Proceedings of the Sixth International Conference on
Railway Technology: Research, Development and Maintenance
Edited by: J. Pombo
Civil-Comp Conferences, Volume 7, Paper 8.11
Civil-Comp Press, Edinburgh, United Kingdom, 2024
ISSN: 2753-3239, doi: 10.4203/ccc.7.8.11

Predictive Maintenance Optimisation for CCTV Systems in Electric Multiple Unit Trains Using Machine Learning Techniques

**M. M. Rahman[1], B. Alkali[2], A. K. Jain[2],
J. M. Parrilla Gutierrez[2], C. Mcneil[3] and J. Nelson[3]**

**[1]Research in Computing, Department of Mechanical Engineering, Glasgow Caledonian University, United Kingdom
[2]Department of Mechanical Engineering, Glasgow Caledonian University Glasgow, United Kingdom
[3]Siemens Mobility Limited, Glasgow, United Kingdom**

Abstract

This paper presents the study conducted to investigate the maintenance strategy to improve the reliability of Closed-Circuit Television (CCTV) systems in railway rolling stock Electric Multiple Unit trains. The project attempts to optimise maintenance procedures by assessing 1214 sample datasets collected from sensors and control units during fleet data processing to identify and forecast CCTV faults. The analysis indicates that the CCTV system is the worst-performing system leading to delay and cancellation of service operations. The study attempts to address the pattern of CCTV failures using predictive modelling tools, and machine learning techniques such as Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, and Decision Tree Regressor used for modelling and prediction. The results exhibit satisfactory predictive accuracy of the incident reported days, starting date, and issue date for each incident, the results show important performance indicators such as Mean Squared Error, R-squared and Mean Absolute Error, which indicate promising outcomes. The results emphasise the capability of predictive modelling to improve the dependability of CCTV systems in railway rolling equipment, leading to enhanced operational efficiency and passenger safety.

Keywords: CCTV system, rolling stock, machine learning, predictive model, reliability, electric multiple unit.

1 Introduction

Siemens UK has its first Desiro trains introduced in Scotland. These EMU units considered in this paper were constructed between 2009 and 2010 at the Siemens facility in Krefeld, Germany. Later, they underwent additional testing and were put into service at Wildenrath, Germany. Additionally, in-service failure raises maintenance costs, which has an impact on train fleets' total availability and reliability. This study aims to optimise the maintenance of the CCTV system by using condition-based maintenance, considering the number of years they have been in operation. The report does not disclose the design shortcomings of the CCTV system components due to confidentiality.

2 Fleet Data Performance

The highest level of technical reported incidents leading to a decrease in performance of the fleet shows 33 incidents of CCTV malfunctions were reported from 2020 to March 2024. The technical data for the CCTV is classified into Three types namely Primary, Secondary, and Tertiary data sets. The graph in Figure 1 displays the breakdown of the top 10 systems technological incidents and the CCTV incidents have the highest number of incidents recorded.

