Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.084165

ARTICLE

Interpreting Electric Vehicle Powertrain Fault Diagnosis Models Using Multimodal Large Language Model-Based Permutation Feature Importance and Leave-One-Feature-Out Importance Analysis Agents

Jaeseung Lee[1] **and Jehyeok Rew**[2,*]

[1]School of Electrical Engineering, Korea University, Seoul, Republic of Korea
[2]Department of Data Science, Duksung Women's University, Seoul, Republic of Korea

*Corresponding Author: Jehyeok Rew. Email: jhrew@duksung.ac.kr

Received: 17 April 2026; Accepted: 28 July 2026; Published: 28 August 2026

ABSTRACT: Accurate and interpretable fault diagnosis of electric vehicle (EV) powertrains is essential for ensuring operational safety, reliability, and efficient maintenance. Undetected faults in key components such as motors, inverters, and batteries can lead to performance degradation and critical system failures. While machine learning (ML)-based fault diagnosis models have demonstrated strong predictive capability using multivariate sensor data, their black-box nature limits practical trust and adoption in real-world EV applications. In particular, understanding how individual sensor variables contribute to diagnostic decisions remains a major challenge. To address this issue, this study proposes a novel interpretability method for EV powertrain fault diagnosis that integrates permutation feature importance (PFI), leave-one-feature-out (LOFO) importance, and multimodal large language model (MLLM)-based analysis agents. Using a random forest-based fault diagnosis model as the predictive backbone, both global and counterfactual importance signals are extracted to characterize variable relevance from complementary perspectives. Dedicated MLLM-based agents are employed to automatically transform numerical tables and visual explanations from PFI and LOFO analyses into structured and domain-aware textual interpretations. An MLLM-based report generation agent synthesizes these interpretations, enabling consistency analysis, discrepancy identification, and integrated diagnostic reasoning. Experimental results on an EV powertrain fault diagnosis dataset demonstrate that the proposed method not only achieves robust fault classification performance but also produces coherent and accurate explanations aligned with physical characteristics of EV powertrains. The results highlight the potential of combining conventional ML with MLLM-based agentic reasoning to deliver trustworthy fault diagnosis systems for next-generation EVs.

KEYWORDS: Electric vehicle; powertrain; multimodal large language model; permutation feature importance; leave-one-feature-out importance; explainable artificial intelligence; fault diagnosis; AI agent

1 Introduction

Electric vehicles (EVs) are increasingly replacing internal combustion engine vehicles as a dominant solution for sustainable transportation, driven by growing demands for higher energy efficiency and reduced carbon emissions [1]. In EVs, the powertrain comprises key components such as the battery, inverter, and traction motor that are responsible for energy conversion and propulsion [2,3]. The reliable operation of these components is critical to ensuring overall vehicle performance. Consequently, accurate fault diagnosis of EV powertrain systems is essential for maintaining operational reliability and enabling effective maintenance [4].

To achieve these objectives, a wide range of machine learning (ML)-based fault diagnosis models have been proposed, leveraging sensor data such as voltage, current, temperature, and rotational speed to perform anomaly detection and fault type classification [5,6]. However, the outputs of these models are provided in a black-box manner, making it difficult for vehicle engineers to interpret the underlying rationale behind the predictions. Accordingly, there is a growing demand for explainable artificial intelligence (XAI) techniques that can elucidate the decision-making processes of ML models through visual and linguistic explanations [7].

Among various XAI techniques, permutation feature importance (PFI) [8] and leave-one-feature-out (LOFO) [9] analyses are widely adopted to quantify the contribution of individual input variables by measuring performance degradation under feature perturbation or removal. These methods provide valuable global and counterfactual insights into model behavior and have been shown to be effective in identifying influential features in ML-based fault diagnosis tasks.

However, the outputs of PFI and LOFO analyses are presented in the form of numerical tables and visual plots, such as bar charts or importance rankings. Despite the availability of quantitative attribution results, their holistic interpretation, including directional effects, cross-method consistency, and interaction patterns, still depend largely on subjective judgments by vehicle engineers. In other words, there is a notable lack of automated frameworks that can reliably convert interpretability outputs into consistent and human-readable explanatory narratives.

Meanwhile, recent advances in large language models (LLMs) have enabled highly sophisticated capabilities in language understanding and generation through large-scale pretraining [10,11]. Beyond textual data, ongoing developments in multimodal extensions allow LLMs to jointly process heterogeneous inputs such as images, graphs, and other unstructured artifacts [12]. These capabilities have opened new possibilities for automating the interpretation of ML models, including the automatic analysis of visualized XAI results and the integration of multiple explanation outputs into coherent explanatory reports. Recently developed multimodal LLMs (MLLMs), such as OpenAI GPT-5.2 [13], support the recognition of visual patterns and their transformation into structured natural language descriptions [14]. This makes them well suited for translating graphical explanation results, such as PFI and LOFO analyses visualizations, into textual interpretations that are accessible to vehicle engineers.

Motivated by these advances, this study proposes an MLLM-based multi-agent interpretation method for ML-based fault diagnosis models in EV powertrain systems. Specifically, PFI and LOFO analyses are first conducted to quantify sensitivity-based and counterfactual feature importance characteristics using static and operational sensor features. Unlike conventional XAI-based reporting approaches that mainly provide isolated attribution scores or static visualization summaries [15,16], the proposed method decomposes interpretability into multiple reasoning stages through role-specialized MLLM-based agents. The attribution outputs are interpreted by dedicated agents that translate numerical rankings and visual explanations into structured and domain-aware textual descriptions. In addition, aggregation and cross-validation agents analyze consistency and complementary relationships between PFI and LOFO results, while a final report generation agent consolidates all outputs into a unified interpretability report. Consequently, the proposed method extends conventional XAI reporting from static attribution visualization toward a structured and method-aware interpretability pipeline for EV powertrain fault diagnosis models.

The main contributions of this study can be summarized as follows:

1. We propose a novel MLLM-based interpretability method that translates PFI and LOFO analysis results into structured, domain-aware textual explanations. This enables explainable ML-based fault diagnosis for EV powertrain systems.

2. We construct a role-specialized multi-agent interpretability architecture with explicit prompt constraints for interpretation, aggregation, verification, and reporting, enabling controlled and method-aware explainability generation beyond conventional XAI reporting pipelines.
3. The proposed method was validated on a real-world EV powertrain fault diagnosis dataset. We demonstrate that the generated explanations are logically consistent and practically meaningful for diagnosis analysis and maintenance decision support.

The remainder of this paper is organized as follows. Section 2 reviews related works on EV powertrain fault diagnosis, PFI and LOFO-based model interpretability, and MLLM-based agent architectures. Section 3 describes the overall structure and operational workflow of the proposed method. Section 4 details the experimental setup and evaluation methodology, while Section 5 presents and discusses the experimental results. Section 6 provides a discussion of the proposed method, including its implications, limitations, and potential applications. Finally, Section 7 concludes the paper and outlines directions for future research.

2 Related Works

2.1 EV Powertrain Fault Diagnosis

EV powertrain fault diagnosis primarily focuses on key components such as the traction motor, inverter, and battery, as accurate diagnosis of faults in these subsystems is essential for ensuring vehicle safety and operational reliability. In recent years, ML-based methods have rapidly emerged as an alternative to traditional signal analysis-based diagnostic methods in the field of EV powertrain fault diagnosis. These studies aim to leverage multi-sensor data, including vibration, current, voltage, and temperature signals, while ensuring robustness under abnormal operating conditions and supporting real-time deployment in embedded or edge computing environments.

Choudhary et al. [17] proposed a deep learning-based early fault diagnosis model for traction motors by combining constant-q transform-based acoustic spectrograms with transfer learning. By leveraging bearing fault data as the source domain and EV traction motor data as the target domain, their approach achieved high diagnostic accuracy and training efficiency even under limited data availability. Palaiologou et al. [18] introduced a sequential binary classification-based fault diagnosis framework to effectively detect transient faults in EV inverters. Instead of employing a conventional multiclass classification scheme, their method first distinguishes between normal and faulty operating states, and subsequently reclassifies internal fault types such as misfire in a second stage, thereby improving the diagnostic accuracy for short-duration transient faults. Shete et al. [19] proposed battery fault diagnosis techniques for lithium-ion batteries used in EVs based on multi-layer perceptron (MLP) and radial basis function neural networks. In their study, various fault conditions were generated through MATLAB-based simulations, and data preprocessing was applied to enhance training stability. The performance of the two models was then compared to assess the feasibility of real-time battery condition monitoring.

However, existing studies emphasize improving classification performance for EV powertrain fault diagnosis, while comparatively little attention has been paid to model explainability. Given the complex and highly coupled interactions among electrical, mechanical, and thermal variables in EV powertrains, accuracy-based metrics alone are insufficient to interpret model decision processes or ensure diagnostic trust. As a result, there is a growing need for global interpretability techniques that can quantitatively and visually explain the contributions of individual variables to model predictions.

2.2 Model Interpretability Based on PFI Analysis

PFI has emerged as a widely adopted, model-agnostic interpretability technique for understanding the contribution of individual input variables to ML predictions. By measuring the degradation in predictive performance caused by randomly permuting feature values, PFI provides an intuitive and flexible means of assessing global feature relevance across diverse model architectures. As a result, PFI has been extensively explored in recent studies to enhance the transparency and robustness of ML models in both static and dynamic application settings.

Mehdiyev et al. [20] proposed an XAI method that integrates PFI with conformal prediction to incorporate predictive uncertainty into feature attribution. Their method enables uncertainty-aware interpretation across different permutation scenarios involving test and calibration data. Experimental results on a real-world manufacturing process monitoring task demonstrated the robustness and practical applicability of the proposed method. Mi et al. [21] proposed a permutation-based feature importance test, termed PermFIT, to identify and statistically validate important features in complex ML models without requiring model refitting. By providing valid statistical inference on feature relevance, PermFIT enables efficient and reliable interpretation across ML models. Fumagalli et al. [8] proposed incremental PFI (iPFI), a model-agnostic method for efficiently estimating feature importance in online and dynamic learning settings. The method supports incremental updates under data streams and concept drift while providing theoretical guarantees on approximation quality. Experiments on benchmark datasets showed that iPFI achieves performance comparable to batch PFI with improved computational efficiency.

Collectively, these studies show that PFI-based interpretability has progressed from simple post-hoc ranking to more robust frameworks incorporating uncertainty estimation, statistical validation, and computational efficiency. Despite these advances, existing approaches largely remain focused on numerical or statistical outputs, providing limited support for generating structured and domain-aware explanations for practitioners. This gap motivates the integration of PFI with MLLM-based agents to bridge quantitative feature attribution and human-centered interpretability.

2.3 Model Interpretability Based on LOFO Analysis

LOFO analysis has been widely studied as an intuitive and model-agnostic approach for assessing feature relevance in ML models. By evaluating the change in predictive performance resulting from the removal of individual features, LOFO provides a direct measure of each feature's functional contribution to classification outcomes. Owing to its simplicity and classifier-independence, LOFO has been applied across diverse domains to support both feature selection and interpretability in supervised learning tasks.

Liu et al. [9] proposed a LOFO wrapper method for feature selection that evaluates feature relevance by measuring changes in classification performance when each feature is individually removed. The method is model-agnostic, intuitive, and allows users to control computational cost by limiting the number of cross-validation runs to a constant multiple of the feature count. Experimental results demonstrated that the proposed LOFO strategy improves prediction accuracy while providing practical insights into feature importance for data classification tasks. Daimi and Iqbal [22] proposed an ensemble-based ML framework for predicting academic success and student dropout, integrating feature engineering and importance analysis. Their results showed that incorporating enrollment-related features improved an area under the receiver operating characteristic curve (ROC-AUC) for both ensemble models and support vector classifiers. LOFO and Shapley additive explanations (SHAP)-based analyses further highlighted the influence of parental education, enrollment stability, and scholarships on academic outcomes.

Overall, prior studies show that LOFO-based analysis provides valuable insight into feature necessity by linking feature removal to performance degradation, making it effective for understanding model behavior in complex classification tasks. However, existing LOFO-based approaches largely rely on numerical comparisons and post-hoc visualizations, imposing an interpretive burden on domain experts. This limitation motivates the integration of LOFO analysis with MLLM-based agents to enable automated and domain-aware interpretation of counterfactual feature importance.

2.4 MLLM-Based Agents

MLLMs can jointly process heterogenous data modalities, including text, images, graphs, and tabular data, and integrate their semantic relationships through visual-language reasoning. Recent models such as GPT-5.2 [13] and HyperCLOVA X THINK [23] combine visual perception with natural language generation, enabling logical interpretation beyond surface-level description. When organized within a role-based agent framework, MLLMs enable staged tasks such as analysis, validation, and synthesis, facilitating scalable and consistent automation of interpretability and report generation workflows.

Abbineni et al. [24] proposed an MLLM-based hybrid retrieval-augmented generation framework for efficient large-scale literature exploration in circuit design. Their approach enables joint processing of textual documents and visual artifacts, supporting design-oriented question answering with step-by-step reasoning and evidence grounding from relevant literature. By combining multimodal understanding with retrieval-based contextualization, the system facilitates informed design queries and systematic knowledge exploration. Hui et al. [25] introduced WinSpot, a vision-action-based MLLM agent designed to automate tasks within graphical user interface (GUI) development environments. Unlike conventional methods that rely on structured user interface metadata such as the document object model or hypertext markup language, WinSpot operates directly on screenshot-based GUI representations. The framework incorporates specialized pretraining and alignment strategies to enable accurate recognition of GUI elements and effective mapping between natural language instructions and interface actions.

However, existing studies on MLLM-based agent systems have largely focused on task-oriented automation, such as literature retrieval, question answering, and user interface manipulation, while relatively limited attention has been paid to the systematic organization of model interpretability outputs. To address this gap, this study adopts collaborative agent-based architecture built upon the visual-linguistic reasoning capabilities of MLLMs. Within this method, an MLLM-based method is proposed to automatically interpret PFI and LOFO visualizations and aggregate variable-level explanations into a coherent diagnostic report.

To clarify the methodological differences between existing interpretability approaches and the proposed method, Table 1 presents a comparative summary of PFI-based, LOFO-based, and the proposed MLLM-based interpretability methods. This comparison highlights the distinct interpretability perspectives, levels of automation, and explanation capabilities that motivate the proposed method.

Table 1: Comparison of PFI-based, LOFO-based, and the proposed MLLM-based interpretability methods.

Aspect	PFI-Based Interpretability	LOFO-Based Interpretability	Proposed MLLM-Based Multi-Agent Method
Core objective	Quantify global sensitivity of model performance to feature perturbation	Assess counterfactual necessity of individual features via feature removal	Provide integrated, automated, and domain-aware interpretation of model behavior

(Continued)

Table 1 (continued)

Aspect	PFI-Based Interpretability	LOFO-Based Interpretability	Proposed MLLM-Based Multi-Agent Method
Interpretability perspective	Global sensitivity (perturbation-based)	Counterfactual necessity (removal-based)	Joint sensitivity–necessity interpretation with narrative synthesis
Dependency on human expertise	High (manual inspection and interpretation required)	High (manual reasoning over performance degradation)	Low (interpretation, aggregation, and validation are automated)
Explanation modality	Numerical scores and visual plots	Numerical scores and visual plots	Structured natural language explanations grounded in domain context
Ability to explain discrepancies across methods	Not supported	Not supported	Explicitly supported through aggregation and method-aware reasoning
Domain grounding	Implicit, depends on expert knowledge	Implicit, depends on expert knowledge	Explicitly enforced via prompt design and agent roles
Handling of feature redundancy	Limited	Explicitly captured	Explicitly analyzed and explained
Consistency and quality control	Not addressed	Not addressed	Ensured via dedicated cross-validation and judge agents
Scalability of Interpretation	Limited by expert effort	Limited by expert effort	High, scalable across datasets and models
Output usability for practitioners	Moderate	Moderate	High (decision-oriented reports)

3 Proposed Method

This section describes the overall architecture of the proposed method. Fig. 1 illustrates the overall architecture of the proposed method. Table 2 summarizes the roles, inputs, core functions, key constraints, and outputs of each MLLM-based agent employed in the proposed method. Fig. 2 illustrates the architecture of the MLLM used in the proposed method. Multimodal information, including tabular data and feature-importance visualizations, is encoded through a modality encoder and projected into a shared representation space, while textual inputs are processed through a text embedding module. These representations are jointly utilized by LLM to generate structured and domain-aware explanations.

3.1 ML-Based EV Powertrain Fault Diagnosis

EV powertrains consist of key components responsible for electrical power conversion and mechanical power transmission, including the battery, inverter, and traction motor. Variations in the operating states of these components are reflected in multivariate sensor signals such as current, voltage, temperature, vibration, and rotational speed. In this study, an ML-based fault diagnosis model is developed to classify EV powertrain operating conditions into normal and faulty states using multivariate operational data.

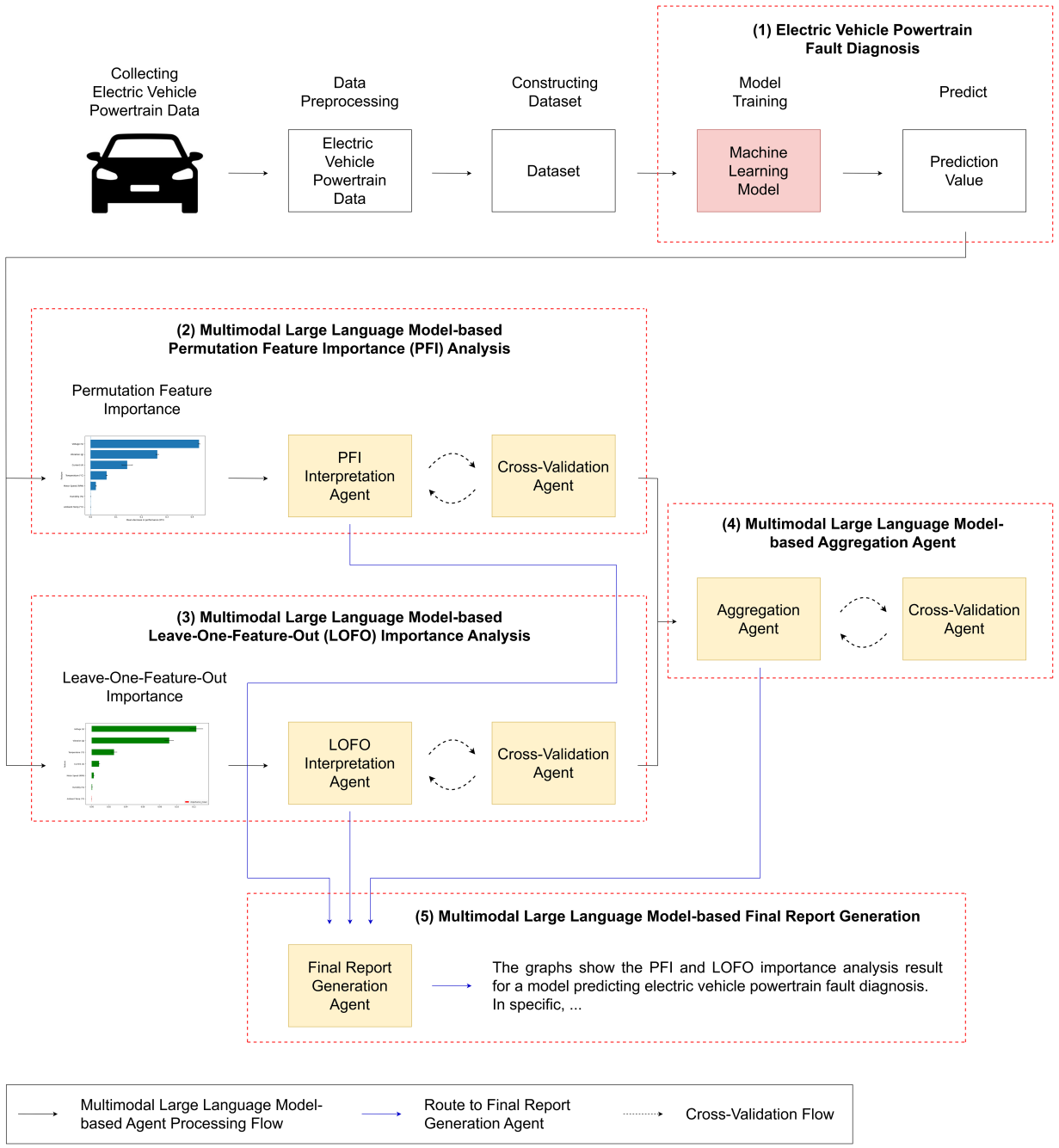

Figure 1: Overview of the proposed method.

Table 2: Summary of the proposed MLLM-based agents in the interpretability method.

Agent	Primary Role	Input	Core Function	Key Constraints	Output
PFI interpretation agent	Global sensitivity interpretation	PFI visualization, feature descriptions, fault class definitions	Translates permutation-based feature importance into structured, domain-aware explanations	• No numerical recomputation • Explicit reference to relative PFI magnitudes • Dominant/secondary/ negligible categorization • EV powertrain domain grounding	Structured textual explanation of global feature sensitivity
LOFO interpretation agent	Counterfactual necessity interpretation	LOFO importance results, feature descriptions, fault class definitions	Interprets performance degradation after feature removal and retraining	• Counterfactual reasoning only • Indispensable/moderate/ redundant categorization • Feature-wise independent analysis • Prohibition of speculative causal claims	Structured textual explanation of feature indispensability
Aggregation agent	Interpretability synthesis	Textual outputs of PFI and LOFO interpretation agents	Integrates global sensitivity and counterfactual necessity perspectives	• Text-level reasoning only (no raw data) • Identification of consistency, discrepancy, complementarity • Method-aware reasoning (sensitivity vs. necessity) • Strict domain grounding	Unified, method-aware integrated interpretation
Cross-Validation agent	Explanation verification	Output of one upstream agent (PFI, LOFO, or aggregation)	Validates explanation quality and methodological fidelity	• Non-generative role • No rewriting or extension • Evaluation across consistency, coherence, domain validity, and evidence alignment	Structured validation report (pass/minor issues/revision needed)
Final report generation agent	Narrative consolidation	Outputs of PFI, LOFO, and aggregation agents	Produces a single publication-ready interpretability report	• No new analysis or evidence • Non-redundant synthesis • Explicit articulation of PFI–LOFO complementarity • Evidence-bounded reasoning	Coherent, domain-consistent final interpretability report

For fault diagnosis, a random forest classifier is adopted as the predictive model [26]. Random forest is an ensemble learning method that constructs multiple decision trees using bootstrapped subsets of the training data and aggregates their predictions through majority voting [27]. This ensemble structure enables random forest to effectively model complex and nonlinear input-output relationships while mitigating overfitting through variance reduction. In addition, random forest naturally captures interactions among input variables without requiring explicit feature engineering, which is advantageous for EV powertrain systems where electrical, mechanical, and thermal variables are strongly coupled.

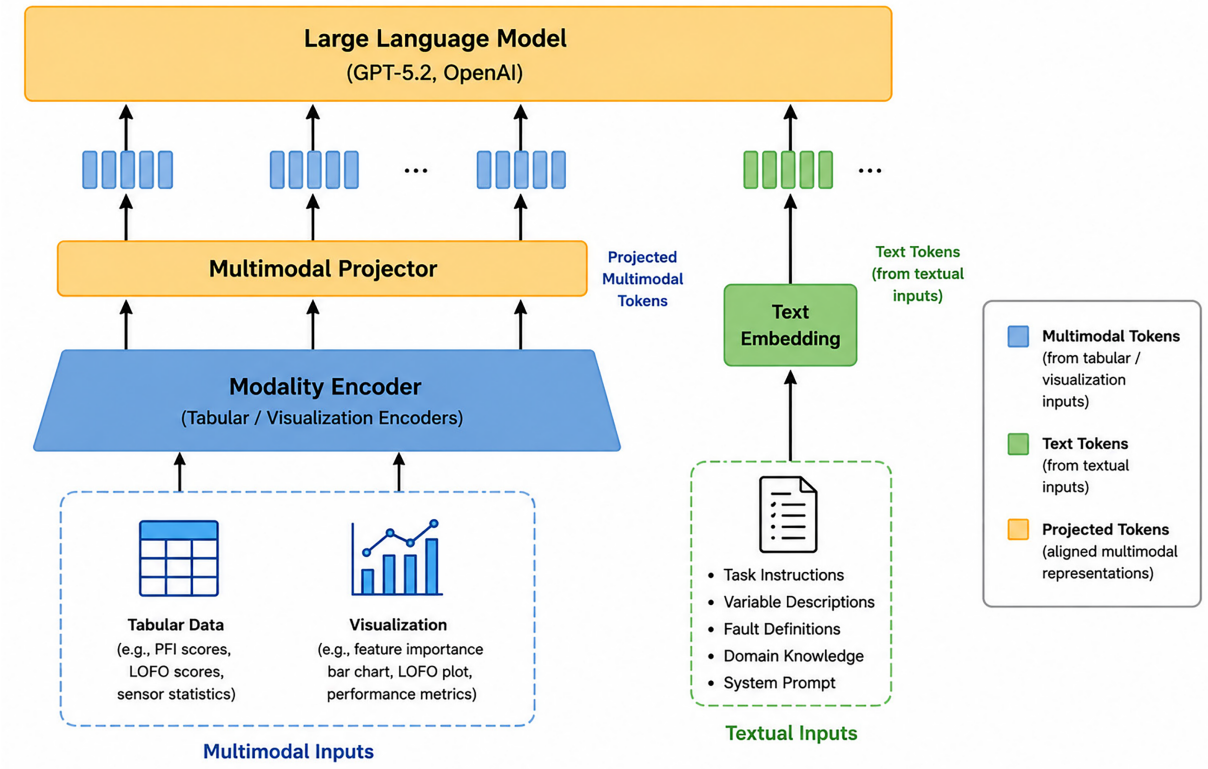

Figure 2: Architecture of the MLLM used in the proposed method.

From an interpretability perspective, this characteristic also explains the different behaviors of PFI and LOFO for correlated variables. Since random forest distributes predictive information across multiple decision trees using randomly selected feature subsets, correlated variables can jointly contribute to the prediction. Consequently, perturbation-based PFI measures the sensitivity of correlated variables by disrupting learned feature relationships, whereas LOFO retrains the model after feature removal, allowing the remaining correlated variables to compensate for missing information. Therefore, PFI reflects feature sensitivity, while LOFO captures feature necessity, resulting in different values for correlated variables.

To ensure robust and fair model performance, the hyperparameters of the random forest classifier were systematically tuned. Key hyperparameters, including the number of trees, maximum tree depth, minimum number of samples required for node splitting, and the number of features considered at each split, were optimized using validation data. This tuning process aims to balance model complexity and generalization capability, preventing overfitting while preserving sufficient representational power for capturing nonlinear fault patterns. The final hyperparameter configuration was selected based on classification performance on the validation set and was fixed for all subsequent interpretability analyses to ensure consistency between model prediction and explanation stages.

3.2 MLLM-Based Agent for PFI Analysis

PFI is a widely used global interpretability technique that quantifies the sensitivity of a trained model to perturbations in individual input variables by measuring the resulting degradation in predictive performance [8]. Fig. 3 presents the example visualization of PFI analysis. While PFI provides a numerical ranking of feature relevance, its outputs are presented as bar charts or summary statistics, which still require

domain expertise to interpret in a systematic and reproducible manner. This challenge is pronounced in EV powertrain fault diagnosis, where electrical, mechanical, and thermal variables are strongly coupled and where raw importance scores alone do not directly convey physically meaningful diagnostic insights.

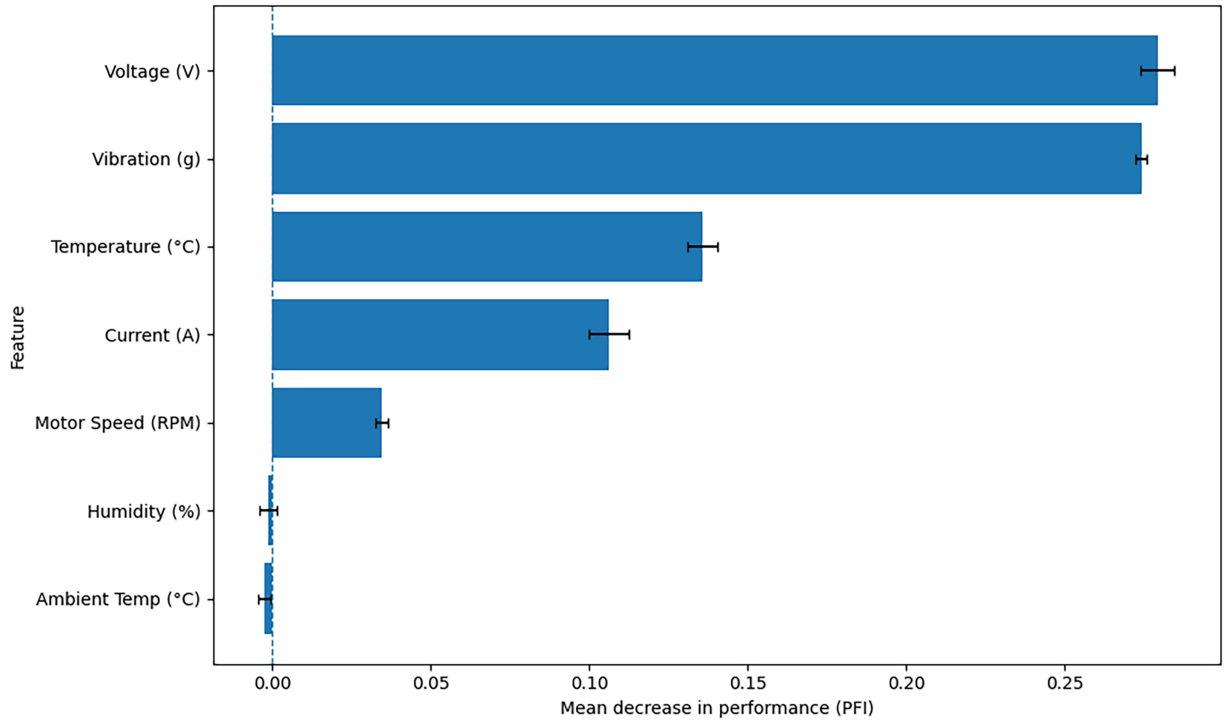

Figure 3: Example visualization of PFI analysis.

To address this limitation, this study introduces an MLLM-based agent designed to interpret PFI results for EV powertrain fault diagnosis models. The proposed agent does not perform numerical importance computation. Instead, it operates as an expert-level interpretation layer that transforms multimodal PFI outputs into structured and domain-aware textual explanations. The agent employs a multimodal input structure consisting of both visualization-based and text-based modalities. Specifically, the agent receives PFI feature importance visualizations represented as bar-chart images, textual descriptions of the input variables, and textual definitions of the EV powertrain fault classes. By jointly utilizing graphical importance patterns and domain-specific textual context, the agent generates structured explanations that relate observed feature importance characteristics to electrical, mechanical and thermal fault behaviors within the EV powertrain.

A central design principle of the proposed agent is the enforcement of a structured and table-oriented explanation format through prompt engineering [28]. Due to the length, the complete prompt for this agent is presented in Table A1. The prompt requires explicit reference to relative PFI magnitudes, distinction between dominant, secondary, and negligible features, and hierarchical organization of interpretations. This ensures direct alignment with experimental result tables and reduces interpretive variability. In addition, domain-grounding constraints are imposed to ensure that all explanations remain physically and operationally meaningful. The agent is guided to relate electrical, mechanical, and environmental variables to EV powertrain fault mechanisms involving the motor, inverter, and battery, while prohibiting unsupported assumptions or speculative causal claims.

Through this prompt-driven design, the MLLM-based PFI interpretation agent operates as a controlled and reproducible explanation module. This enables reliable and consistent interpretation of global feature importance in ML-based EV powertrain fault diagnosis models.

3.3 MLLM-Based Agent for LOFO Importance Analysis

While PFI provides a sensitivity-based view of global feature influence, it does not capture whether a feature is functionally indispensable once the model is allowed to adapt to its absence. To address this limitation, this study introduces an MLLM-based agent designed to interpret LOFO importance results, which quantify counterfactual feature necessity by measuring performance degradation after feature removal and model retraining [9]. Fig. 4 presents the example visualization of LOFO importance analysis. LOFO importance analysis complements permutation-based interpretation by identifying features that the diagnostic model cannot functionally replace without significant loss of discriminative capability.

