven. This insight was shown by the superior performance of models relying on static feature representations.

*3) Future Improvements***:** Just because the LSTM was scrapped for this study, doesn't mean it is out of the frame for future work. One of the key takeaways from this study is that LSTM exploration is suggested for areas of future research. Incorporating additional temporal features such as rolling averages or lagged differences could lead to improved modeling efforts.

V. Future Work

This study has demonstrated advancements in predictive maintenance for metro systems. From all of the studies I have looked at, not many took time to look at the noise in data (variation of >10 second data readings). There still remain several avenues for further exploration and refinement.

A. Experimenting with Other ML Models

There are several other machine learning models/techniques that could be implemented that were not covered by this paper Autoencoder architectures for anomaly detection could provide a complementary approach to identifying failure events based on reconstruction errors. Implementing Isolation Forest algorithms may offer efficient and effective means of detecting anomalies in large-scale datasets. This was another option the early stages of this project considered. Ensemble techniques such as stacking or boosting could enhance predictive accuracy and robustness. These ideas have been covered extensively in other research papers, hence why they were not in this one.

B. Integration of Additional Data Sources

Incorporating new sensors such as humidity, external temperature, and operational load sensors could provide new insights on the factors influencing equipment failures. If not, it would at least provide more context to boost other data's value. On the same note as additional sensors, integrating historical maintenance records and logs could enrich the dataset and provide more insight to types of maintenance/breakdowns.

C. Scalability and Generalizability

The two things that will always remain true are – this type of proect will continue to scale and become more general. As data amounts continue to grow, and more datasets like the Metro-3T become available, it is clear that generalization is key. Testing the developed models across different metro systems and equipment types can assess their generalizability to new operational contexts. Handling scalability of models to deal with increasing volumes of sensor data also will become an important issue to solve.

References

[1] Vishak Nair, Premalatha M, Srinivasa Perumal R, and Braveen M, "Enhancing Metro Rail Efficiency: A Predictive Maintenance Approach Leveraging Machine Learning and Deep Learning Technologies," May 2024. [Online]. Available: https://doi.org/10.21203/rs.3.rs-4319916/v1

[2] Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine Learning for Predictive Maintenance: A Multiple Classifiers Approach. IEEE Transactions on Industrial Informatics, 11(3), pp. 812-820, 2015. Available: 10.1109/TII.2014.2349359

[3] Davari, Narjes & Veloso, Bruno & Ribeiro, Rita & Pereira, Pedro & Gama, João. (2021). Predictive maintenance based on anomaly detection using deep learning for air production unit in the railway industry. 1-10. Available: 10.1109/DSAA53316.2021.9564181.

[4] Veloso, B., Ribeiro, R.P., Gama, J. *et al.* The MetroPT dataset for predictive maintenance. *Sci Data* **9**, 764 (2022). Available: https://doi.org/10.1038/s41597-022-01877-3

[5] Davari, N.; Veloso, B.; Costa, G.d.A.; Pereira, P.M.; Ribeiro, R.P.; Gama, J. A Survey on Data-Driven Predictive Maintenance for the Railway Industry. Sensors 2021, 21, 5739. Available: https://doi.org/10.3390/s21175739

[6] Najjar, Ayat & Ashqar, Huthaifa & Hasasneh, Ahmad. (2023). Predictive Maintenance of Urban Metro Vehicles: Classification of Air Production Unit Failures Using Machine Learning. Available: https://www.researchgate.net/publication/371035904

[7] Ran, Yongyi & Zhou, Xin & Lin, Pengfeng & Wen, Yonggang & Deng, Ruilong. (2019). A Survey of Predictive Maintenance: Systems, Purposes and Approaches. 10.48550/arXiv.1912.07383. Available: