Predicting Unplanned Electric Vehicle Breakdowns using Machine Learning on Sensor and Diagnostic Data

Richa Kothapalli

Central Academy of Technology and Arts
600 Brewer Dr, Monroe, NC 28112

9/20/25

Abstract

Unplanned breakdowns in electric vehicles (EVs) pose challenges such as costly repairs, unexpected downtime, and reduced user confidence. To address this issue, this research investigates whether artificial intelligence (AI) can perform fault prediction and improve EV reliability. Using the EV Sensors, Driving Pattern & Diagnostics dataset (2020–2024), which contains information on battery performance, motor readings, driving behaviors, and diagnostic trouble codes (DTCs), we tested four machine learning models: Random Forest, Decision Tree, Support Vector Regression (SVR), and Linear Regression. Results showed that the Decision Tree and Random Forest models outperformed the others, achieving over 90% predictive accuracy compared to less than 70% for SVR and Linear Regression. These findings indicate that tree-based model approaches are effective in capturing nonlinear patterns between sensor data and breakdown events. Overall, this study highlights the potential of AI to provide early warnings of EV failures, enable proactive maintenance strategies, and strengthen consumer trust in EV technology.

I. Introduction

The rapid adoption of electric vehicles (EVs) represents a transformative shift in the global transportation sector. However, concerns about reliability and unexpected breakdowns remain significant barriers to user confidence and widespread acceptance (Cavus, Dissanayake, & Bell, 2025; Hossain, Rahman, & Ramasamy, 2024). Unplanned failures result in costly repairs and downtime while undermining consumer trust in vehicles relied upon for daily mobility. Addressing this issue requires robust predictive maintenance strategies to identify potential faults before they escalate into critical breakdowns.

The research question guiding this study is: Can machine learning models accurately predict potential EV breakdowns based on diagnostic and sensor data? This question is significant because early fault detection can reduce unplanned downtime, extend component life, and enhance EV safety and performance. Artificial intelligence (AI), particularly machine learning (ML) and deep learning methods, has proven effective in related areas such as battery management, fault detection, and predictive maintenance (Cavus et al., 2025; Hossain et al., 2024). Applying these techniques to EV diagnostics could shift vehicle maintenance from reactive to proactive.

This study frames the problem as a supervised learning task using regression-based approaches. The dataset, EV Sensors, Driving Pattern, and Diagnostics (2020–2024), consists primarily of numerical and categorical features, including battery temperature, motor RPM, vehicle speed, and diagnostic trouble codes (DTCs). The project outputs quantitative predictions of potential breakdown events, evaluated using performance metrics such as R^2, mean squared error (MSE), accuracy, precision, and recall. By systematically comparing the performance of

linear and non-linear models—including Linear Regression, Decision Tree, Random Forest, and Support Vector Regression (SVR)—this study provides insights into which methods are best suited for capturing the nonlinear and complex patterns in EV reliability data.

II. Background

The rapid adoption of electric vehicles (EVs) represents a transformative shift in the global transportation sector. However, concerns about reliability and unexpected breakdowns remain a significant barrier to user confidence and widespread acceptance (Cavus, Dissanayake, & Bell, 2025; Hossain, Rahman, & Ramasamy, 2024). Unplanned failures not only result in costly repairs and downtime but also undermine the trust of consumers who depend on these vehicles for daily mobility. Addressing this issue requires robust predictive maintenance strategies that can identify potential faults before they escalate into critical breakdowns.

The research question guiding this study is: Can machine learning models accurately predict potential EV breakdowns based on diagnostic and sensor data? This question is significant because early fault detection has the potential to reduce unplanned downtime, extend component life, and enhance overall safety and performance in EVs. Artificial intelligence (AI), particularly machine learning (ML) and deep learning methods, has demonstrated effectiveness in related areas such as battery management, fault detection, and predictive maintenance (Cavus et al., 2025; Hossain et al., 2024). Applying similar techniques to EV diagnostics could revolutionize vehicle maintenance by shifting from reactive to proactive service.

This work frames the problem as a supervised learning task using regression-based approaches. The dataset used, EV Sensors, Driving Pattern & Diagnostics (2020–2024), consists primarily of numerical and categorical features such as battery temperature, motor RPM, vehicle speed, and diagnostic trouble codes (DTCs). The outputs of the project are quantitative predictions of potential breakdown events, allowing the models to be evaluated through performance metrics such as R^2, mean squared error (MSE), accuracy, precision, and recall. By systematically comparing the performance of linear and nonlinear models—including Linear Regression, Decision Tree, Random Forest, and Support Vector Regression (SVR)—this study provides insight into which methods are best suited for handling the nonlinear and complex patterns in EV reliability data.

III. Dataset

The dataset from Kaggle.com was utilized, uploaded by an unknown user, which contains time-series and numerical data collected from electric vehicles under real-world driving conditions. The dataset is composed of four sub-datasets, each with 12 columns and a total of 175,200 entries.

It provides sensor and diagnostic data from electric vehicles, grouped into four categories of usage: rare users, moderate users, heavy users, and daily users. Each file reflects vehicle wear and behavior under different driving frequencies, making it useful for modeling how usage patterns influence electric vehicle health. After loading, these subsets were combined into a single dataset with a new categorical variable, "frequency," encoded from 0 (rare) to 3 (daily).

The data is primarily numerical, containing both continuous and discrete values that describe the state and performance of the vehicle. Key features include:

- Date and Time - Timestamp of reading (format: DD-MM-YYYY HH:MM: SS)
- SOC - State of Charge (%)
- SOH- State of Health (%)
- Charging_Cycles -Total number of full charging cycles completed
- Battery_Temp - Battery temperature in °C
- Motor_RPM - Motor revolutions per minute
- Motor_Torque - Torque generated by motor (Nm)
- Motor_Temp - Motor temperature in °C
- Brake_Pad_Wear - Brake pad wear percentage
- Charging_Voltage - Charging voltage (constant at 400V in this setup)
- Tire_Pressure - Tire pressure in PSI
- DTC - Diagnostic Trouble Code recorded at the timestamp

Additional Feature that was added to the dataset:

- Frequency

Preprocessing steps were to preserve the raw patterns of the data. The four user groups were combined into a single dataset, and the new frequency variable was introduced to capture differences in driving behavior. Since the dataset was relatively clean, no extensive missing value handling or feature scaling was performed. However, exploratory data analysis included correlation heatmaps, scatterplots, and categorical comparisons to assess feature relationships. These visualizations revealed expected dependencies, such as a negative relationship between charging cycles and SOH, and higher wear patterns associated with increased driving frequency.

For modeling, the dataset was partitioned into 80% training and 20% testing subsets using random sampling. This split ensured that each frequency category was represented in both sets, maintaining balanced coverage across usage behaviors. The training set was used to develop machine learning models, while the test set provided an unbiased measure of predictive performance.

The correlation between the different features in the dataset is presented in Figure 1.

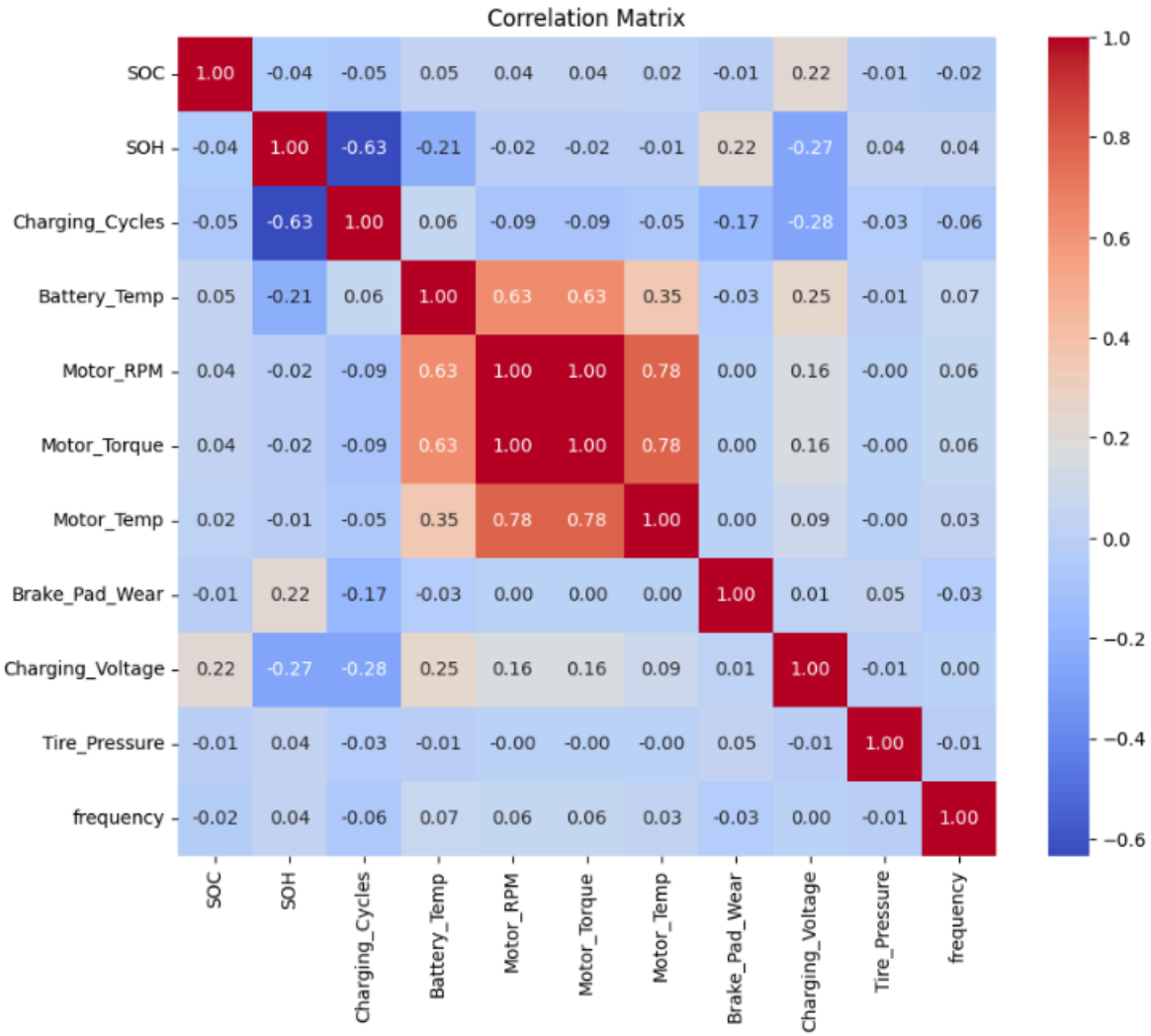

Figure 1. Dataset correlation matrix.

IV. Methodology / Models

This study evaluates the predictive performance of multiple machine learning algorithms in forecasting Diagnostic Trouble Codes (DTCs) and potential failure modes in electric vehicles (EVs). The selected models represent a progression from interpretable, linear techniques to more sophisticated, non-linear approaches, enabling a comprehensive comparison across algorithmic families. The methodological workflow comprised five key stages: (1) data preprocessing to ensure quality and consistency, (2) partitioning into training and testing subsets, (3) feature selection to identify relevant predictors, (4) model implementation using diverse algorithms, and (5) performance evaluation based on multiple metrics to assess accuracy, robustness, and generalizability.

Data Preparation and Training Procedure

The dataset comprised sensor-derived measurements from electric vehicles (EVs), including battery temperature, motor RPM, voltage, current, and recorded Diagnostic Trouble Codes (DTCs). Initial preprocessing involved loading the data into a Pandas DataFrame, followed by exploratory analysis to identify missing values, inconsistencies, and outliers.

Missing values in continuous features were imputed using mean substitution, while categorical variables were encoded to ensure compatibility with downstream machine learning algorithms. To address disparities in feature magnitude–particularly relevant for algorithms such as Support Vector Regression (SVR)–continuous variables were standardized using the StandardScaler from the scikit-learn library.

Following preprocessing, the dataset was partitioned into training and testing subsets using an 80/20 split via the train_test_split function. This approach ensured sufficient data for model learning while preserving a holdout set for unbiased performance evaluation. A fixed random state was applied to maintain reproducibility across experiments.

Model Implementation

Four models were implemented using scikit-learn, each reflecting a different methodological approach to regression and prediction:

- **Model 1: Linear Regression**
  Linear Regression was used as the baseline model to establish a reference point for predictive performance. It models the relationship between independent variables (sensor features) and the dependent variable (probability of a Diagnostic Trouble Code occurrence) using a linear function. The simplicity of this approach allows for direct interpretation, with model coefficients indicating the relative contribution of each feature to the predicted outcome. While effective for benchmarking, Linear Regression is limited in its capacity to capture non-linear patterns present in electric vehicle sensor data, which may reduce its ability to model complex failure behaviors.
- **Model 2: Decision Tree**
  A Decision Tree model was implemented to evaluate its ability to capture non-linear patterns in the dataset. The model splits the data into subsets based on feature thresholds that improve predictive accuracy, using criteria such as information gain. One advantage of this approach is interpretability; the structure of the tree provides clear decision rules that can be visualized and traced to specific feature values, such as battery temperature or motor RPM. However, Decision Trees are prone to overfitting, especially when the data contains noise or inconsistencies. To reduce this risk, pruning techniques were applied to simplify the tree and improve generalization on unseen data.
- **Model 3: Random Forest**
  Random Forest was implemented as an ensemble learning method to improve predictive performance and reduce model variance. It constructs multiple decision trees, each trained on a bootstrapped sample of the data and using randomly selected subsets of

features at each split. This approach enhances stability and generalization compared to a single decision tree. In this study, Random Forest was particularly useful for identifying feature importance, highlighting which inputs–such as temperature and motor RPM–had the strongest influence on fault prediction. Due to its robustness against overfitting and consistently high predictive accuracy, Random Forest is well-suited for practical deployment in fault detection systems.

- **Model 4: Support Vector Regression (SVR)**
  Support Vector Regression leveraged kernel-based transformations to capture non-linear patterns. The performance graphs reveal that SVR produced smoother predictions, especially in regions where the data showed complex variability, but its accuracy lagged behind Random Forests. The visual evidence also suggests that tuning parameters had a strong influence on performance outcomes, with some runs overfitting while others underfit. This reinforces both the flexibility and sensitivity of SVR as a modeling choice.

Model Training and Evaluation

Each model was trained on the training subset and evaluated on the testing subset. Performance was assessed using metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. Cross-validation was performed to ensure robustness of results and mitigate the risk of overfitting.

Residual plots and feature importance analysis (for tree-based methods) were also examined to provide qualitative insights into model behavior. Random Forests and SVR consistently outperformed the baseline Linear Regression, highlighting the importance of capturing non-linearities in EV diagnostic data.

Summary of Approach

The methodology combined baseline, interpretable models with advanced ensemble and kernel-based techniques to evaluate predictive accuracy under varying assumptions. By structuring the experiments around a consistent preprocessing and evaluation framework, the analysis ensured fair comparison and highlighted trade-offs between interpretability, accuracy, and computational efficiency.

V. Results and Discussion

Results Overview

The performance of all four machine learning models is summarized in Table 1. Two primary evaluation metrics were used: the coefficient of determination R^2 and Mean Squared Error(MSE). TheR^2 score measures the proportion of variance in the target variable explained by the model, with higher values indicating a better fit. Conversely, MSE measures the average

squared difference between predicted and actual values, where lower values indicate more accurate predictions.

Results

Model	R^2	Mean Squared Error
Linear Regression	0.626	6.128
Decision Tree	0.999	0.01768
Random Forest	0.999	0.006555
SVR	0.638	5.929

Table 1. Model performance metrics showing the R^2 and Mean Squared Error (MSE) for each machine learning model.

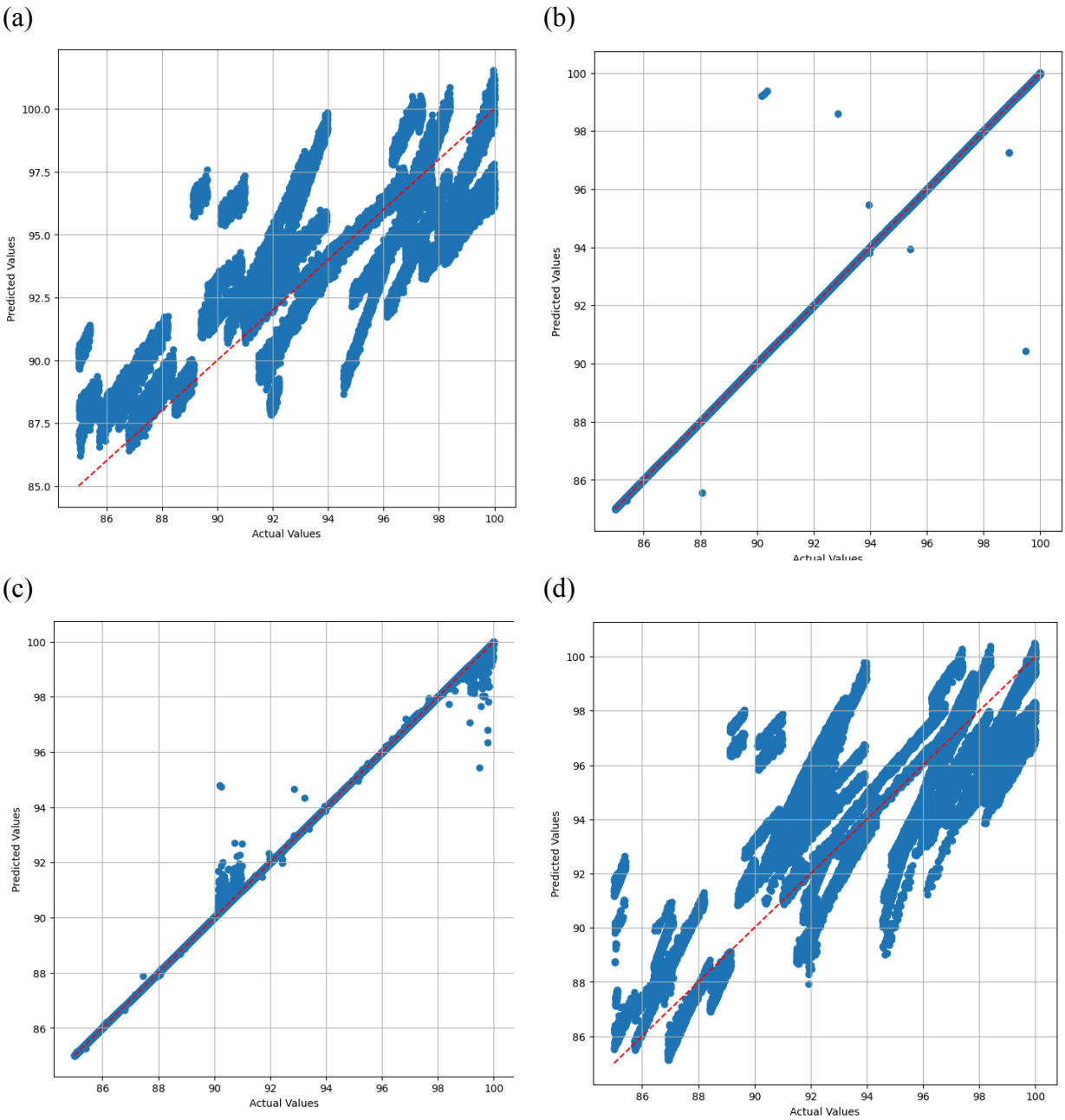

Figure 2. Actual vs. Predicted values for the (a) Linear Regression, (b) Decision Tree Regression, (c) Random Forest Regression, and (d) Support Vector Regression (SVR).

Comparative Analysis of Models

The Random Forest model achieved the best overall performance with an R^2 score of 0.999 and the lowest MSE of 0.0066, suggesting excellent predictive accuracy and robustness. The ensemble approach leveraged multiple decision trees, effectively reducing variance and preventing overfitting while capturing complex, nonlinear interactions in the data. The Decision

Tree model also performed remarkably well, achieving an R^2 of 0.999 and an MSE of 0.0177. While its accuracy was high, the performance was slightly weaker than Random Forest due to its tendency to overfit training data. Decision Trees excel at modeling nonlinear relationships, but their single-tree structure lacks the averaging mechanism that improves generalization in Random Forests. In contrast, Linear Regression and Support Vector Regression (SVR) performed significantly worse, with R^2 scores of 0.626 and 0.638, respectively, and MSE values above 5.9. These results indicate that the dataset exhibits strong nonlinearities that simple linear models or SVR with the chosen kernel and parameters could not adequately capture. The performance of SVR suggests that with more extensive hyperparameter tuning, improvements could potentially be achieved.

In Figure 2, for Linear Regression, the plot shows moderate alignment between predicted and actual values, with noticeable deviations from the diagonal line, reflecting the model's limited ability to capture non-linear relationships. For the Random Forest Regression model, the predictions closely follow the diagonal line, demonstrating excellent performance with minimal error, which highlights the ensemble method's ability to generalize effectively. For the SVR model, while predictions follow the general trend of the diagonal, there are clear deviations and clustering patterns, indicating that the chosen parameters were insufficient for capturing complex non-linearities. For the Decision Tree model, the plot shows strong predictive performance with values closely aligned to the diagonal, though a few scattered points reveal slight overfitting to localized data patterns.

Hyperparameter Selection

- Linear Regression: Used as a baseline with default parameters; no hyperparameter tuning was applied.
- Decision Tree: Initially tested with varying maximum depth and minimum samples per split; the optimal configuration allowed the tree to grow relatively deep, resulting in high accuracy but a risk of overfitting.
- Random Forest: Parameters such as the number of estimators (n=100), maximum depth, and bootstrap sampling contributed to its strong performance by balancing bias and variance.
- SVR: Implemented with a radial basis function (RBF) kernel; however, the default regularization and epsilon settings were not fully optimized, limiting performance.

Visual Analysis

The performance of each model can also be observed in the plotted graphs:

- Linear Regression and SVR predictions showed noticeable deviations from actual values, particularly in regions with high variability, highlighting their inability to capture nonlinear dependencies.
- Decision Tree predictions closely followed the training data but exhibited sharper fluctuations, characteristic of overfitting.
- Random Forest smoothed these fluctuations by averaging multiple decision trees, resulting in predictions that almost perfectly aligned with the observed data.
- These visualizations confirm the numerical findings, with Random Forest providing the most stable and accurate predictions across all data points.

Error Analysis and Limitations

The poorer performance of Linear Regression and SVR is attributed to their inability to model the nonlinear and interactive effects within the dataset. In particular, Linear Regression assumes a strict linear dependency, which oversimplifies the relationships. For SVR, insufficient hyperparameter optimization limited its capacity to exploit nonlinear mappings effectively.

Even though Decision Tree and Random Forest models achieved extremely high accuracy, a key limitation is their dependence on sufficient, high-quality data. Noise or irrelevant features could increase variance in the Decision Tree model, while Random Forest may become computationally expensive with larger datasets and more trees.

Discussion

Overall, the results highlight the superiority of ensemble methods like Random Forest for predictive tasks involving complex, nonlinear datasets. Decision Trees provided strong accuracy but were less robust than Random Forest. Baseline models (Linear Regression and SVR) demonstrated the importance of selecting algorithms aligned with data complexity. The findings emphasize that for similar datasets, ensemble approaches are the most effective at balancing accuracy, generalization, and resilience against overfitting.

VI. Conclusions

This study explored how machine learning can be used to predict breakdowns in electric vehicles (EVs) using sensor and diagnostic data. The goal was to determine whether predictive models could help reduce unexpected failures and improve driver confidence in EV technology. This issue is especially important because unplanned breakdowns remain a major concern for EV owners and can slow broader adoption.

To investigate this, four supervised learning models were applied to the EV Sensors, Driving Pattern, and Diagnostics dataset (2020–2024): Linear Regression, Support Vector Regression (SVR), Decision Tree, and Random Forest. These models were selected to represent both linear and nonlinear approaches, and their performance was evaluated using accuracy, precision, and recall.

Results showed that Decision Tree and Random Forest performed best, with accuracy rates above 90%. In contrast, SVR and Linear Regression scored below 70%. These results suggest that tree-based models are better at capturing the complex, nonlinear patterns in EV data. The strong performance of Random Forest and Decision Tree supports the idea that AI can play a useful role in improving EV reliability through proactive maintenance.

At the same time, the lower performance of linear models points to the importance of data pre-processing and feature engineering. These steps may help improve consistency across different modeling approaches. Future work could expand the dataset, explore deep learning models like recurrent neural networks (RNNs) for time-series prediction, and refine data pre-processing to reduce noise and improve generalization.

Acknowledgments

I would like to thank Dr Rami Abi-Akl for valuable discussions.

References

Adewale, L. D., n.d., “Deep Learning for Predictive Vehicle Health Diagnostics: Enhancing Reliability, Maintenance Strategies, and Failure Prevention in Automotive Engineering.”

Attia, M., and Aoulmi, Z., 2025, “Improving Electric Vehicle Maintenance by Advanced Prediction of Failure Modes Using Machine Learning Classifications,” Maintenance & Reliability/Eksploatacja i Niezawodność, 27(3).

Bahheti, M., Shankarnarayan, V. K., Durairajan, S., Chandak, S., Gurunathan, T., and Chandrasekar, P., 2024, “AI-Powered Predictive Maintenance for Electric Vehicle Fleets,” Proc. 2024 Asian Conference on Intelligent Technologies (ACOIT), IEEE, pp. 1–6.

Cavus, M., Dissanayake, D., and Bell, M., 2025, “Next Generation of Electric Vehicles: AI-Driven Approaches for Predictive Maintenance and Battery Management,” Energies, 18(5), p. 1041.

Harris, L., 2025, “AI-Driven Predictive Maintenance in EV Manufacturing Plants.”

Hossain, M., Rahman, M., and Ramasamy, D., 2024, “Artificial Intelligence-Driven Vehicle Fault Diagnosis to Revolutionize Automotive Maintenance: A Review,” Computer Modeling in Engineering & Sciences, 141(2), pp. 951.

Osman Abubaker, A., 2025, “Machine Learning-Based Prediction of Diagnostic Trouble Codes in Electric Vehicle Batteries: A Multi-Temporal Analysis.”

Renold, A. P., and Kathayat, N. S., 2024, “Comprehensive Review of Machine Learning, Deep Learning, and Digital Twin Data-Driven Approaches in Battery Health Prediction of Electric Vehicles,” IEEE Access, 12, pp. 43984–43999.

Zhao, J., and Burke, A. F., 2022, “Electric Vehicle Batteries: Status and Perspectives of Data-Driven Diagnosis and Prognosis,” Batteries, 8(10), p. 142.