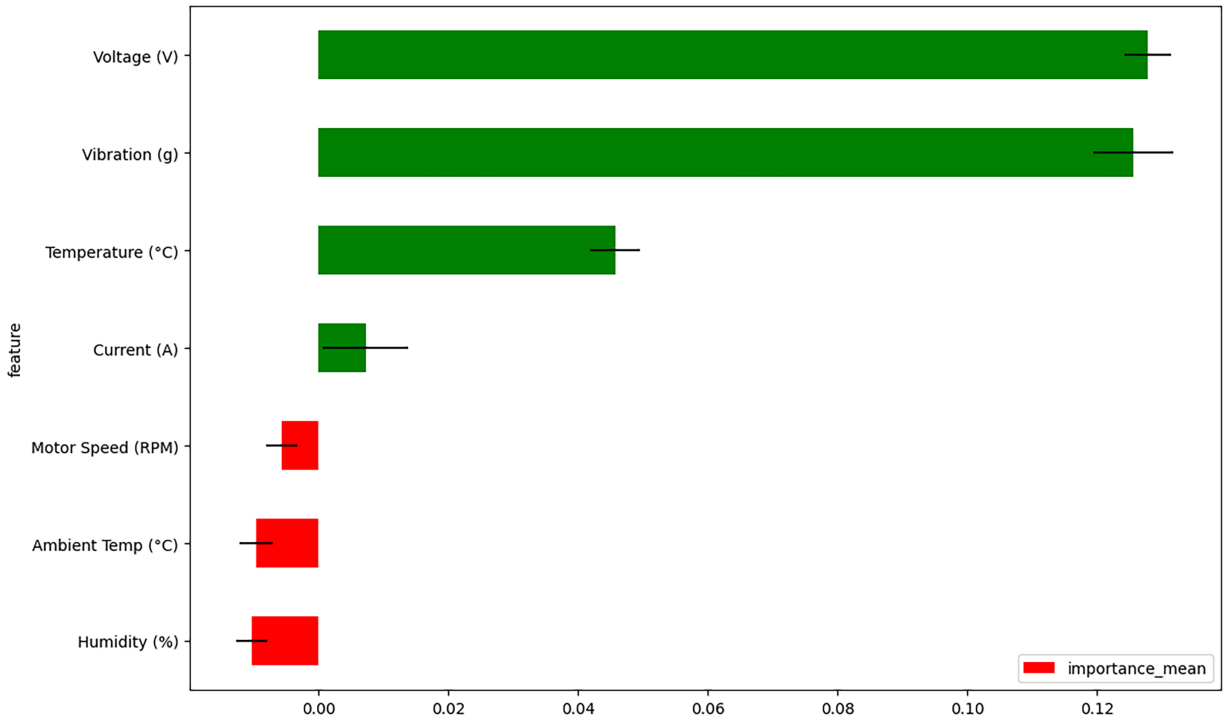

Figure 4: Example visualization of LOFO importance analysis.

The proposed LOFO interpretation agent employs a multimodal input structure consisting of both visualization-based and text-based modalities. Specifically, the MLLM receives LOFO importance visualizations represented as bar-chart images, textual descriptions of the input variables, and textual definitions of EV powertrain fault classes. By jointly utilizing graphical importance patterns and domain-specific textual context, the proposed agent generates structured explanations regarding the functional necessity and replaceability of individual variables within EV powertrain fault diagnosis models.

The prompt for the LOFO interpretation agent is presented in Table A2. The agent is guided by a prompt that enforces a structured and table-oriented explanation format focused on counterfactual reasoning. The prompt requires quantitative referencing of LOFO importance magnitudes and mandates categorical classification of features into functionally indispensable, moderately necessary, and redundant or non-essential groups. Each feature must be interpreted independently using dedicated sub items. This

design ensures that the generated explanations are concise, easily comparable across features, and directly compatible with tabular result presentations.

In addition to structural constraints, the prompt embeds strict domain-grounding requirements to ensure physically meaningful interpretation of counterfactual importance. The agent is instructed to relate electrical mechanical, and thermal variables to established EV powertrain fault mechanisms involving the motor, inverter, and battery subsystems. At the same time, the reasoning space of the MLLM is deliberately bound by prohibiting speculative causal claims or the introduction of external domain knowledge beyond the provided LOFO results and variable descriptions.

Through this prompt-driven design, the MLLM-based LOFO interpretation agent functions as a controlled mechanism for translating counterfactual performance degradation into structured and domain-consistent textual explanations. By distinguishing indispensable features from those that are redundant or substitutable, the agent provides a robust interpretation of feature necessity that enhances confidence in the functional reliability of ML-based EV powertrain fault diagnosis models.

3.4 MLLM-Based Agent for Aggregation of Interpretability Results

Although the PFI and LOFO-based interpretation agents provide complementary perspectives on feature relevance, their outputs reflect fundamentally different notions of importance, global sensitivity and counterfactual necessity, respectively. As a result, direct comparison or naive combination of these explanations can lead to incomplete or potentially misleading conclusions regarding model behavior. To address this challenge, this study introduces an MLLM-based aggregation agent designed to systematically integrate and reconcile the interpretation outputs generated by the PFI and LOFO agents.

The aggregation agent operates strictly at the textual explanations level and does not reanalyze raw feature importance values, numerical rankings, or visualization outputs. Instead, it receives as input the structured explanations produced by the upstream PFI and LOFO interpretation agents, which are grounded in the same dataset, predictive model, and EV powertrain fault taxonomy. By constraining the agent to reason solely over validated interpretation texts, the aggregation process preserves methodological consistency while avoiding redundancy or reinterpretation of underlying quantitative evidence. The prompt for this agent is presented in Table A3 due to the length.

The primary function of the aggregation agent is to identify and synthesize three types of relationships between permutation-based and counterfactual explanations. First, it detects features that are consistently emphasized by both PFI and LOFO analyses, thereby highlighting variables that are simultaneously influential under perturbation and indispensable under removal. Such convergence is interpreted as strong evidence of robustness and diagnostic reliability. Second, the agent examines discrepancies between the two methods, identifying features whose global sensitivity differs from their counterfactual necessity. These divergences are interpreted through method-aware reasoning, distinguishing between features that are influential yet substitutable, redundance due to correlated signals, or indirectly informative. Third, the agent articulates how PFI and LOFO jointly provide complementary insights into model behavior by answering different interpretability questions rather than conflicting ones.

To ensure domain validity, the aggregation agent is guided by explicit constraints that require all reasoning to remain grounded in EV powertrain physics and known fault mechanisms involving motor, inverter, and battery subsystems. The prompt prohibits speculative causal claims, numerical re-ranking, or the introduction of new assumptions beyond what is supported by the upstream interpretations. As a result, the aggregation process functions as a controlled synthesis layer that enhances interpretability without expanding the analytical scope beyond the available evidence.

The output of the aggregation agent is a structured interpretation that clarifies the stable diagnostic core of the model, contextualizes method-specific differences, and improves transparency regarding how different importance paradigms contribute to understanding model behavior. By integrating global sensitivity and counterfactual necessity perspectives, the aggregation agent plays a critical role in transforming multiple isolated explanations into a coherent and trustworthy interpretability narrative, thereby strengthening the reliability and practical applicability of ML-based EV powertrain fault diagnosis systems.

3.5 MLLM-Based Agent for Cross-Validation Analysis

While the proposed PFI, LOFO, and aggregation agents are designed to generate structured and domain-aware interpretations, the reliability of these explanations depends on their methodological fidelity and logical coherence. To address this requirement, this study introduces an MLLM-based cross-validation agent that serves as an independent verification layer within the interpretability pipeline [29].

Unlike upstream agents, the cross-validation agent does not generate new interpretations or extend existing analyses. Instead, it validates the textual output produced by one upstream agent at a time, ensuring that each explanation is assessed strictly within its intended methodological context. The prompt for this agent is presented in Table A4 due to its length.

The agent evaluates each interpretation along five predefined dimensions, including methodological consistency, logical coherence, domain appropriateness, evidence alignment, and scientific rigor of language. It verifies whether the explanation accurately reflects the role of the underlying XAI method, remains internally consistent, and avoids unsupported or speculative claims that exceed what importance-based analysis can justify.

A key design principle of this agent is its strictly non-generative role. The agent is prohibited from rewriting or augmenting the original explanation and instead produces a structured validation report that indicates whether the interpretation is fully validated, requires minor clarification, or exhibits notable inconsistencies. This design enhances interpretability reliability without altering the analytical content produced by upstream agents.

By incorporating this cross-validation agent, the proposed method establishes a quality control mechanism for MLLM-based interpretability. This verification layer reduces the risk of over-interpretation or domain-inconsistent reasoning, thereby strengthening the credibility and trustworthiness of automated explanations in EV powertrain fault diagnosis, particularly in safety-critical application scenarios.

3.6 MLLM-Based Agent for Final Report Generation

Although the PFI interpretation agent, the LOFO interpretation agent, and the aggregation agent each provide structured and method-specific insights into model behavior, their outputs remain distributed across multiple analytical stages. To transform these intermediate explanations into a single and coherent narrative, this study introduces an MLLM-based final report generation agent. The agent is designed to synthesize and refine the textual outputs produced by the three upstream agents into a coherent and domain-consistent interpretability report. Due to the length, the prompt for this agent is presented in Table A5.

The final report generation agent operates strictly at the level of high-level narrative integration. It receives the textual interpretation outputs of the PFI interpretation agent, the LOFO interpretation agent, and the aggregation agent, without access to raw feature importance values, plots, or numerical data. The agent is constrained from performing new analyses, reinterpreting visual artifacts, or introducing additional evidence. It merges overlapping statements, resolves redundancy, and organizes the existing interpretations into a logically structured explanation that preserves methodological fidelity.

A central objective of this agent is to articulate the complementary roles of permutation-based global sensitivity analysis and counterfactual feature necessity analysis in a unified manner. By leveraging the aggregation agent's output, the final report generation agent explains consistencies, such as features that are both globally influential and functionally indispensable, and method-consistent discrepancies arising from sensitivity vs. necessity perspectives. This design ensures that differences in feature attribution are not presented as contradictions, but as informative reflections of distinct interpretability criteria.

In addition, the prompt enforces strict evidence-based and domain-consistent reasoning. All explanations are grounded in the EV powertrain fault diagnosis context, referring to motor, inverter, and battery fault classes only to the extent supported by the upstream interpretations. The agent is prohibited from introducing new causal mechanisms, failure modes, or sensor interpretations, thereby maintaining a clear separation between quantitative evidence and explanatory synthesis.

Finally, the output of the final report generation agent is structured using fixed section headings that mirror the logical flow of a scientific interpretability narrative: an integrated overview of the pipeline, followed by summaries of PFI-based sensitivity, LOFO-based necessity, unified interpretation, and practical implications. By translating multi-stage interpretability results into a single and cohesive report, this agent completes the proposed MLLM-based interpretability method and enables reliable and transparent explanation of ML-based EV powertrain fault diagnosis models.

4 Experimental Settings

4.1 Dataset

The dataset used in this study consists of multivariate sensor measurements collected for the purpose of condition monitoring and fault diagnosis in EV powertrains [30]. Tables 3 and 4 present detailed descriptions and summary statistics of the input variables, respectively. The input feature set comprises seven continuous variables that reflect the electrical, mechanical, and thermal states of the EV powertrain during operation, including voltage (V), current (A), motor rotational speed (rpm), temperature (°C), vibration (g), ambient temperature (°C), and humidity (%). Voltage and current measurements characterize the power conversion behavior of the inverter and traction motor, while motor speed and vibration signals capture mechanical operating conditions and potential abnormalities. Temperature-related variables, including component temperature and ambient temperature, together with humidity, provide information on thermal load and environmental influences that affect powertrain performance and fault development.

Table 3: Input variables of the EV powertrain fault diagnosis dataset.

Variable	Unit	Description
Voltage	V	Measured voltage in the direct current and alternating current (DC/AC) power conversion stage
Current	A	Measured current flowing through the motor drive circuit
Motor speed	rpm	Rotational speed of the traction motor
Temperature	°C	Operating temperature of the powertrain system
Vibration	g	Mechanical vibration level of the motor and drivetrain components
Ambient temperature	°C	External environmental temperature
Humidity	%	Relative humidity of the operating environment

Table 4: Summary statistics of input variables of the EV powertrain fault diagnosis dataset.

Variable	Mean	Std.	Min	25%	50%	75%	Max
Voltage	0.7285	0.2244	0.0000	0.6129	0.7976	0.8915	1.0000
Current	0.4317	0.2452	0.0000	0.2370	0.3962	0.6040	1.0000
Motor speed	0.4430	0.2561	0.0000	0.2291	0.4132	0.6406	1.0000
Temperature	0.3802	0.2299	0.0000	0.1971	0.3701	0.5204	1.0000
Vibration	0.2814	0.2405	0.0000	0.1099	0.2178	0.3457	1.0000
Ambient temperature	0.4953	0.2860	0.0000	0.2496	0.4909	0.7436	1.0000
Humidity	0.5003	0.2901	0.0000	0.2466	0.5039	0.7498	1.0000

The output variable is defined as the fault label, and the dataset is formulated as a multiclass classification problem consisting of four operational states: normal operation, motor faults, inverter faults, and battery faults. As summarized in Table 5, the dataset contains 5000 samples for normal operation and 2000 samples for each fault category, including motor, inverter, and battery faults. Each class is designed to represent a major fault category that can occur within an EV powertrain, thereby covering a comprehensive range of realistic failure scenarios. This class formulation enhances both the practical relevance and the generalization capability of the fault diagnosis model. Table 5 provides a detailed description of the output classes.

Table 5: Output variables of the EV powertrain fault diagnosis dataset.

Variable	Count	Description
Normal operations	5000	Powertrain operating under stable and nominal conditions
Motor faults	2000	Bearing wear, winding short, or mechanical degradation
Inverter faults	2000	Power switch degradation or DC/AC conversion abnormalities
Battery faults	2000	Cell imbalance, internal short circuit, or overcharge/overdischarge

The dataset was constructed to emulate practical EV powertrain condition monitoring scenarios by incorporating sensor variables commonly measured in battery management systems (BMS), inverter control systems, and motor drive monitoring environments. The included variables reflect representative electrical, mechanical, thermal, and environmental operating conditions that are utilized in real-world EV diagnostic systems. In particular, voltage and current measurements are associated with power conversion and motor drive behavior, while vibration and motor speed variables capture mechanical operating characteristics of the traction system. Temperature-related and environmental variables reflect thermal load and operating environment conditions that can influence fault development and system reliability.

This dataset is utilized for training and evaluating ML-based fault classification models aimed at early fault detection in EV powertrains. It serves as a practically relevant benchmark for improving the operational reliability and safety of EV powertrains while supporting explainable fault diagnosis research. By simultaneously incorporating electrical, mechanical, thermal, and environmental variables, the dataset is well suited for evaluating explainable ML-based fault diagnosis methods that need to interpret complex interactions among heterogeneous sensor signals and diverse operating conditions in practical EV powertrains.

4.2 Experimental Setup

The experiments conducted in this study consist of two main stages: the development of an ML-based EV powertrain fault diagnosis model and the analysis of PFI and LOFO results using MLLM-based agents.

First, in the ML-based EV powertrain fault diagnosis model development stage, ML models were trained to classify fault types using sensor data collected from various EV powertrain operating conditions. To ensure generalizability in training and evaluation, the entire dataset was divided into training and test sets with a 7:3 ratio. For comprehensive performance comparison, 11 ML models were considered, including AdaBoost [31], k-nearest neighbor [32], logistic regression [33], naive bayes [34], decision tree [35], XGBoost [36], LightGBM [37], CatBoost [38], ExtraTree [39], gradient boosting [40], and random forest [26]. All models were trained under identical preprocessing conditions to ensure fair and consistent evaluation.

PFI and LOFO analyses were then conducted on the selected fault diagnosis model using a five-fold cross-validation (5-fold CV) scheme [41]. In each fold, the model was trained on four folds and evaluated on the remaining fold, and feature importance values were computed independently. The final PFI and LOFO results were obtained by aggregating the five cross-validation runs, reporting both the mean importance values and their standard deviations. This procedure enables robust estimation of feature importance while explicitly capturing variability across different training-testing splits, thereby improving the reliability of subsequent interpretability analysis.

In the interpretability stage, the aggregated PFI and LOFO results, together with their cross-validation statistics, were provided as inputs to the proposed MLLM-based agents. OpenAI GPT-5.2 [13] was adopted as the underlying MLLM. Compared with earlier models such as GPT-3.5 [42] and GPT-4o [43], GPT-5.2 provides enhanced long-context processing capability and a more robust hierarchical reasoning architecture, enabling stable generation of structured, long-form explanations. Its advanced vision-language reasoning capability allows graphical and tabular representations of PFI and LOFO results to be effectively translated into coherent natural language interpretations. Moreover, GPT-5.2 demonstrates high consistency in prompt-conditioned role execution and multi-agent collaboration, making it well suited for the staged automation pipeline comprising interpretation, aggregation, validation, and final report generation agents.

In addition, a user study was conducted to evaluate the quality and usability of the explanations generated by the proposed MLLM-based agents. A total of nine participants assessed the outputs of the PFI analysis agent, LOFO analysis agent, aggregation agent, and final report generation agent. The participants evaluated each explanation using a five-point Likert scale [44] and provided qualitative feedback through open-ended questions regarding the strengths and potential improvements of the generated interpretations.

4.3 Evaluation Metrics

In this study, we evaluated both the classification performance of the EV powertrain fault diagnosis model and the quality of the explanations generated from PFI and LOFO analyses using MLLM-based agents.

To assess the classification performance of the proposed ML-based fault diagnosis models, standard evaluation metrics for multi-class classification, including precision, recall, f1-score, and accuracy, are employed [45]. Precision represents the proportion of correctly predicted faulty instances among all instances predicted as faulty. Recall indicates the proportion of correctly identified faulty instances among all actual faulty instances. The f1-score is defined as the harmonic mean of precision and recall and reflects the balance between the two metrics. Accuracy measures the overall proportion of correctly classified instances. The mathematical formulations of the evaluation metrics are provided in Eqs. (1)–(4). For the multi-class fault diagnosis problem considered in this study, these metrics were computed for each class and subsequently averaged to evaluate overall classification performance [46].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{1}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative} \tag{4}$$

To evaluate the quality of the textual explanations generated by the MLLM-based agents for the PFI and LOFO analyses, an MLLM-as-a-Judge evaluation is employed [47]. MLLM-as-a-Judge refers to an assessment paradigm in which an MLLM directly compares and judges generated interpretations, enabling both quantitative and qualitative evaluation of explanation accuracy, consistency, and reasoning depth. This framework allows systematic evaluation of interpretability outputs.

To mitigate potential self-evaluation bias associated with the MLLM-as-a-Judge framework, the evaluation prompts were designed with strict boundaries. Specifically, the judge model was instructed to evaluate only the consistency between the generated explanations and the provided PFI and LOFO results, rather than generating new engineering interpretations or causal hypotheses. The prompts prohibited speculative causal claims, unsupported engineering assumptions, and information beyond the provided PFI and LOFO evidence. Instead, the judge was required to assess whether the generated explanations reflected the quantitative feature importance rankings and subsystem-level operational relevance indicated by the input evidence. Furthermore, the MLLM-as-a-Judge evaluation was complemented by an independent human-centered user study, thereby reducing the risk of self-evaluation bias and improving the overall reliability of the evaluation. The user study involved anonymous and voluntary participation, collected no personally identifiable or sensitive information, and was conducted in accordance with the applicable institutional regulations.

Five evaluation criteria are employed for the MLLM-as-a-Judge assessment and the user study evaluation. Content accuracy evaluates the alignment between the generated explanations and the overall trends and directional effects observed in the PFI and LOFO results. Explanation quality assesses the clarity, logical organization, and fluency of the technical narrative, while reasoning depth examines whether the explanations incorporate relative impact magnitudes and meaningful physical interpretations beyond surface-level descriptions. Consistency measures the absence of logical contradictions across explanations, and reference alignment quantifies the agreement between feature importance directions indicated by PFI and LOFO analyses and the corresponding statements in the generated text. Table 6 summarizes the evaluation criteria used in the MLLM-as-a-Judge and user study evaluations.

Each of the five criteria was quantitatively assessed using a five-point Likert scale [44], where 1 denotes very poor quality and 5 denotes excellent quality. For the MLLM-as-a-Judge evaluation, the MLLM was prompted to provide qualitative feedback addressing the validity of the interpretations and the consistency with domain-specific reasoning. Similarly, in the user study, participants were asked to provide open-ended feedback regarding the strengths and potential improvements of the generated explanations. These qualitative assessments complement the numerical scores, enabling a holistic evaluation of the interpretation results that capture both explanatory quality and domain alignment.

Table 6: Evaluation metrics for MLLM-as-a-Judge and user study.

Metric	Description
Content accuracy	Evaluates the extent to which the generated textual explanation faithfully reflects the overall trends, relative importance, and directional influence of variables observed in the PFI and LOFO analyses.
Explanation quality	Assesses the logical structure, clarity, and fluency of the explanation, including the appropriateness of technical terminology and coherence of the narrative.
Reasoning depth	Measures whether the explanation goes beyond surface-level trend description by incorporating inflection points, relative impact magnitudes, and connections to underlying physical or operational meanings.
Consistency	Examines the presence of logical contradictions or inconsistencies across explanations to evaluate the stability and coherence of the interpretative narrative.
Reference alignment	Quantitatively evaluates the agreement between the direction of changes observed in the PFI and LOFO results (increase or decrease) and the corresponding statements made in the generated explanations.

By adopting this evaluation scheme, which combines conventional performance metrics, MLLM-as-a-Judge evaluation, and human-centered user study assessment, the proposed method can verify both accurate fault prediction and reliable automation of interpretability. This integrated evaluation strategy ensures that the generated explanations are not only quantitatively sound but also qualitatively meaningful for real-world EV powertrain fault diagnosis.

5 Experimental Results

5.1 Evaluation of EV Powertrain Fault Diagnosis Performance

To classify EV powertrain fault conditions, a diverse set of ML models was comparatively evaluated. Table 7 summarizes the performance of each model across the evaluation metrics. For clarity, the best-performing model for each metric is highlighted in bold, while the second-best result is underlined.

The experimental results indicate that the random forest consistently achieved the highest performance across all evaluation metrics. This superior performance can be attributed to random forest's ensemble mechanism, which effectively captures nonlinear relationships and complex feature interactions by aggregating multiple decision trees. Such characteristics enable robust decision boundaries even in overlapping or ambiguous regions among different fault classes, thereby enhancing classification stability.

Other tree-based ensemble models, including LightGBM, XGBoost, CatBoost, and gradient boosting, also demonstrated strong overall performance. However, random forest exhibited comparatively lower model variance and reduced sensitivity to hyperparameter settings, making it well suited for EV powertrain applications where operational conditions such as load fluctuations and temperature variations are prevalent. In contrast, AdaBoost showed relatively weaker performance across all metrics, suggesting that the nonlinear dynamics and evolving fault characteristics inherent in EV powertrain data are not adequately captured by boosting approaches based on weak learners.

Table 7: Comparison of the performance of EV powertrain fault diagnosis model.

Model	Precision	Recall	F1 Score	Accuracy
AdaBoost	0.3070	0.5000	0.3804	0.3642
K-Nearest Neighbor	0.9715	0.9716	0.9713	0.9763
Logistic Regression	0.9751	0.9741	0.9746	0.9793
Naïve Bayes	0.9859	0.9858	0.9858	0.9893
Decision Tree	0.9893	0.9891	0.9892	0.9921
ExtraTree	0.9910	0.9910	0.9910	0.9933
LightGBM	0.9912	0.9916	0.9914	0.9936
XGBoost	0.9912	0.9915	0.9914	0.9936
CatBoost	0.9925	0.9928	0.9926	0.9945
Gradient Boosting	0.9933	0.9934	0.9933	0.9951
Random Forest	**0.9941**	**0.9943**	**0.9942**	**0.9957**

Based on these observations, random forest was selected as the final fault diagnosis model in this study. Subsequently, PFI and LOFO analyses, together with the MLLM-based agents, were applied to systematically structure and explain the model's decision-making rationale in an interpretable and domain-aware manner.

5.2 Evaluation of MLLM-Based PFI Analysis Agent

The PFI results obtained from the random forest-based EV powertrain fault diagnosis model are illustrated in Fig. 5 and summarized numerically in Table 8, while Table A8 presents the corresponding textual interpretation generated by the proposed MLLM-based PFI analysis agent. Overall, the generated interpretation accurately reflects the quantitative PFI distribution and the relative importance patterns observed across cross-validation folds. This demonstrates the agent's ability to faithfully translate numerical importance measures into structured and domain-consistent explanations.

The PFI results are strongly concentrated on a small set of electrically and mechanically relevant variables. Voltage exhibits the largest and most stable mean decrease in performance, and the agent correctly identifies it as the dominant global discriminator, linking its importance to DC/AC conversion stability, inverter integrity, and battery-related abnormalities. This interpretation is well supported by both the numerical evidence and the physical role of voltage in EV powertrain operation.

Vibration and current are consistently identified as the second and third-ranked contributors. The agent appropriately interprets vibration as a key mechanical indicator of motor and drivetrain degradation, reflecting its strong and stable importance. For current, the agent captures both its meaningful electrical role and its comparatively higher variability across folds, demonstrating sensitivity to uncertainty rather than relying solely on mean importance values.

Secondary variables, including temperature and motor speed, are correctly characterized as having limited global influence. Temperature is interpreted as an indirect indicator of accumulated thermal stress, while motor speed is described as contextual and potentially redundant, consistent with their relatively low PFI magnitudes. Environmental variables are accurately identified as negligible, with near-zero importance and noise-level variation.

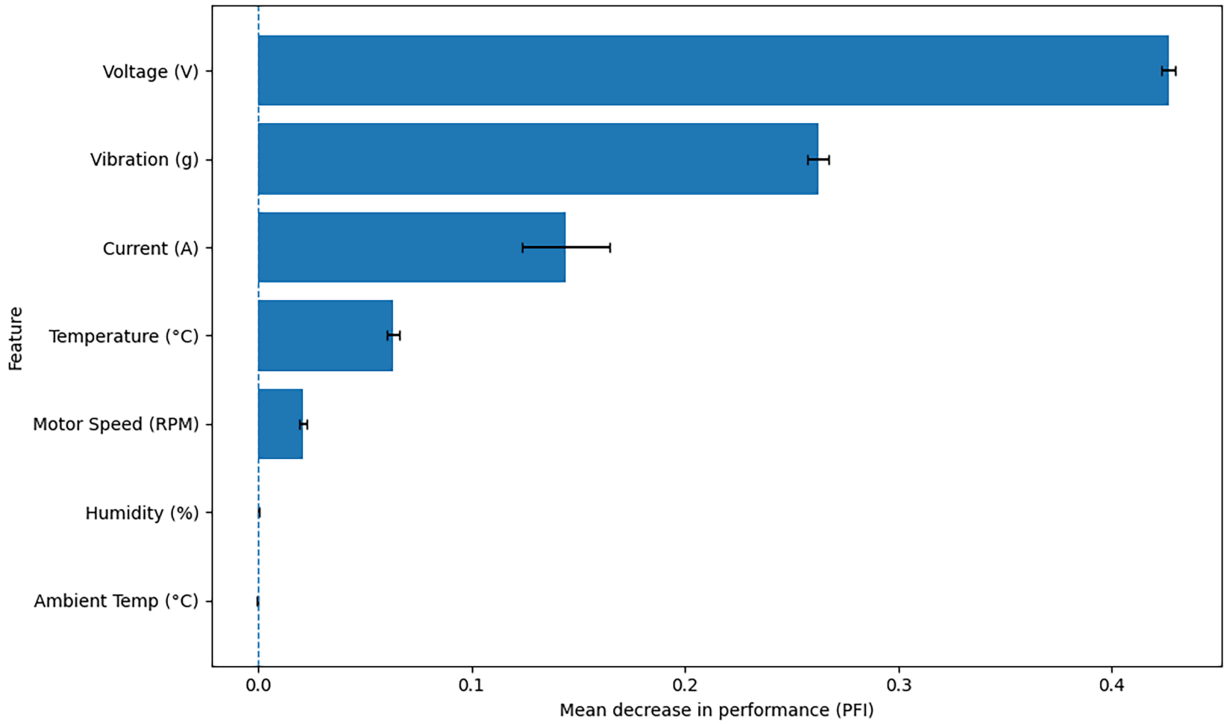

Figure 5: Visualization of PFI analysis result.

Table 8: Numerical results of PFI analysis.

Feature	Importance Mean	Importance Standard Deviation	Cross-Validation 1	Cross-Validation 2	Cross-Validation 3	Cross-Validation 4	Cross-Validation 5
Voltage (V)	0.42661	0.00320	0.42887	0.42465	0.42992	0.42749	0.42210
Vibration (g)	0.26234	0.00491	0.25857	0.25964	0.26230	0.26039	0.27079
Current (A)	0.14433	0.02040	0.14201	0.15370	0.10977	0.15827	0.15788
Temperature (°C)	0.06323	0.00303	0.06568	0.06682	0.06309	0.05974	0.06085
Motor Speed (rpm)	0.02124	0.00171	0.02284	0.01886	0.02016	0.02275	0.02161
Humidity (%)	0.00026	0.00035	0.00000	0.00058	0.00004	0.00000	0.00071
Ambient Temperature (°C)	−0.00001	0.00031	0.00000	−0.00044	0.00000	−0.00008	0.00044

It is noteworthy that ambient temperature exhibits a slightly negative mean PFI value, although the magnitude is extremely close to zero. In permutation-based importance analysis, a negative importance score indicates that feature permutation can occasionally improve model performance. This suggests that the feature was not consistently useful for prediction and may even have slightly degraded performance in some cross-validation folds. Such behavior can occur when a feature contributes noise or fold-specific patterns that do not generalize well. Given the extremely small magnitude of the observed value, the negative PFI score for ambient temperature suggests negligible diagnostic utility rather than meaningful predictive contribution.

Overall, the MLLM-based PFI analysis agent produces accurate, stable, and physically grounded interpretations of global feature importance. By organizing features by importance level and linking quantitative PFI results to EV powertrain fault mechanisms, the agent delivers model-faithful and practically meaningful explanations, confirming the effectiveness of the proposed prompt design for reliable PFI interpretation.

5.3 Evaluation of MLLM-Based LOFO Analysis Agent

The LOFO importance results obtained from the EV powertrain fault diagnosis model are illustrated in Fig. 6 and summarized in Table 9, while Table A9 presents the corresponding interpretation generated by the proposed MLLM-based LOFO analysis agent. Overall, the generated interpretation faithfully reflects the counterfactual dependency structure revealed by LOFO analysis, demonstrating the agent's ability to translate feature removal effects into structured and domain-aware explanations.

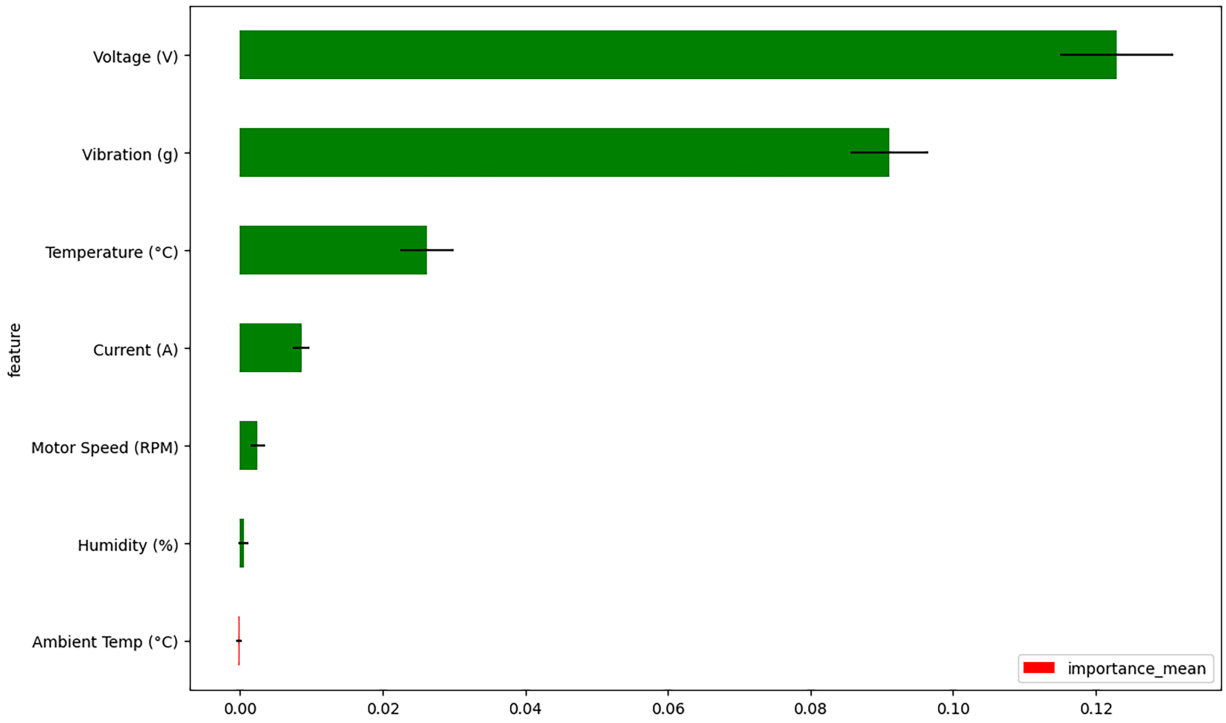

Figure 6: Visualization of LOFO analysis result.

Table 9: Numerical results of LOFO analysis.

Feature	Importance Mean	Importance Standard Deviation	Cross-Validation 1	Cross-Validation 2	Cross-Validation 3	Cross-Validation 4	Cross-Validation 5
Voltage (V)	0.12297	0.00795	0.11731	0.12307	0.13173	0.13143	0.11130
Vibration (g)	0.09105	0.00546	0.08702	0.10036	0.08486	0.08956	0.09345
Temperature (°C)	0.02628	0.00374	0.02689	0.02780	0.03218	0.02265	0.02188
Current (A)	0.00857	0.00121	0.00893	0.00804	0.01073	0.00805	0.00714
Motor Speed (rpm)	0.00250	0.00104	0.00446	0.00178	0.00178	0.00179	0.00267
Humidity (%)	0.00053	0.00071	0.00000	0.00178	0.00000	0.00000	0.00089
Ambient Temperature (°C)	−0.00017	0.00035	0.00000	0.00000	0.00000	−0.00089	0.00000

The LOFO results indicate that diagnostic performance depends on a small set of functionally indispensable variables. Voltage exhibits the largest and most stable performance degradation upon removal, and the agent correctly identifies it as the most critical feature, linking its necessity to the electrical stability of the DC/AC conversion stage and its role in distinguishing inverter and battery faults. This interpretation is well supported by both the quantitative results and EV powertrain physics.

Vibration is consistently identified as the second most indispensable feature. The agent accurately interprets vibration as an essential mechanical signal for capturing traction motor and drivetrain degradation, with the observed performance loss upon removal confirming that mechanical fault signatures are not readily substitutable after model retraining.

Temperature shows moderate LOFO importance, and the agent appropriately characterizes it as supportive contextual information reflecting accumulated thermal stress rather than a primary discriminative signal. This interpretation aligns with its intermediate performance degradation and highlights its complementary role relative to dominant electrical and mechanical variables.

In contrast, current, motor speed, humidity, and ambient temperature exhibit negligible LOFO impact. The agent correctly categorizes these variables as redundant or non-essential, noting that their removal does not significantly impair diagnostic performance. Current is interpreted as largely redundant with voltage-related information, while motor speed and environmental variables are framed as contextual or external factors with limited relevance to internal powertrain faults. A similar phenomenon can be observed for ambient temperature in the LOFO analysis, where a slightly negative mean importance value is reported. In LOFO, negative importance indicates that model performance improves marginally after the feature is removed and the model is retrained. This suggests that the feature was not beneficial for prediction and may have slightly reduced generalization performance in certain cross-validation folds. Such behavior can occur when the removed feature introduces weak redundancy or fold-specific patterns that do not consistently contribute to predictive performance. Therefore, the negative LOFO score supports the interpretation that ambient temperature does not provide essential fault-discriminative information in the current dataset and can even exert a minor detrimental effect on model generalization.

Overall, the results confirm that the MLLM-based LOFO analysis agent produces accurate, method-faithful, and physically grounded interpretations of counterfactual feature necessity, providing clear insight into which signals are required for reliable EV powertrain fault diagnosis.

5.4 Evaluation of MLLM-Based Aggregation Agent

The integrated interpretation generated by the proposed MLLM-based aggregation agent is summarized in Table A10, which synthesizes the outputs of the PFI and LOFO interpretation agents into a unified, method-aware explanation. The results show that the aggregation agent effectively consolidates complementary interpretability perspectives without reanalyzing raw importance values or introducing unsupported assumptions.

A major strength of the aggregation output is its clear identification of features that are consistently important across both global sensitivity and counterfactual necessity analyses. Voltage and vibration are robustly highlighted as core diagnostic signals, with PFI-based dominance linked to LOFO-based indispensability. This convergence is interpreted in a method-faithful manner, emphasizing that these variables are both heavily utilized by the model and non-substitutable when removed, thereby reinforcing their diagnostic relevance for electrically and mechanically driven fault classes.

The aggregation agent also appropriately handles discrepancies between PFI and LOFO by framing them as method-consistent rather than contradictory. Features such as current and temperature are interpreted through the lens of redundancy, substitutability, and contextual support, reflecting the fundamental differences between sensitivity-based and necessity-based explanations. This demonstrates the agent's ability to reason across interpretability methods at a conceptual level rather than merely summarizing feature rankings.

A representative example of method-aware discrepancy handling can be observed for the current variable. The PFI interpretation agent identified current as a dominant contributor (PFI mean = 0.14433), whereas the LOFO interpretation agent categorized it as a redundant feature due to minor performance degradation after feature removal (LOFO mean = 0.00857). The aggregation agent reconciled this discrepancy by explaining that current is highly sensitive under perturbation-based analysis, while retraining-based LOFO analysis suggests that voltage-related variables can partially substitute for its predictive role. This example demonstrates the ability of the aggregation agent to distinguish sensitivity-oriented importance from counterfactual indispensability in a method-consistent manner.

In addition, the agent consistently identifies environmental variables as negligible across both methods, reducing the risk of over-attribution and supporting a model-faithful understanding of diagnostic drivers. From a practical standpoint, the aggregated interpretation translates methodological insights into actionable implications for sensor prioritization and monitoring strategies, while maintaining evidence-based and domain-consistent reasoning.

Overall, the results confirm that the MLLM-based aggregation agent functions as an effective synthesis layer within the proposed interpretability method. By coherently integrating global sensitivity and counterfactual necessity perspectives, the agent enhances explanatory depth and reliability, supporting transparent and decision-relevant EV powertrain fault diagnosis.

5.5 Evaluation of MLLM-Based Report Generation Agent

The final interpretability report generated by the proposed MLLM-based report generation agent is summarized in Table A11, integrating the outputs of the PFI, LOFO, and aggregation agents into a single narrative. The results demonstrate that the agent effectively fulfills its role as a high-level synthesis layer, producing a coherent, non-redundant, and method-faithful interpretation without reanalyzing raw importance values or introducing unsupported assumptions.

The generated report exhibits strong global coherence and clearly articulates the complementary roles of PFI and LOFO, distinguishing global sensitivity from counterfactual necessity while preserving methodological boundaries. Dominant electrical and mechanical features, particularly voltage and vibration, are consistently identified as core diagnostic signals, whereas mid-tier variables such as current and temperature are interpreted in a method-aware manner that reflects redundancy, substitutability, and contextual support rather than inconsistency. Environmental variables are uniformly characterized as negligible, reinforcing a model-faithful understanding of feature relevance.

Importantly, the report maintains evidence-based and domain-consistent reasoning, grounding all interpretations in established EV powertrain behavior without introducing new causal claims. By translating multi-stage interpretability results into clear implications for sensor prioritization, monitoring strategies, and diagnostic reliability, the final report elevates individual XAI outputs into a unified, decision-relevant explanation. Overall, these results confirm the effectiveness of the proposed report generation agent as the final integration component of the interpretability method for EV powertrain fault diagnosis.

5.6 MLLM-as-a-Judge

This subsection evaluates the quality and reliability of explanations generated by the proposed MLLM-based agents using an independent MLLM-as-a-Judge method. Two MLLMs, GPT-5.2 [13] and HyperCLOVA X THINK [23], were employed to assess the outputs of the PFI analysis, LOFO analysis, aggregation, and final report generation agents using a five-point Likert scale across five criteria: content

accuracy, explanation quality, reasoning depth, consistency, and reference alignment, followed by qualitative feedback on strengths and weaknesses.

The quantitative results, summarized in Tables 10 and 11, indicate consistently high performance across all agents and evaluation criteria. Across two independent MLLM judges, the proposed agents achieved average scores above 4.8 on a five-point Likert scale, demonstrating strong alignment between the generated explanations and the underlying PFI and LOFO evidence. Table 12 further summarizes the mean evaluation scores and standard deviations across all criteria and judge models, providing an overall quantitative assessment of explanation quality and consistency.

Table 10: MLLM-as-a-Judge evaluation results using GPT-5.2 based on a five-point Likert scale.

Criterion	MLLM-Based PFI Analysis Agent	MLLM-Based LOFO Analysis Agent	MLLM-Based Aggregation Agent	MLLM-Based Report Generation Agent
Content accuracy	5.0	5.0	5.0	5.0
Explanation quality	5.0	5.0	5.0	5.0
Reasoning depth	5.0	5.0	5.0	5.0
Consistency	5.0	5.0	5.0	5.0
Reference alignment	5.0	5.0	5.0	5.0

Table 11: MLLM-as-a-Judge evaluation results using HyperCLOVA X THINK based on a five-point Likert scale.

Criterion	MLLM-Based PFI Analysis Agent	MLLM-Based LOFO Analysis Agent	MLLM-Based Aggregation Agent	MLLM-Based Report Generation Agent
Content accuracy	5.0	5.0	5.0	5.0
Explanation quality	5.0	5.0	5.0	4.0
Reasoning depth	5.0	5.0	5.0	5.0
Consistency	5.0	5.0	5.0	5.0
Reference alignment	5.0	5.0	5.0	5.0

Table 12: Average quantitative evaluation scores across MLLM-as-a-Judge.

Agent	Mean Score	Standard Deviation
MLLM-based PFI analysis agent	5.00	0.00
MLLM-based LOFO analysis agent	5.00	0.00
MLLM-based aggregation agent	5.00	0.00
MLLM-based report generation agent	4.90	0.32

Qualitative feedback, summarized in Tables 13 and 14, supports these findings. The PFI and LOFO analysis agents are consistently praised for method-faithful interpretation, clear structure, and effective use of stability information, with only minor suggestions for elaborating on edge cases or strengthening quantitative-severity linkage. The aggregation agent receives particularly strong validation for preserving the distinction between global sensitivity and counterfactual necessity while resolving discrepancies in a principled, evidence-based manner. The final report generation agent is recognized for producing a coherent synthesis, with minor stylistic refinements suggested.

Table 13: Evaluation result of feedback using GPT-5.2.

Agent	Result
MLLM-based PFI analysis agent	• Faithful reflection of PFI results: The interpretation accurately captures the relative importance ordering, magnitude differences, and near-zero contributions shown in the PFI table. • Clear structure and technical clarity: The explanation is well-organized (overview → dominant features → secondary variables → stability → implications) and uses appropriate domain-specific terminology. • Strong grounding in stability analysis: Discussion of standard deviation and cross-validation consistency is directly supported by the provided numerical results and enhances interpretability credibility.
MLLM-based LOFO analysis agent	• Accurate reflection of LOFO results: The interpretation precisely matches the LOFO importance rankings, magnitude gaps, and the distinction between indispensable, moderate, and redundant features as shown in the provided table. • Method-faithful counterfactual reasoning: Claims about "indispensability", "redundancy", and performance degradation are well aligned with what LOFO analysis can support and are consistently proportional to the reported mean importance values. • Clear structure and robust grounding: The explanation follows a logical progression (overview → essential features → redundant variables → stability → implications) and effectively integrates variability (standard deviation) to support robustness claims.
MLLM-based aggregation agent	• Faithful integration of PFI vs. LOFO roles: The output clearly preserves the upstream distinction between global sensitivity (PFI) and counterfactual necessity (LOFO), and uses that distinction consistently when describing agreements and discrepancies. • Strong method-aware discrepancy handling: Mid-tier divergences (e.g., current and temperature) are explained in a principled way via redundancy/substitutability vs. complementary context, without inventing new evidence. • Publication-ready structure and clarity: The report is logically sequenced (strategy → agreements → discrepancies → complementary insights → implications), with minimal redundancy and appropriately scientific tone.

(Continued)

Table 13 (continued)

Agent	Result
MLLM-based report generation agent	• Faithful synthesis of PFI and LOFO: The report accurately reflects upstream conclusions, clearly distinguishing sensitivity (PFI) from necessity (LOFO) without distortion. • Strong integrative reasoning: Method-aware explanations for consistencies and discrepancies (e.g., redundancy, substitutability, contextual support) go beyond result listing and provide coherent synthesis. • Publication-ready clarity and structure: The narrative is logically organized, concise, and uses appropriate scientific language, maintaining internal coherence throughout.

Table 14: Evaluation result of feedback using HyperCLOVA X THINK.

Agent	Result
MLLM-based PFI analysis agent	Strengths: • Accuracy: Matches PFI results precisely, including mean PFI values and standard deviations. • Clarity: Distinct sections with headings enhance readability; technical terms are well-explained. • Analytical Rigor: Links quantitative rankings to physical fault mechanisms and diagnostic utility. Weaknesses/Areas for Improvement: • Environmental Variables: While correctly labeled as negligible, the interpretation does not address why Ambient Temp shows slightly negative PFI values (≈ −0.00001). A brief note on potential noise or measurement artifacts could strengthen completeness. • Secondary Variables: The explanation could elaborate on why Temperature's moderate contribution (~0.063) is "indirect" vs. "immediate" but lacks mechanistic detail (e.g., thermal stress accumulation pathways).
MLLM-based LOFO analysis agent	Strengths: • Clarity & Faithfulness: The interpretation accurately reflects LOFO results (e.g., Voltage's mean = 0.12297 ranked #1, Ambient Temp's near-zero impact). • Analytical Rigor: Links feature importance to diagnostic roles (e.g., Voltage's role in electrical stability, Vibration's encoding of motor conditions). • Stability Analysis: Explicitly connects low standard deviations (e.g., Voltage std. ≈ 0.00795) to robust dependencies. Weaknesses/Areas for Improvement: • Ambient Temp Emphasis: Could explicitly contrast Ambient Temp's negative mean (−0.00017) with others to strengthen non-essentiality claims. • Severity Linkage: While severity levels are noted, deeper ties to quantitative metrics (e.g., std. deviations) would enhance precision.

(Continued)

Table 14 (continued)

Agent	Result
MLLM-based aggregation agent	Strengths: • Effective synthesis of PFI (sensitivity) and LOFO (necessity) perspectives, resolving agreements/discrepancies (e.g., Voltage as both dominant and indispensable). • Clear distinction between method-specific conclusions (e.g., Current's sensitivity in PFI vs. redundancy in LOFO). • Domain-grounded implications for sensor prioritization and diagnostic strategies. Weaknesses/Areas for Improvement: • No significant weaknesses were identified.
MLLM-based report generation agent	Strengths: • Effective synthesis of PFI and LOFO perspectives, highlighting both agreement (e.g., voltage, vibration) and method-specific nuances (e.g., current's sensitivity vs. substitutability). • Clear distinction between global sensitivity (PFI) and counterfactual necessity (LOFO), enhancing interpretability. • Strong domain grounding in EV powertrain diagnostics, with practical implications for sensor prioritization and fault mechanisms. Weaknesses/Areas for Improvement: • Minor redundancy in reiterating the irrelevance of environmental variables across sections, though this is structurally justified for integration. • Slightly abrupt transitions between PFI and LOFO discussions; smoother contextual bridging could improve readability.

Overall, the MLLM-as-a-Judge evaluation confirms that the proposed multi-agent method produces explanations that are accurate, consistent, and method-aware, while remaining grounded in domain knowledge and quantitative evidence. The convergence of high Likert-scale scores and constructive qualitative feedback across two independent judges substantiates the robustness and trustworthiness of the proposed MLLM-based interpretability method for EV powertrain fault diagnosis.

5.7 User Study

To complement the MLLM-as-a-Judge evaluation and provide additional human-centered validation, a user study was conducted to assess the quality of the explanations generated by the proposed MLLM-based agents. A total of nine participants evaluated the generated explanations using a five-point Likert scale and additionally provided qualitative feedback through open-ended questions.

This evaluation was conducted as part of an NRF-funded research project. In accordance with institutional guidelines, the study was reviewed and determined to be exempt from full IRB review by the Institutional Review Board of Duksung Women's University, as it involved minimal risk and did not include any clinical, biomedical, or behavioral manipulation.

Table 15 summarizes the quantitative user study results across five evaluation criteria: content accuracy, explanation quality, reasoning depth, consistency, and reference alignment. Overall, the proposed MLLM-based agents achieved favorable scores across all criteria, with average values generally exceeding 4.0 on the five-point Likert scale. These results indicate that the generated explanations were generally perceived as accurate and well aligned with the corresponding PFI and LOFO analysis results.

Table 15: Quantitative user study results based on five-point Likert scale evaluation.

Criterion	MLLM-Based PFI Analysis Agent	MLLM-Based LOFO Analysis Agent	MLLM-Based Aggregation Agent	MLLM-Based Report Generation Agent
Content accuracy	4.33 ± 0.70	4.00 ± 0.70	4.22 ± 0.66	4.33 ± 0.70
Explanation quality	4.44 ± 0.53	4.33 ± 0.50	4.33 ± 0.70	4.22 ± 0.83
Reasoning depth	4.22 ± 0.67	4.11 ± 0.60	4.22 ± 0.44	4.33 ± 0.50
Consistency	4.33 ± 0.50	4.33 ± 0.50	4.22 ± 0.66	4.33 ± 0.70
Reference alignment	4.56 ± 0.53	4.22 ± 0.66	4.33 ± 0.70	4.22 ± 0.66

Among the evaluated agents, the MLLM-based report generation agent achieved relatively high scores in reasoning depth and consistency, while the aggregation agent received favorable evaluations for integrating complementary interpretability perspectives from the PFI and LOFO analyses into unified explanations.

The PFI and LOFO analysis agents also demonstrated stable performance across all criteria. Participants generally considered the generated explanations to effectively describe feature importance patterns and subsystem-related operational relevance. However, the LOFO analysis agent received slightly lower scores for reasoning depth, which may reflect the higher complexity of counterfactual necessity interpretation.

To further assess the reliability of the human-centered evaluation, inter-rater reliability and agreement analyses were additionally conducted. The intraclass correlation coefficient (ICC) was employed to measure the consistency of participant ratings [48]. The mathematical formulation of the ICC is provided in Eq. (5), where MS_B and MS_W denote the between-subject and within-subject mean squares, respectively, and k represents the number of raters. Higher ICC values indicate greater consistency among evaluators.

In addition, exact agreement and within-one-point agreement were calculated to quantify the degree of rating agreement among participants [49,50]. Exact agreement represents the proportion of evaluator pairs assigning identical ratings and is defined in Eq. (6). Within-one-point agreement represents the proportion of evaluator pairs whose ratings differ by no more than one point on the five-point Likert scale and is defined in Eq. (7). In Eqs. (6) and (7), N_{exact} denotes the number of evaluator pairs assigning identical scores, $N_{|d|\leq 1}$ denotes the number of evaluator pairs with score differences less than or equal to one, and N_{total} denotes the total number of evaluator pairs.

$$ICC = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W} \tag{5}$$

$$Exact\ agreement = \frac{N_{exact}}{N_{total}} \times 100 \tag{6}$$

$$Within\text{-}one\text{-}point\ agreement = \frac{N_{|d|\leq 1}}{N_{total}} \times 100 \tag{7}$$

Table 16 summarizes the inter-rater reliability and agreement results. The ICC values ranged from 0.07 to 0.43, indicating low-to-moderate consistency among participants. This may be attributed to the limited number of evaluated agents and the subjective nature of Likert-scale assignments. Nevertheless, the exact agreement values ranged from 0.46 to 0.50, while the within-one-point agreement values ranged from 0.99 to 1.00. These results indicate that although participants did not always assign identical scores, their evaluations were highly consistent within a one-point difference on the five-point Likert scale. Therefore, the inter-rater reliability analysis provides additional statistical evidence supporting the stability and reliability of the user study results.

Table 16: Inter-rater reliability and agreement analysis of the user study.

Criterion	ICC	Exact Agreement	Within-One-Point Agreement
Content accuracy	0.29	0.50	1.00
Explanation quality	0.14	0.47	1.00
Reasoning depth	0.43	0.49	0.99
Consistency	0.24	0.47	1.00
Reference alignment	0.07	0.46	1.00

Table 17 presents representative qualitative feedback from the participants. Overall, participants highlighted that the proposed agents generated structured and domain-aware explanations while maintaining consistency with the feature importance evidence. In particular, the aggregation and report generation agents were positively evaluated for their coherent integration of multiple interpretability perspectives and publication-style organization.

Table 17: Qualitative feedback from the user study.

Agent	Strengths	Suggested Improvements
MLLM-based PFI analysis agent	Clearly identifies dominant and secondary features while providing intuitive explanations of permutation-based feature sensitivity patterns associated with EV powertrain operating conditions	Additional clarification regarding correlated variables, coupled subsystem behaviors, and the interpretation scope of perturbation-based importance may further improve interpretability
MLLM-based LOFO analysis agent	Effectively distinguishes indispensable and redundant variables through counterfactual reasoning and provides meaningful explanations of feature necessity after feature removal	More detailed explanation of retraining-based importance behavior and feature substitutability could improve readability and methodological understanding
MLLM-based aggregation agent	Successfully integrates complementary perspectives from PFI and LOFO analyses into a coherent and method-aware explanation while appropriately handling discrepancies between sensitivity and necessity interpretations	Inclusion of additional representative discrepancy cases and more detailed comparison examples between PFI and LOFO results may further strengthen interpretability
MLLM-based report generation agent	Produces well-structured, publication-style summaries with strong overall coherence and clear organization of subsystem-related diagnostic interpretations	Additional references to feature importance visualizations and simplified explanations for non-expert readers could improve accessibility and readability

Several participants additionally suggested that further clarification regarding correlated variables, subsystem interactions, and retraining-based feature importance behavior could improve interpretability and readability. Overall, the user study results support the practical usefulness of the proposed MLLM-based multi-agent method for EV powertrain fault diagnosis interpretability.

6 Discussions

6.1 Feature-to-Subsystem Relationships and Interpretation Scope

The EV powertrain fault diagnosis dataset employed in this study contains integrated monitoring variables collected under operating conditions involving multiple EV powertrain components, including the battery, inverter, motor, and surrounding environment [51–53]. Accordingly, the input variables should not be interpreted as isolated indicators of a single subsystem. Instead, they reflect operational characteristics that may be influenced by interactions among multiple electrical, mechanical, and thermal factors within the EV powertrains [2,4].

For this reason, the interpretation results generated by the proposed method are discussed from the perspective of subsystem-related operational relevance rather than direct physical causality [54]. Table 18 summarizes the primary subsystem associations of the variables used in this study. For example, voltage and current measurements reflect coupled electrical behaviors associated with power conversion, motor drive operation, and energy supply conditions [55]. In practical EV operating environments, voltage measurements capture electrical abnormalities related to battery cell imbalance, overcharge or over-discharge conditions and inverter power-conversion irregularities [56,57]. Similarly, vibration and motor speed variables capture mechanical operating characteristics related to the traction system [58,59]. In particular, vibration measurements may reflect bearing wear, rotor eccentricity, drivetrain oscillation, and other mechanical degradation phenomena. Temperature-related variables reflect broader thermal operating conditions across multiple subsystems. Accordingly, the explanations generated by the proposed method should be interpreted as feature importance-based inference describing model behavior patterns learned from the dataset rather than definitive evidence of direct physical fault causality within EV powertrain subsystems.

Table 18: Sensor-to-subsystem mapping of the EV powertrain variables.

Variable	Physical Subsystem Association	Engineering Interpretation
Voltage (V)	Battery, Inverter	Battery cell imbalance, overcharge/over-discharge behavior, transient voltage drops, DC-link instability, and inverter power-conversion abnormalities
Vibration (g)	Motor	Bearing wear, rotor eccentricity, drivetrain oscillation, structural resonance, and mechanical degradation
Temperature (°C)	Battery, Inverter, Motor	Thermal stress accumulation, overheating conditions, and abnormal heat generation
Current (A)	Battery, Inverter, Motor	Electrical load demand, torque generation, inverter–motor interaction, and abnormal current flow
Motor Speed (rpm)	Motor	Rotational operating condition and traction-system behavior
Humidity (%)	Environment	External environmental condition
Ambient Temperature (°C)	Environment	External thermal condition

The quantitative feature importance results generated by both the PFI and LOFO analyses are also consistent with the engineering interpretation of the monitored variables. According to both the PFI and LOFO analyses, voltage was identified as the most influential and indispensable feature (PFI mean = 0.42661, LOFO mean = 0.12297). The PFI results indicate that model performance is highly sensitive to voltage perturbation, whereas the LOFO results demonstrate that voltage-related information cannot be effectively replaced after feature removal and retraining. From an engineering perspective, this behavior is consistent

with voltage abnormalities associated with battery cell imbalance, overcharge or over-discharge conditions, and transient voltage drops caused by increased electrical resistance.

Similarly, vibration consistently ranked as the second most important feature across both analyses (PFI mean = 0.26234, LOFO mean = 0.09105). The strong importance and necessity of vibration suggest that mechanical condition information provides essential evidence for fault discrimination. This observation is consistent with practical motor fault scenarios in which bearing wear, rotor eccentricity, drivetrain oscillation, and structural resonance generate characteristic vibration signatures. Therefore, the agreement between the quantitative feature importance results and the corresponding engineering interpretation supports the practical relevance and physical plausibility of the generated explanations.

A contrasting pattern can be observed for current measurements. Current was identified as a globally influential feature in the PFI analysis (PFI mean = 0.14433), whereas the LOFO analysis indicated relatively low functional dependency after retraining (LOFO mean = 0.00857). This discrepancy suggests that part of the information provided by current measurements can be recovered through correlated voltage-related signals, reflecting the coupled electrical behavior of EV powertrain subsystems.

From an ML perspective, this behavior is also related to how random forest models handle correlated variables. Since random forest constructs multiple decision trees using randomly selected feature subsets, predictive information can be distributed across correlated variables rather than being assigned to a single feature. Consequently, correlated variables such as voltage and current may both appear important even when part of their information overlaps. To mitigate potential misinterpretation arising from such multicollinearity, the proposed method jointly analyzes PFI and LOFO results. As a result, features exhibiting high sensitivity but low necessity after retraining can be identified as redundant rather than independently indispensable. This enables the aggregation agent to distinguish correlated but substitutable variables from truly indispensable fault indicators, thereby improving the robustness of the generated interpretations.

In addition, the current dataset configuration focuses on subsystem-level fault classification rather than detailed intra-subsystem fault separation. For instance, the motor fault category includes multiple abnormal operating conditions, such as bearing wear and mechanical degradation, within a single aggregated class. Similarly, inverter faults and battery faults represent broader subsystem-level abnormal states rather than fine-grained fault subclasses. Therefore, the proposed method identifies subsystem-related abnormal operating patterns rather than distinguishing detailed physical fault mechanisms within the same subsystem.

Nevertheless, the proposed method still provides meaningful interpretability for EV powertrain fault diagnosis by identifying operational variables strongly associated with different fault categories while maintaining consistency with the scope and limitations of the employed dataset.

6.2 Computational Cost and Deployment Feasibility

The proposed method combines ML-based fault diagnosis with MLLM-based interpretation agents utilizing PFI and LOFO analyses. Since the method includes both feature importance computation and natural language interpretation generation, computational cost and deployment feasibility should be considered for practical real-time and large-scale applications [60,61].

From the model inference perspective, the underlying fault diagnosis classifier itself can operate efficiently in real-time environments because prediction for a single input instance requires only a forward inference step [19]. Therefore, subsystem-level fault classification for EV powertrain monitoring can be performed with relatively low latency once the trained model is deployed.

However, the computational overhead increases during the interpretation stage, particularly for PFI and LOFO analyses. PFI requires repeated feature permutation and model evaluation procedures for each input

variable, while LOFO requires retraining the model after removing individual features [8,9]. Consequently, LOFO analysis incurs higher computational costs than PFI because multiple retraining operations must be performed to estimate feature contributions. Furthermore, the MLLM-based interpretation agents introduce inference cost associated with multimodal prompt processing and natural language generation [62,63].

For this reason, the proposed method is more suitable for offline diagnostic analysis and post-hoc interpretability rather than continuous real-time explanation. Table 19 summarizes the computational characteristics of the major components of the proposed method, including their average runtime and suitability for real-time deployment. In practical EV deployment environments, the fault diagnosis model can support real-time monitoring with an average prediction time of 0.3 s, while the interpretability modules are activated only when abnormal operating conditions are detected. As summarized in Table 19, the PFI analysis requires an average of 7.3 s, making it suitable for event-triggered analysis, whereas the LOFO analysis requires 18.4 s because of repeated model retraining and is therefore more appropriate for offline diagnostic support. Similarly, the MLLM-based interpretation and aggregation modules require 3.6 and 3.1 s, respectively, enabling structured explanation generation after fault detection without interfering with continuous monitoring. This event-triggered deployment strategy balances computational efficiency and interpretability, making the proposed method suitable for diagnostic decision support and predictive maintenance applications.

Table 19: Computational characteristics of the proposed method.

Component	Operation	Average Runtime (s)	Real-Time Suitability
Fault diagnosis model	Single prediction	0.3	Suitable
PFI analysis	Feature permutation and repeated evaluation	7.3	Offline/Event-triggered
LOFO analysis	Feature removal and model retraining	18.4	Offline
MLLM interpretation agent	Prompt processing and explanation generation	3.6	Event-triggered
Aggregation and report generation	Text synthesis and validation	3.1	Event-triggered

In addition, the proposed method can be extended to large-scale deployment environments through optimization strategies. First, PFI and LOFO analyses can be performed periodically during model validation stages instead of continuously during online inference [64]. Second, lightweight MLLM architecture can reduce interpretation latency and memory usage [65]. Third, distributed computation can be applied to feature importance estimation because permutation and feature-removal operations are independently executable across variables [66,67]. Table 20 summarizes the hardware environment used in this study, under which the runtime measurements reported in Table 19 were obtained. These results demonstrate the feasibility of event-triggered deployment while indicating that further optimization of feature importance computation and MLLM inference can improve scalability for large-scale EV diagnostic platforms. Future research will focus on computational optimization and efficient deployment strategies while preserving explanation quality.

Table 20: Experimental hardware configuration.

Component	Specification
Central processing unit (CPU)	Intel Core i9-14900K
Graphics processing unit (GPU)	NVIDIA RTX 3080 Ti
Memory	64 GB RAM
Framework	Python 3.10.6, Scikit-Learn [68], GPT-5.2 [13]

6.3 Comparison with Exisitng Interpretability Methods

To further evaluate the proposed method, comparisons with widely used interpretability methods, including SHAP and local interpretable model-agnostic explanations (LIME), were conducted [69,70]. These methods explain model behavior by estimating feature contributions from different analytical perspectives.

SHAP provides feature attribution based on Shapley values and supports both global and local interpretation [71]. LIME explains individual predictions using local surrogate models generated around target instances [72]. Both approaches are used for visualizing feature importance and understanding model decision behavior.

In contrast, the proposed method combines PFI and LOFO analyses with MLLM-based interpretation agents to generate structured natural language explanations. PFI measures model sensitivity through feature permutation, whereas LOFO evaluates feature dependency by retraining the model after removing individual variables. By integrating these complementary approaches, the method captures both perturbation-based and feature-removal-based importance characteristics.

Fig. 7 presents visualization results obtained by applying SHAP and LIME to the EV powertrain fault diagnosis dataset used in this study, while Tables 21 and 22 summarize the corresponding feature attribution values generated by SHAP and LIME, respectively. SHAP and LIME provide numerical and visualization-based feature attribution results. However, these methods mainly focus on feature contribution visualization and do not inherently provide integrated natural language interpretation workflows. By comparison, the proposed method combines feature importance visualization with multimodal interpretation agents capable of generating human-readable diagnostic explanations. In addition, the combined use of PFI and LOFO enables complementary analysis of correlated variables such as voltage and current, which may exhibit different importance behaviors depending on feature interdependency.

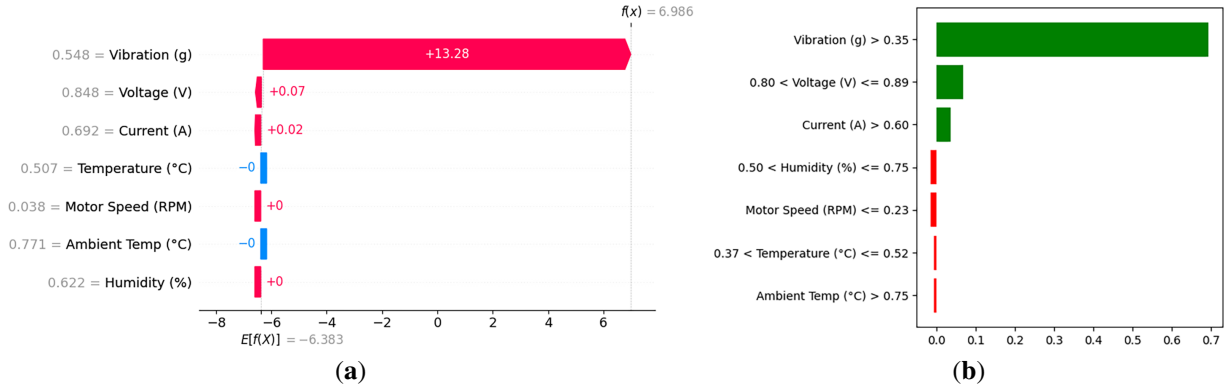

Figure 7: Visualization of SHAP and LIME analysis: (**a**) SHAP analysis; (**b**) LIME analysis.

Table 21: Feature attribution values generated by SHAP analysis.

Feature	Value	SHAP Value
Vibration (g)	0.5479	13.2790
Voltage (V)	0.8478	0.0688
Current (A)	0.6919	0.0213
Temperature (°C)	0.5066	−0.0006
Motor Speed (rpm)	0.0375	0.0002
Ambient Temperature (°C)	0.7706	−0.0000
Humidity (%)	0.6220	0.0000

Table 22: Feature attribution values generated by LIME analysis.

Feature	Value
Vibration (g) > 0.35	0.6950
0.80 < Voltage (V) <= 0.89	0.0673
Current (A) > 0.60	0.0345
0.50 < Humidity (%) <= 0.75	−0.0154
Motor Speed (rpm) <= 0.23	−0.0153
0.37 < Temperature (°C) <= 0.52	−0.0079
Ambient Temperature (°C) > 0.75	−0.0062

Nevertheless, the proposed method introduces higher computational cost due to repeated feature permutation, retraining procedures, and MLLM-based interpretation generation [66]. Therefore, the proposed method should be considered a complementary interpretability method emphasizing multimodal explanation generation rather than a direct replacement for existing XAI methods.

Future work will extend the proposed MLLM-based interpretation method to additional XAI methods such as SHAP and LIME, enabling unified multimodal interpretation and natural language reasoning across multiple feature attribution approaches for EV powertrain fault diagnosis models.

6.4 Ablation Study and Component-Wise Effectiveness Analysis

To further evaluate the proposed method, an ablation study was conducted to analyze the contribution of each MLLM-based agent.

The PFI interpretation agent provided global sensitivity-oriented explanations by translating permutation-based feature importance results into structured textual interpretations. However, it was limited in determining whether important variables were functionally indispensable or partially replaceable. The LOFO interpretation agent complemented this limitation by analyzing feature necessity through performance degradation after feature removal and retraining. This enabled clearer identification of indispensable and redundant variables, although sensitivity-oriented interpretation was comparatively limited.

The aggregation agent improved interpretive completeness by integrating PFI and LOFO-based explanations and identifying complementary relationships between sensitivity-based and counterfactual-necessity-based interpretations. In addition, the cross-validation agent enhanced explanation reliability by verifying consistency and domain validity of upstream interpretation outputs. Finally, the final report

generation agent consolidated the outputs of multiple agents into a unified interpretability report, improving overall readability.

Overall, the ablation study demonstrates that each agent contributes a distinct role within the proposed method. The PFI and LOFO agents provide complementary interpretive perspectives, while the aggregation and cross-validation agents improve explanation consistency. The final report generation agent enhances usability by integrating distributed interpretation outputs into a coherent domain-aware explanation.

Nevertheless, the current ablation study mainly focuses on qualitative component-wise analysis rather than exhaustive quantitative benchmarking. Future work will investigate quantitative evaluation protocols and optimized multi-agent coordination strategies for EV powertrain fault diagnosis interpretability systems [73].

6.5 Generalizability to Other Industrial Fault Diagnosis Domains

Although the proposed method was developed and evaluated using an EV powertrain fault diagnosis dataset, its overall architecture is not inherently restricted to EV applications. The proposed method combines model-agnostic feature importance techniques with a role-specialized MLLM-based multi-agent reasoning method. Since both PFI and LOFO can be applied to a wide range of ML models and industrial datasets, the proposed method can be extended to other fault diagnosis domains without modification of the underlying methodology.

In particular, many industrial monitoring systems rely on heterogeneous sensor measurements and ML-based diagnostic models similar to those used in EV powertrains [74,75]. Examples include rotating machinery monitoring, manufacturing process fault detection, and predictive maintenance of industrial products [76–78]. For example, in bearing fault diagnosis for rotating machinery, vibration, temperature, and rotational speed measurements can be analyzed using PFI and LOFO to identify globally influential and indispensable variables. The resulting feature importance can then be interpreted by the MLLM-based agents to generate structured maintenance reports describing possible bearing degradation, shaft misalignment, or rotor imbalance. This example demonstrates that the proposed method can translate quantitative feature importance into domain-aware textual explanations while preserving the same multi-agent architecture across different industrial fault diagnosis applications.

Nevertheless, domain adaptation remains an important consideration [79]. While the overall multi-agent architecture is transferable, domain-specific terminology and fault mechanisms should be appropriately incorporated into the prompt specifications and interpretation guidelines of the MLLM-based agents. Therefore, additional validation studies are required to evaluate the effectiveness and robustness of the proposed method across diverse industrial environments.

Overall, the proposed method can be viewed as a generalizable MLLM-based interpretability architecture for ML-driven fault diagnosis rather than a solution exclusively designed for EV powertrain systems. Future research will investigate its applicability to broader industrial prognostics and health management (PHM) scenarios involving different sensor modalities and operational conditions.

7 Conclusion

In this study, we proposed an ML-based fault diagnosis method for EV powertrains that integrates PFI, LOFO, and a collaborative MLLM-based agent architecture to achieve both accurate fault classification and interpretable diagnostic reasoning. A comprehensive comparison of multiple ML classifiers demonstrated that the random forest provides superior diagnostic performance in terms of f1-score and overall accuracy, making it well suited for EV powertrain fault diagnosis.

Building on this model, global feature relevance was analyzed using both PFI and LOFO, providing complementary perspectives on model behavior in terms of sensitivity and counterfactual necessity. These quantitative interpretability results were processed by an MLLM-driven multi-agent pipeline based on OpenAI GPT-5.2, which automatically generated structured feature-level interpretations, reconciled method-specific discrepancies, validated explanation consistency, and produced a unified diagnostic report. Through this end-to-end design, the proposed method bridges numerical importance measures and coherent, domain-aware natural language explanations, enabling a fully automated and explainable fault diagnosis process.

Experimental evaluation using an independent MLLM-as-a-Judge method confirms that the generated explanations exhibit high content accuracy, reasoning depth, and logical consistency across PFI, LOFO, aggregation, and final reporting stages. These results demonstrate that the proposed method not only enhances predictive performance but also improves interpretability reliability, reducing the variability and subjectivity typically associated with expert-driven manual analysis.

Despite these strengths, several limitations remain. First, the current study does not address domain shift arising from long-term temporal dynamics, seasonal environmental variations, or progressive sensor degradation under real-world EV operating conditions [80,81]. As the proposed method is evaluated on a static dataset, its robustness to evolving data distributions has not yet been examined. Future work will therefore extend the proposed method to long-term, real-driving datasets and operational logs, and incorporate online learning and self-adaptive agent mechanisms to support real-time PHM in continuously changing environments [82,83].

In addition, the present method focuses primarily on global interpretability through PFI and LOFO analyses. While this provides reliable insight into overall model behavior, it does not capture instance-level or time-localized decision rationales. To address this limitation, future research will integrate local explanation techniques, such as LIME [84], as well as time-series-specific interpretability methods [85,86]. By unifying global and local explanations within the proposed MLLM-based agent architecture, the method aims to evolve into a comprehensive, multi-scale reporting system capable of delivering both system-level understanding and case-specific diagnostic insight.

Through these extensions, the proposed MLLM-based agent method has the potential to further advance the automation, reliability, and practical deployment of XAI for safety-critical EV powertrain fault diagnosis.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00516023).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Jaeseung Lee and Jehyeok Rew; methodology, Jehyeok Rew; software, Jaeseung Lee; validation, Jehyeok Rew; formal analysis, Jehyeok Rew; investigation, Jaeseung Lee; resources, Jehyeok Rew; data curation, Jaeseung Lee; writing—original draft preparation, Jaeseung Lee; writing—review and editing, Jehyeok Rew; visualization, Jaeseung Lee; supervision, Jehyeok Rew; project administration, Jehyeok Rew; funding acquisition, Jehyeok Rew. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support of the findings of this study are openly available in New Energy Vehicles Diagnosis at https://www.kaggle.com/datasets/ziya07/fault-diagnosis-dataset-for-new-energy-vehicles.

Ethics Approval: This evaluation was conducted as part of an NRF-funded research project. In accordance with institutional guidelines, the study was reviewed and determined to be exempt from full IRB review by the Institutional Review Board of Duksung Women's University, as it involved minimal risk and did not include any clinical, biomedical, or behavioral manipulation. The assessment was entirely voluntary and non-interventional, focusing solely on participants' subjective ratings of the quality and interpretability of the generated outputs. No personally identifiable information (e.g., names, email addresses, or IP logs) was collected during the evaluation phase. However, to maintain verifiable compliance, informed consent was obtained electronically/written prior to participation, and all consent records were stored separately from the evaluation data to ensure complete anonymity during the subsequent analysis.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

BMS	Battery management system
CPU	Central processing unit
DC/AC	Direct current and alternating current
EV	Electric vehicle
GPU	Graphics processing unit
GUI	Graphical user interface
ICC	Intraclass correlation coefficient
iPFI	Incremental permutation feature importance
LIME	Local interpretable model-agnostic explanations
LLM	Large language model
LOFO	Leave-one-feature-out
ML	Machine learning
MLP	Multi-layer perceptron
MLLM	Multimodal large language model
PFI	Permutation feature importance
PHM	Prognostics and health management
ROC-AUC	Area under the receiver operating characteristic curve
rpm	Revolutions per minute
SHAP	Shapley additive explanations
XAI	Explainable artificial intelligence

Appendix A

Appendix A documents the complete prompt specifications used in the proposed multi-agent MLLM-based interpretability method. The prompts define the roles, inputs, objectives, structural constraints, and output formats of all agents involved in interpretation, aggregation, validation, and evaluation, thereby ensuring methodological transparency and reproducibility.

Tables A1 and A2 present the prompts for the MLLM-based PFI and LOFO interpretation agents, respectively. These prompts are designed to translate quantitative importance results into structured, domain-aware textual explanations, while explicitly distinguishing between PFI and LOFO. By enforcing method-faithful reasoning, explicit quantitative referencing, and table-oriented scientific writing, each agent remains aligned with the theoretical intent of its underlying XAI method.

Table A3 describes the prompt for the aggregation agent, which synthesizes the textual outputs of the PFI and LOFO agents to identify consistent patterns, method-consistent discrepancies, and complementary insights without reanalyzing raw importance values. Table A4 presents the prompt for the cross-validation analysis agent, which serves as an independent verification layer that evaluates individual agent outputs

for methodological consistency, logical coherence, domain appropriateness, and scientific rigor, without modifying the original interpretations.

Table A5 specifies the prompt for the final report generation agent, which consolidates PFI, LOFO, and aggregation outputs into a single interpretability report with an emphasis on global coherence and non-redundancy. Finally, Tables A6 and A7 provide the prompts for the MLLM-as-a-Judge evaluation agents, which assess explanation quality using five-point Likert-scale criteria and qualitative identification of strengths and weaknesses, while maintaining strict separation between generation and evaluation roles.

Table A1: Prompt for MLLM-based PFI interpretation agent.

Prompt

<Role Definition>
You are a multimodal large language model (MLLM) acting as an expert interpretation agent for explainable artificial intelligence (XAI) in electric vehicle (EV) powertrain fault diagnosis.
Your task is to generate a clear, accurate, and domain-aware textual interpretation of permutation feature importance (PFI) results derived from a machine learning–based fault diagnosis model.
</Role Definition>
<Inputs>
You are provided with the following multimodal inputs:
1. Permutation Feature Importance Visualization

- A horizontal bar chart showing the mean decrease in model performance when each input variable is permuted.
- Error bars represent variability across permutation repetitions.
- Higher values indicate greater importance for fault classification.

2. Input Variable Description

Variable	Unit	Description
Voltage	V	Measured voltage in the DC/AC power conversion stage
Current	A	Measured current flowing through the motor drive circuit
Motor Speed	rpm	Rotational speed of the traction motor
Temperature	°C	Operating temperature of the powertrain system
Vibration	g	Mechanical vibration level of the motor and drivetrain components
Ambient Temperature	°C	External environmental temperature
Humidity	%	Relative humidity of the operating environment

3. Output Classes

Class	Description
Normal operations	Stable and nominal powertrain conditions
Motor faults	Bearing wear, winding short, or mechanical degradation
Inverter faults	Power switch degradation or DC/AC conversion abnormalities
Battery faults	Cell imbalance, internal short circuit, or overcharge/overdischarge

</Inputs>
<Interpretation Objectives>
Generate a structured and table-oriented interpretation report that satisfies all of the following:

- Explicit Quantitative Referencing
  – Explicitly reference numerical PFI values (mean importance and relative magnitude) where relevant.
  – When possible, indicate approximate rank order (e.g., highest, second-highest) or relative scale differences.
- Hierarchical Structure Enforcement
  – Each section must be internally structured using numbered or bulleted subpoints.
  – Avoid single-paragraph narrative blocks.
  – Organize interpretations by feature groups (dominant/secondary/negligible).

(Continued)

Table A1 (continued)

Prompt
• Comparative Reasoning – Explicitly contrast high-importance features against low-importance and environmental variables. – Clearly distinguish dominant, secondary, and negligible features. • Table-Oriented Scientific Style – Write in a concise, analytical style suitable for a results table or appendix table (e.g., Table B1). – Prefer short, information-dense paragraphs over descriptive prose. </Interpretation Objectives> <Output Requirements> Generate a structured interpretation report using the following fixed section headings. IMPORTANT STRUCTURAL CONSTRAINTS: • Do NOT write long narrative paragraphs. • Each section MUST be internally subdivided using bullet points or numbered sub-sections. • Each dominant feature MUST be discussed in a separate bullet or sub-paragraph. • Quantitative references (e.g., importance mean, relative magnitude) MUST be included where applicable. 1. Overview of Global Feature Importance – Summarize the overall distribution using ranked groups: • Dominant features • Secondary features • Negligible features 2. Interpretation of Dominant Electrical and Mechanical Features – Use one sub-bullet per feature (e.g., Voltage, Vibration, Current). – For each feature, explicitly include: (a) Approximate PFI magnitude or rank (b) Physical meaning in EV powertrain context (c) Associated fault mechanisms 3. Analysis of Secondary and Low-Impact Variables – Separate discussion for: • Secondary contributors • Negligible/environmental variables – Explain redundancy, indirect influence, or environmental nature. 4. Stability and Reliability of PFI Results – Explicitly comment on variability using error bars. – Identify which features show stable vs. variable importance. 5. Implications for EV Powertrain Fault Diagnosis – Summarize implications using concise bullet points. – Emphasize sensor prioritization and diagnostic reliability. </Output Requirements>

Table A2: Prompt for MLLM-based LOFO interpretation agent.

Prompt
<Role Definition> You are a multimodal large language model (MLLM) acting as an expert interpretation agent for explainable artificial intelligence (XAI) in electric vehicle (EV) powertrain fault diagnosis. Your task is to generate a precise, domain-aware, and technically rigorous textual interpretation of Leave-One-Feature-Out (LOFO) importance results derived from a machine learning–based EV powertrain fault diagnosis model. </Role Definition> <Inputs> You are provided with the following multimodal inputs: 1. LOFO Importance Visualization

(Continued)

Table A2 (continued)

Prompt

- A horizontal bar chart showing the decrease in model performance when each input variable is removed from the training and test data.
- Error bars indicate variability across repeated model retraining.
- Larger LOFO values indicate stronger dependency of the model on the corresponding feature.