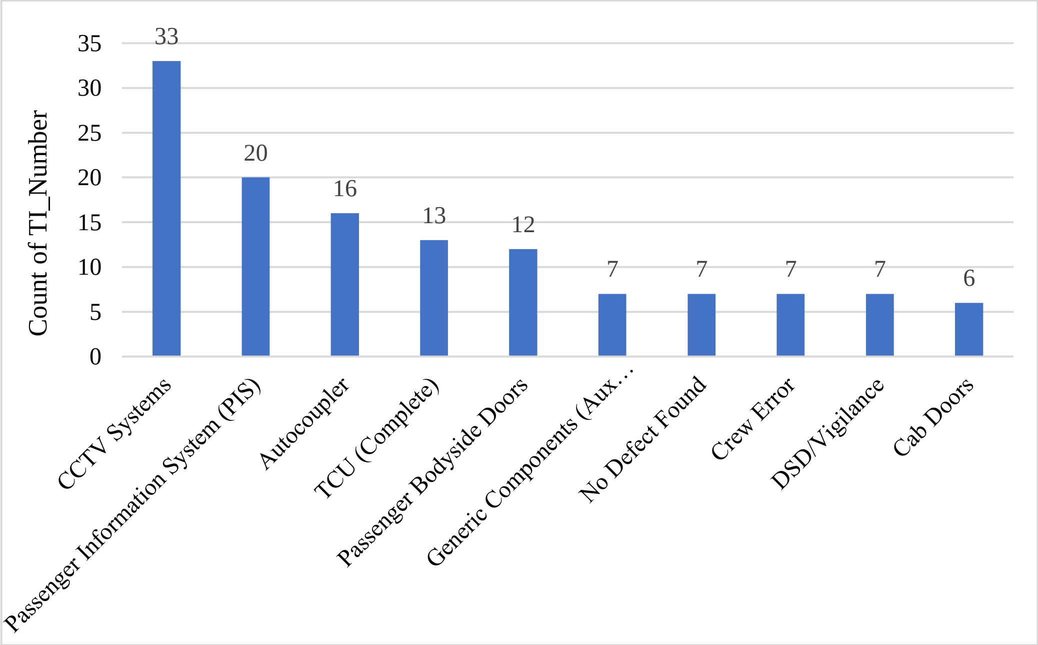

Figure 1: Top 10 System Technical Incident

The CCTV system recorded 51 camera faults, followed by 33 monitor faults as shown in Figure 2

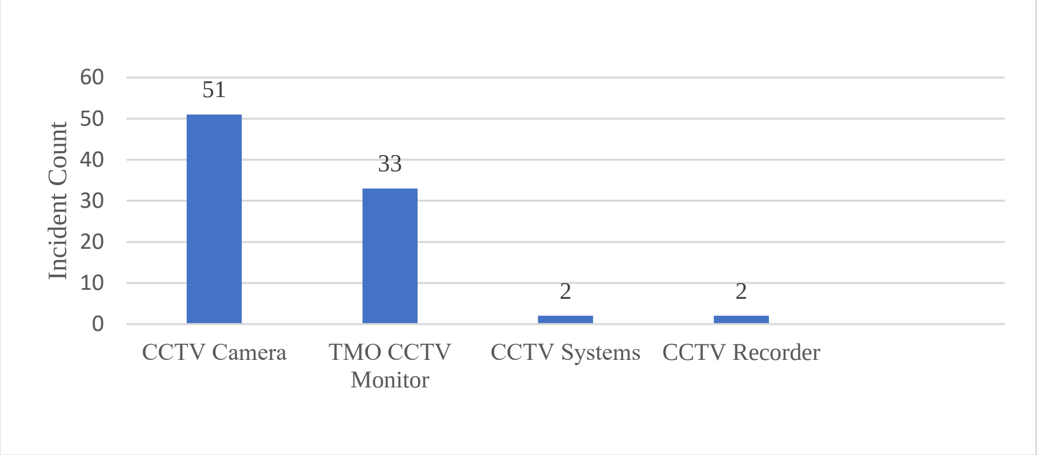

Figure 2: Secondary failure of CCTV

Figure 3 shows fourteen types of failure and Jetview is the highest failure for tertiary failure.

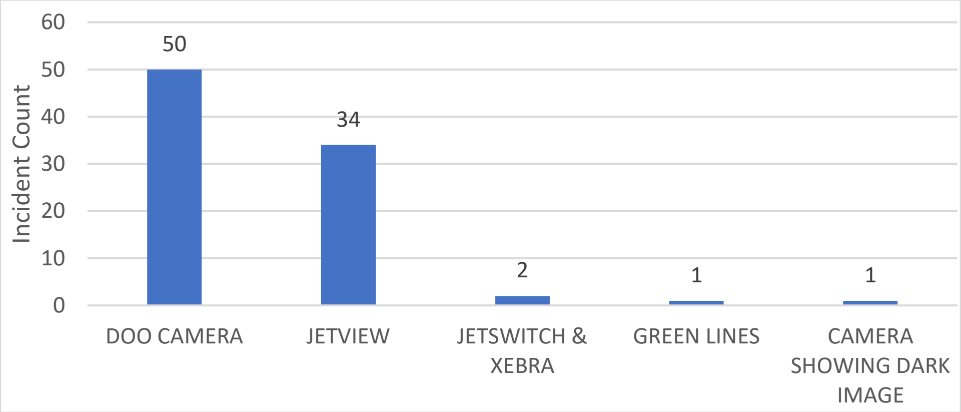

Figure 3: CCTV Tertiary failure incident

Tables 1& 2 below show the grouped Mean Time Between Failure (MTBF) and MTTR (Downtime days) of all CCTV system failures and the percentage distribution

MTBF_Range(Days)	Fault Count	Percentage
0-20	33	64.70
21-40	8	15.68
41-60	6	11.76
61-80	2	3.92
81-100	1	1.96
101-120	1	1.96

Table 1: Mean Time Between Failure

Table 1 shows approximately 65% of MTBF recorded incidents occur between (0-20) days with 2 incidents taking (81-120) days. For example, one CCTV incident (defective on the DMOSA) takes 114 days.

MTTR_Range (Days)	Fault Count	Percentage
0-10	45	88.23
11-20	5	9.80
21-30	0	0.0
31-40	0	0.0
41-50	1	1.96

Table 2: Mean Time To Repair

Table 2 on the other hand shows 88% of the incident range between (0-10) days with one incident taking between (41-50 days) which is the highest number of days to solve a reported CCTV camera incident. For example, misted pods show a long duration. However, in practice, most CCTV incidents are resolved within 24 hours and the data indicate the patterns of incidence which is used as a measure of the overall performance of the EMU fleet.

3 Methods

The use of machine learning techniques has generated a lot of interest in the railway industry. For example, the paper in [1] developed a predictive model for rolling stock vehicle service life prediction using machine learning approaches, namely gradient boosting regression and random forest regression. The Random Forest Regressor model is utilised in the data analysis and prediction. In the energy sector, Ahmad [2] used the random forest model and support vector regressor to predict the failure of solar thermal energy systems and propose a machine-learning approach for effective day-ahead forecasting of solar power generation using retrospective metering data and open-source weather information provided by meteorological services using a random forest algorithm. Predictive maintenance (PDM) [3] utilising machine learning is centred around three fundamental elements: algorithm selection, feature engineering and data pre-processing. Random Forest and Decision Tree algorithms perform with good accuracy in the predictive model. The random forest classifier is found to be more appropriate than other machine learning algorithms. The regression creates several decision trees during training and outputs show the average prediction made by each tree.

3.1 Data Acquisition and Preprocessing

This section describes the process of retrieving the dataset provided and arranges it in a format that is suitable for analysis. It is common to encounter data cleaning challenges such as managing observations' null values and transforming incident date and time columns into a numerical format to facilitate analysis using machine learning models. As exploratory analysis of the data is conducted one of the most prevalent issues identified is a malfunction in the closed-circuit television (CCTV) as a result of misting of the CCTV camera pods and this incident tends to have a longer duration to solve.

3.2 Feature Engineering

In order to further sanitize the data, we utilise the CountVectorizer() function to derive numerical features from the text data, namely the "SUMMARY" field. This process transforms textual summaries into numerical vectors that are well-suited for our selected machine-learning techniques used for data analysis. We also utilise the pd. concat() function to merge numerical and text data, resulting in a comprehensive feature set for the modelling and prediction.

3.3 Data Splitting

The data splitting involves partitioning the dataset into separate sets for training and testing purposes. The training set, typically comprising 70% of the data, is used to train the model. Conversely, the testing set, typically accounting for 30% of the data, is employed to assess the model's performance on data it has not been exposed to previously.

3.4 Model Building and Training

This stage entails the selection and training of a machine-learning model. In this case, employ a RandomForest Regressor, Gradient Boosting Regressor, XGBoost, and Decision tree regressor. Random forests are an ensemble model that enhances prediction accuracy by combining numerous decision trees.

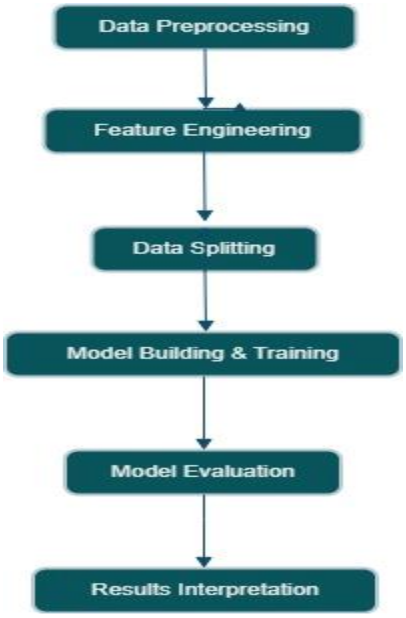

Figure 4: CCTV flow chart failure prediction model

3.5 Modelling and analysis

The Random Forest Regressor is an ensemble learning technique that creates many decision trees during training and calculates the average prediction produced by each tree for regression problems [4]. Every decision tree in the random forest is trained on a subset of the original dataset that is created using the bootstrapping method. At every node in the tree, the algorithm chooses the optimal feature from a randomly selected subset of features. It then divides the node into child nodes depending on this feature, to minimize the variance of the target variable within each node [5].
For this calculation, we assume the following:

Random Forest Prediction:
Y= the final prediction from the Random Forest, calculated as the average output from all decision trees.
n= the quantity of decision trees present in the random forest
Y_i= the prediction made by the i-th decision tree.
i= an index representing the i-th decision tree (from 1 to n).

$$Y = \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (1)$$

XGBoost (Extreme Gradient Boosting) Prediction:
Y_i= represents the output of the i-th decision tree.
α_i= weight.
n = is the total number of trees.

$$Y = \sum_{i=1}^{n} Y_i * \alpha_i \qquad (2)$$

Gradient Boosting Regressor Prediction:
F_θ= is the initial prediction.
F_i= represents the output of the i-th tree.
α_i= weight.
n = is the number of trees.

$$Y = F_\theta + \sum_{i=1}^{n} F_i * \alpha_i \qquad (3)$$

Decision tree prediction:
In a decision tree, predictions are made by traversing the tree from the root to a leaf node based on a set of splitting rules (conditions). The value at the leaf node becomes the predicted output.

Metrics calculation of all algorithms:
The algorithm's objective is to reduce the mean squared error (MSE), which quantifies the average squared discrepancy between the predicted and actual values and expression in equation (4).
n= the number of data points (samples) used in the calculation.
y_i= the predicted value for the i-th data point.
y= the actual value or true value for the.
$(y_i - \mathrm{y})^2$= the squared difference between the predicted and actual values for the i-th data point.
$\frac{1}{n}$ = the normalization factor, representing the mean of the squared differences.

$|y_i - \mathrm{y}|$ = the absolute difference between the predicted and actual values for the i-th data point.
ȳ = the mean of the actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathrm{y})^2 \quad (4)$$

The Mean Absolute Error (MAE) measures the average absolute difference between the predicted and actual maintenance times across all the data samples. The calculation is conducted using equation (5) below.

$$\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} |y_i - \mathrm{y}\,| \quad (5)$$

The R-squared (R^2) statistic quantifies the percentage of the variability in the dependent variable that can be accounted for by the model using equation (6).

$$R^2 = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{\mathrm{y}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y} - \bar{\mathrm{y}})^2)} \quad (6)$$

4 Analysis and results

This section seeks to assess the efficacy of the approach implemented in an actual case study within the railway industry. We evaluate the accuracy of the prediction of the proposed method using the dataset that consists of samples collected from sensors and control units of the fleet data. This investigation focussed on the technical incidents contributing to CCTV camera failure. The primary issue recorded relates to CCTV malfunction. The most time-consuming fault to resolve is mostly associated with the misted CCTV camera pods. The longest duration following an incident recorded between 2020 to March 2024 takes about 41.38 days to resolve. We use the downtime data associated with most incidents and the mean time between failure events to predict the overall downtime following the incidents using the Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, and Decision Tree Regressor. The Prediction of Actual and Real Data for CCTV Failure time and the time taken for in days to repair the longest incident associated with CCTV camera pods. Table (3) shows the algorithm performance of the machine learning model algorithms.

Model Algorithm	**Mean Squared Error (MSE)**	**R-squared (R²)**	**Mean Absolute Error (MAE)**
Random Forest	48.75	0.99	4.97
Gradient Boosting	136.64	0.97	8.27
XGBoost	66.28	0.98	6.88
Decision Tree	138.60	0.97	10.57

Table 3: Performance of Algorithms

The Random Forest algorithm shows the best performance with 99% accuracy in the prediction of the dataset, and it is the highest-performing model compared to other machine learning algorithms. Figures 5, 6, 7 and 8 shows the model performance depicting the actual data vs predicted data in the graph:

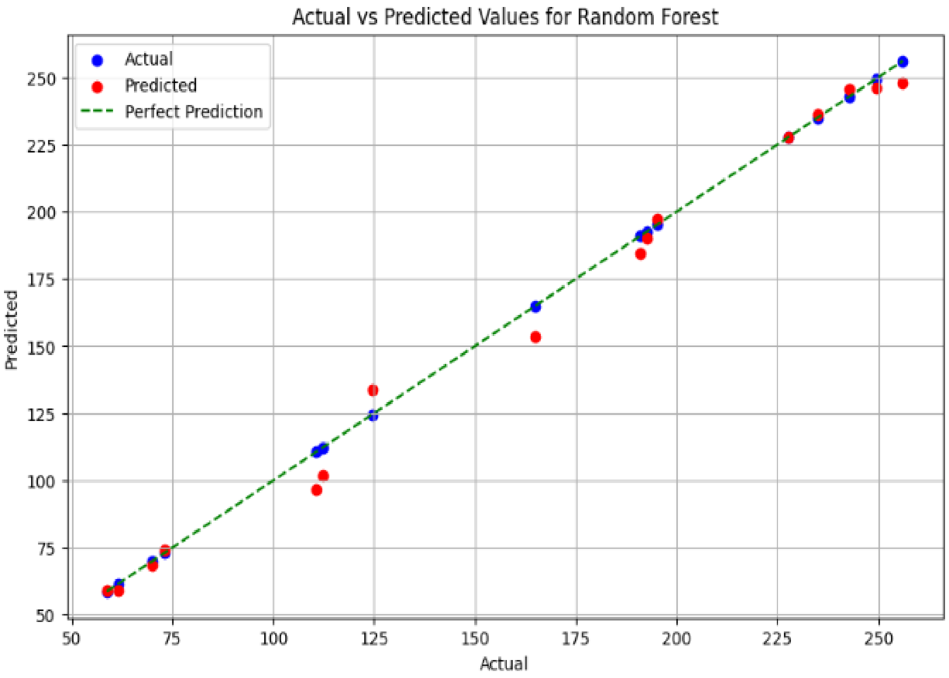

Figure 5: Random Forest Performance graph.

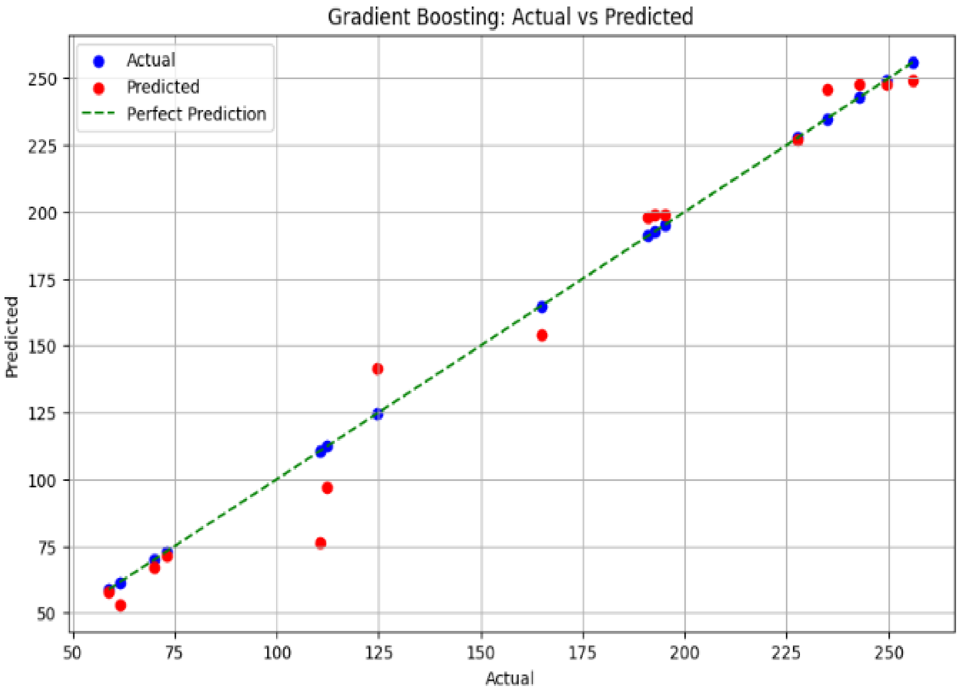

Figure 6: Gradient Boost Performance graph.

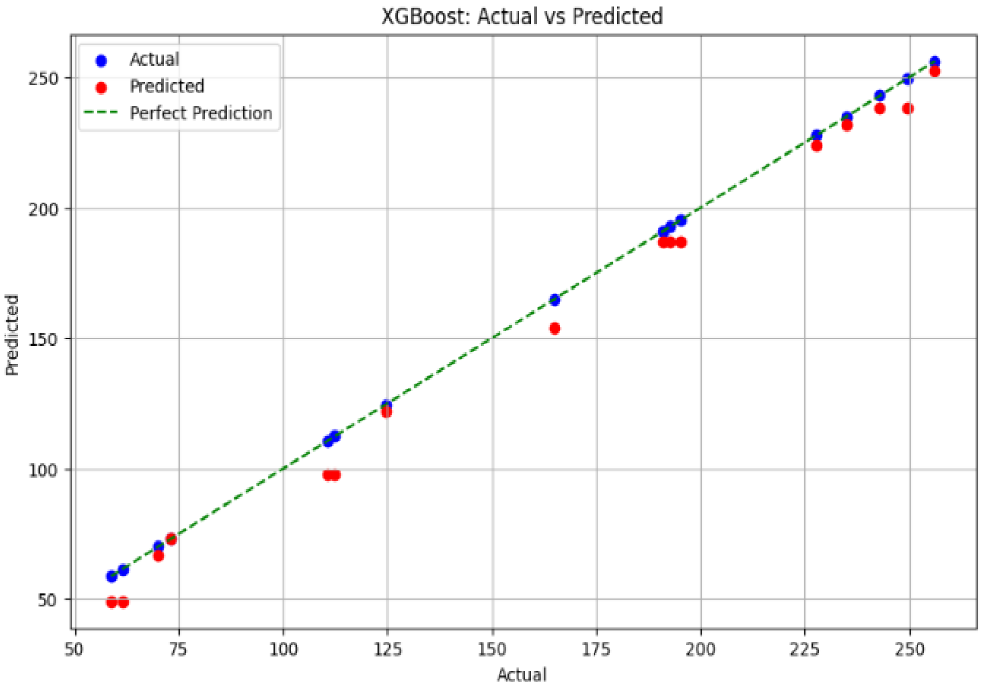

Figure 7: XGBoost performance graph.

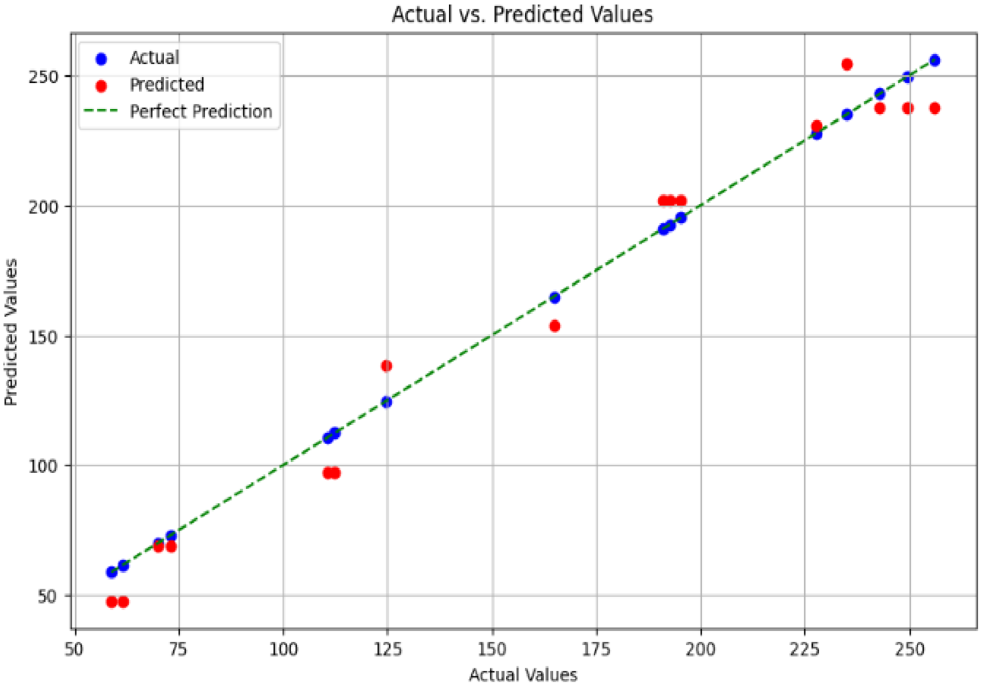

Figure 8: Decision Tree performance graph.

5 Conclusions and Contributions

This paper presents the study of CCTV failure incidents in the railway industry using predictive modelling techniques. The major objective is to identify the patterns of incidents contributing to delays and cancellation of service operations and which the CCTV system is identified as the most critical system and crucial to the overall performance of the fleet. The analysis presented indicated that the act of misting the CCTV camera pods was the fault that required the highest amount of time. The study has shown that predictive modelling is highly useful in analysing and forecasting faults of CCTV systems used in the railway industry. Our modelling attempt and result can boost the reliability of surveillance systems by precisely estimating resolution times and identifying important faults, which in turn supports proactive maintenance planning.

Acknowledgements

The authors would like to thank Siemens Mobility Limited, UK for funding this research project.

References

[1] Mistry, D. and Hough, J., 2024. Data-Driven Approach to State of Good Repair: Predicting Rolling Stock Service Life with Machine Learning for State of Good Repair Backlog Reduction and Long-Range Replacement Cost Estimation in Small Urban and Rural Transit Systems. *Transportation Research Record*, p.03611981241235197.

[2] Ahmad, M.W., Reynolds, J. and Rezgui, Y., 2018. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees, and regression trees. Journal of cleaner production, 203, pp.810-821.

[3] Putra, H.G.P., Supangkat, S.H., Nugraha, I.G.B.B., Hidayat, F. and Kereta, P.T., 2021, August. Designing machine learning model for predictive maintenance of railway vehicle. In 2021 International Conference on ICT for Smart Society (ICISS) (pp. 1-5). IEEE.

[4] Phyo, P.P., Byun, Y.C. and Park, N., 2022. Short-term energy forecasting using machine-learning-based ensemble voting regression. *Symmetry*, *14*(1), p.160.

[5] Yoon, J., 2021. Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. Computational Economics, 57(1), pp.247-265.

[6] De Simone, L., Caputo, E., Cinque, M., Galli, A., Moscato, V., Russo, S., Cesaro, G., Criscuolo, V. and Giannini, G., 2023. LSTM-based failure prediction for railway rolling stock equipment. Expert Systems with Applications, 222, p.119767.

[7] Siergiejczyk, M., Kasprzyk, Z., Rychlicki, M. and Szmigiel, P., 2022. Analysis and assessment of railway CCTV system operating reliability. Energies, 15(5), p.1701.

[8] del Castillo, A.C., Marcos, J.A. and Parlikad, A.K., 2023. Dynamic fleet maintenance management model applied to rolling stock. Reliability Engineering & System Safety, 240, p.109607.

[9] Alkali, B.M., Dinmohammadi, F. and Ramani, A., 2017, April. Towards implementing condition based maintenance policy for rolling stock critical system. In *The Stephenson Conference: Research for Railways*.

[10] Li, J., Doh, S.I. and Manogaran, R., 2023, February. Detection and Maintenance for Railway Track Defects: A Review. In IOP Conference Series: Earth and Environmental Science (Vol. 1140, No. 1, p. 012011). IOP Publishing.