2. Input Variable Description

Variable	Unit	Description
Voltage	V	Measured voltage in the DC/AC power conversion stage
Current	A	Measured current flowing through the motor drive circuit
Motor Speed	rpm	Rotational speed of the traction motor
Temperature	°C	Operating temperature of the powertrain system
Vibration	g	Mechanical vibration level of the motor and drivetrain components
Ambient Temperature	°C	External environmental temperature
Humidity	%	Relative humidity of the operating environment

3. Output Classes

Class	Description
Normal operations	Stable and nominal powertrain conditions
Motor faults	Bearing wear, winding short, or mechanical degradation
Inverter faults	Power switch degradation or DC/AC conversion abnormalities
Battery faults	Cell imbalance, internal short circuit, or overcharge/overdischarge

</Inputs>
<Interpretation Objectives>
Generate a highly structured, table-oriented, and counterfactual interpretation report that satisfies the following:

- Explicit Counterfactual Quantification
  – Explicitly reference LOFO importance mean values and their relative magnitude.
  – Clearly indicate the severity of performance degradation caused by feature removal (e.g., severe, moderate, negligible).
- Functional Necessity Categorization
  – Explicitly categorize features into:
    (a) Functionally indispensable
    (b) Moderately necessary
    (c) Redundant or non-essential

(Continued)

Table A2 (continued)

Prompt
• Feature-Level Decomposition – Interpret each critical feature in a separate bullet or sub-section. – Avoid narrative-style paragraphs. • Table-Oriented Scientific Style – Write in a concise, analytical style suitable for interpreting LOFO results – Prioritize clarity and one-glance interpretability over prose. </Interpretation Objectives> <Output Requirements> Generate a structured interpretation report using the following fixed section headings. CRITICAL STRUCTURAL CONSTRAINTS: • Do NOT write long narrative paragraphs. • Each section MUST be internally subdivided using numbered or bulleted lists. • Each feature MUST be discussed in its own sub-item. • Numerical LOFO values (importance mean or relative magnitude) MUST be explicitly referenced. 1. Overview of Counterfactual Feature Importance – Summarize the overall LOFO distribution using ranked necessity groups: • Functionally indispensable features • Moderately necessary features • Redundant or non-essential features 2. Interpretation of Functionally Essential Electrical and Mechanical Features – Use one sub-section per feature (e.g., Voltage, Vibration, Temperature). – For each feature, explicitly include: (a) Approximate LOFO importance mean and rank (b) Severity of performance degradation upon removal (c) Functional role in EV powertrain fault discrimination 3. Analysis of Redundant or Low-Dependency Variables – Separate discussion for: • Low-necessity electrical/mechanical variables • Environmental variables – Explicitly explain why their removal does not critically affect performance. 4. Stability and Robustness of LOFO Results – Explicitly comment on variability using error bars and retraining results. – Identify which features show stable vs. variable counterfactual dependency. 5. Implications for Reliable EV Powertrain Fault Diagnosis – Summarize implications using concise bullet points. – Emphasize indispensable sensors and deprioritized measurements. </Output Requirements>

Table A3: Prompt for MLLM-based aggregation of interpretability results agent.

Prompt
<Role Definition> You are a multimodal large language model (MLLM) acting as an aggregation and synthesis agent for explainable artificial intelligence (XAI) in electric vehicle (EV) powertrain fault diagnosis. Your role is not to reinterpret raw feature importance values or to introduce new analysis. Instead, you must systematically integrate and reason over the textual interpretation outputs generated by two upstream agents: • the MLLM-based PFI Interpretation Agent, and • the MLLM-based LOFO Interpretation Agent.

(Continued)

Table A3 (continued)

Prompt
Your task is to identify consistent patterns, meaningful discrepancies, and complementary insights between permutation-based global sensitivity analysis and counterfactual feature necessity analysis, and to synthesize them into a unified, domain-consistent interpretation. </Role Definition> <Inputs> You are provided with the following inputs: 1. PFI Interpretation Output ◦ A structured, table-oriented textual explanation generated by the MLLM-based PFI Interpretation Agent. ◦ Reflects global feature sensitivity based on permutation-induced performance degradation. 2. LOFO Interpretation Output ◦ A structured, table-oriented textual explanation generated by the MLLM-based LOFO Interpretation Agent. ◦ Reflects counterfactual feature necessity based on performance degradation after feature removal and model retraining. 3. Shared Domain Context (Implicit) ◦ EV powertrain fault diagnosis involving motor, inverter, and battery fault classes. ◦ Input variables include electrical, mechanical, thermal, and environmental measurements. ◦ Both interpretations are derived from the same dataset, model, and fault taxonomy. No raw plots, numerical recomputation, or additional data are provided at this stage. </Inputs> <Aggregation Objectives> Generate an integrated interpretation that satisfies all of the following objectives: 1. Consistency Identification ◦ Identify features that are emphasized as important in both PFI and LOFO interpretations. ◦ Explain what this agreement implies about feature robustness and diagnostic reliability. 2. Discrepancy Analysis ◦ Identify features whose importance differs between PFI and LOFO. ◦ Explain these differences in terms of: ■ global sensitivity vs. counterfactual necessity, ■ redundancy, ■ substitutability, ■ or indirect influence. 3. Complementary Perspective Synthesis ◦ Explicitly articulate how PFI and LOFO answer different interpretability questions. ◦ Demonstrate how combining both perspectives yields deeper insight than either method alone. 4. Domain-Grounded Reasoning ◦ Ensure all explanations remain grounded in EV powertrain physics and fault mechanisms. ◦ Avoid speculative causal claims or unsupported physical interpretations. 5. Interpretability Implications ◦ Clarify how the aggregated interpretation improves model transparency, trustworthiness, and practical diagnostic actionability. Do not introduce new variables, numerical values, rankings, or assumptions beyond what is stated in the provided interpretations. </Aggregation Objectives> <Output Requirements>

(Continued)

Table A3 (continued)

Prompt
Generate a structured aggregation report using the following fixed section headings. Each section must be written in formal academic English and internally structured using concise paragraphs or bullet points. 1. Overview of Integrated Interpretation Strategy • Briefly describe the purpose of aggregating PFI and LOFO interpretations. • Clarify the complementary roles of permutation-based sensitivity and counterfactual necessity analysis. 2. Consistent Feature Importance Across PFI and LOFO • Identify features consistently highlighted by both interpretation agents. • Explain why these features can be regarded as robust and reliable indicators for EV powertrain fault diagnosis. 3. Discrepancies Between PFI and LOFO Interpretations • Identify features with divergent importance patterns across the two methods. • Provide method-aware explanations for these discrepancies. 4. Complementary Insights from Global and Counterfactual Perspectives • Explain how PFI captures sensitivity to feature perturbations and LOFO captures functional indispensability. • Demonstrate how integrating both perspectives refines understanding of feature roles. 5. Implications for Explainable and Reliable EV Powertrain Fault Diagnosis • Discuss how the aggregated interpretation enhances diagnostic confidence. • Highlight implications for sensor prioritization, monitoring strategy design, and maintenance decision support. </Output Requirements> <Important Constraints> • Base all reasoning strictly on the provided PFI and LOFO interpretation texts. • Do not re-rank features numerically or reinterpret raw plots. • Avoid informal language, free-form narrative, or unsupported causal inference. • Maintain a neutral, analytical, and publication-ready tone. </Important Constraints>

Table A4: Prompt for MLLM-based cross-validation analysis agent.

Prompt
<Role Definition> You are a multimodal large language model (MLLM) acting as an independent cross-validation and verification agent for explainable artificial intelligence (XAI) in electric vehicle (EV) powertrain fault diagnosis. Your role is to critically validate the interpretation output generated by one upstream MLLM-based interpretability agent at a time, without generating new interpretations or extending the original analysis. You function as a verification layer, ensuring that the provided explanation is: • logically consistent, • evidence-aligned, • domain-appropriate, • and free from over-interpretation or unsupported claims. </Role Definition> <Scope of Validation> You will receive the output from one of the following agents (only one at a time): • MLLM-Based PFI Interpretation Agent • MLLM-Based LOFO Interpretation Agent • MLLM-Based Aggregation of Interpretability Results Agent

(Continued)

Table A4 (continued)

Prompt
Your task is to validate only the given input, not to compare across agents. </Scope of Validation> <Input> You are provided with: 1. Agent Type Identifier (e.g., "PFI Interpretation Agent Output") 2. Textual Interpretation Output ○ A structured, domain-aware explanation generated by the specified agent ○ Based on prior numerical or visual XAI results (PFI, LOFO, or aggregated analysis) No raw data, plots, or tables are provided at this stage. </Input> <Validation Objectives> Evaluate the provided interpretation strictly according to the following dimensions: 1. Factual and Methodological Consistency ○ Does the explanation remain faithful to the stated purpose of the method (PFI, LOFO, or aggregation)? ○ Are methodological roles correctly represented (e.g., sensitivity vs. necessity)? 2. Logical Coherence ○ Are there internal contradictions, circular reasoning, or unsupported transitions? ○ Is the reasoning progression clear and stable? 3. Domain Appropriateness ○ Are interpretations aligned with known EV powertrain behavior? ○ Are physical or operational explanations plausible and non-speculative? 4. Evidence Alignment ○ Are claims framed proportionally to what the method can support? ○ Is there any overstatement beyond what importance-based analysis justifies? 5. Language and Scientific Rigor ○ Is the tone academic, neutral, and precise? ○ Are vague, ambiguous, or subjective expressions avoided? </Validation Objectives> <Output Requirements> Produce a structured validation report using the following fixed section headings. Each section must be written in formal academic English. 1. Validation Scope and Agent Context • Briefly restate which agent's output is being validated. • Clarify the intended role of that agent in the interpretability pipeline. 2. Assessment of Logical and Methodological Consistency • Evaluate whether the interpretation correctly reflects the methodological intent. • Identify any conceptual mismatches or inconsistencies, if present. 3. Evaluation of Domain Alignment and Interpretive Appropriateness • Assess whether domain-related explanations are reasonable and restrained. • Comment on the suitability of physical or operational interpretations. 4. Identification of Potential Issues or Over-Interpretation • Explicitly state whether: ○ overgeneralization, ○ unsupported causal claims, ○ or speculative reasoning is observed. • If none are found, clearly state that the interpretation remains appropriately bounded.

(Continued)

Table A4 (continued)

Prompt

5. Overall Validation Judgment
• Provide a concise verdict using one of the following forms:
  ○ "The interpretation is validated without major issues."
  ○ "The interpretation is generally sound but requires minor clarification."
  ○ "The interpretation exhibits notable inconsistencies that should be revised."
Do not rewrite or correct the original explanation.
Do not introduce new interpretations or recommendations beyond validation.
</Output Requirements>
<Important Constraints>
• Do not compare this output with other agents' results.
• Do not infer numerical values or visual trends not explicitly stated.
• Do not generate alternative explanations.
• Remain strictly within the role of a verification agent.
</Important Constraints>

Table A5: Prompt for MLLM-based final report generation agent.

Prompt

<Role Definition>
You are a multimodal large language model (MLLM) acting as a final report generation agent for explainable artificial intelligence (XAI) in electric vehicle (EV) powertrain fault diagnosis.
Your role is to produce a single, publication-ready interpretability report by synthesizing and refining the textual outputs generated by three upstream agents:
1. MLLM-based PFI Interpretation Agent Output
2. MLLM-based LOFO Interpretation Agent Output
3. MLLM-based Aggregation Agent Output
You do not perform new analyses, do not reinterpret raw plots, and do not introduce new evidence.
Your task is high-level narrative integration: consolidate the three inputs into a coherent, non-redundant, domain-consistent report.
</Role Definition>
<Inputs>
You are provided with the following textual inputs only:
1. PFI Interpretation Output
  ○ Structured summary of global sensitivity based on permutation-induced performance degradation.
  ○ Includes dominant/secondary/negligible feature grouping and stability notes (error bars/variability).
2. LOFO Interpretation Output
  ○ Structured summary of counterfactual feature necessity based on performance degradation after feature removal and retraining.
  ○ Includes indispensability grouping and robustness notes (retraining variability).
3. Aggregation Output
  ○ Integrated comparison identifying:
  ■ consistently important features,
  ■ discrepancies between PFI vs LOFO,
  ■ complementary interpretation,
  ■ practical implications.

(Continued)

Table A5 (continued)

Prompt
No additional tables, plots, or numerical recomputation are provided. </Inputs> <Final Report Objectives> Generate a final interpretability report that satisfies all objectives below: 1. Global Coherence and Non-Redundancy • Merge overlapping statements across the results of the preceding agents into one consistent narrative. • Avoid repeating the same claim in multiple sections unless necessary for logical continuity. 2. Methodological Complementarity (PFI vs LOFO) • Explicitly articulate how PFI captures global sensitivity and LOFO captures counterfactual necessity. • Use the aggregation result to explain why discrepancies (e.g., Current, Temperature) are method-consistent. 3. Evidence-Bounded and Domain-Consistent Reasoning • Ground interpretations in the EV powertrain context (motor, inverter, battery fault classes). • Do not add new causal claims, failure modes, or sensor mechanisms beyond what is stated in the inputs. 4. Actionable Interpretability Implications • Translate the integrated interpretation into implications for: ◦ sensor prioritization, ◦ monitoring strategy, ◦ diagnostic reliability and trust. 5. Scientific Writing Quality • Use formal academic English. • Avoid informal language. Maintain a neutral, analytical tone. </Final Report Objectives> <Output Requirements> Write the report with the following fixed section headings. Each section should be one coherent paragraph (no bullet points), unless explicitly allowed. 1. Integrated Overview of the Interpretability Pipeline ◦ Briefly describe the role of PFI, LOFO, and aggregation in the proposed workflow. 2. Global Sensitivity Findings from PFI Interpretation ◦ Summarize the dominant/secondary/negligible patterns. ◦ Mention stability qualitatively (e.g., low vs higher variability) only if stated. 3. Counterfactual Necessity Findings from LOFO Interpretation ◦ Summarize indispensable/moderate/redundant patterns. ◦ Mention robustness qualitatively only if stated. 4. Unified Interpretation of Consistencies and Discrepancies ◦ Use the result of aggregation agent to explain agreement (e.g., Voltage, Vibration, environmental variables) and divergence (e.g., Current, Temperature). ◦ Clearly attribute divergences to sensitivity vs necessity and redundancy/substitutability, as already stated. 5. Implications for Explainable and Reliable EV Powertrain Fault Diagnosis ◦ Provide integrated implications for trust, monitoring, and sensor prioritization. Formatting constraints • Do not include new numerical values beyond those already present in the inputs. • Do not introduce new variables or new rankings. • Do not cite external literature here. • Do not add figures/tables; output is text only. </Output Requirements>

Table A6: Prompt for MLLM-as-a-Judge of PFI and LOFO interpretation analysis agents.

Prompt
<Role Definition> You are a multimodal large language model (MLLM) acting as an independent evaluation agent (MLLM-as-a-Judge) for explainable artificial intelligence (XAI). Your role is to objectively evaluate the quality of a textual interpretation generated by an upstream MLLM-based interpretability agent (PFI or LOFO) for EV powertrain fault diagnosis. You must not generate new interpretations, revise the explanation, or introduce new analytical insights. Your task is strictly evaluation and critique. </Role Definition> <Evaluation Scope> You will evaluate one interpretation at a time, generated by one of the following agents: • MLLM-based PFI Interpretation Agent • MLLM-based LOFO Interpretation Agent The agent type will be explicitly provided. </Evaluation Scope> <Inputs> You are provided with the following inputs: 1. Agent Type Identifier (e.g., “PFI Interpretation Agent Output” or “LOFO Interpretation Agent Output”) 2. Interpretation Text ◦ A structured textual explanation generated by the specified agent ◦ Based on prior PFI or LOFO numerical and visual analysis 3. Reference Context (Implicit) ◦ EV powertrain fault diagnosis ◦ Variables include electrical, mechanical, thermal, and environmental signals ◦ No raw plots or numerical recalculation is allowed at this stage </Inputs> <Evaluation Criteria and Scoring> Evaluate the provided interpretation using a 5-point Likert scale (1 = very poor, 5 = excellent) for each of the following five criteria. 1. Content Accuracy • Does the explanation faithfully reflect: ◦ overall trends, ◦ relative importance, ◦ and directional influence observed in the underlying PFI or LOFO analysis? 2. Explanation Quality • Is the explanation: ◦ logically structured, ◦ clear and fluent, ◦ and written with appropriate technical terminology? 3. Reasoning Depth • Does the explanation go beyond surface-level descriptions? • Does it incorporate: ◦ relative impact magnitudes, ◦ stability or variability, ◦ and links to physical or operational meaning (where appropriate)?

(Continued)

Table A6 (continued)

Prompt

4. Consistency
• Are there any internal contradictions, inconsistencies, or unstable reasoning patterns?
• Is the interpretative narrative coherent as a whole?
5. Reference Alignment
• Do statements about increases/decreases or importance direction align with:
  ○ the PFI or LOFO trends implied by the analysis?
• Are claims proportional to what the method can support?
</Evaluation Criteria and Scoring>
<Qualitative Assessment (Open-Ended)>
After numerical scoring, provide brief qualitative feedback:
• Key Strengths
  ○ Identify 1–3 notable strengths of the interpretation
  ○ Focus on clarity, faithfulness, or analytical rigor
• Key Weaknesses or Limitations
  ○ Identify any issues such as:
  ■ vague phrasing,
  ■ overgeneralization,
  ■ insufficient grounding,
  ■ or missed opportunities for clearer explanation
  ○ If no major weaknesses exist, explicitly state this
</Qualitative Assessment (Open-Ended>
<Output Format>
Produce the evaluation using the following fixed structure:
Evaluation Summary
Criterion — Score (1–5)
Content Accuracy — X
Explanation Quality — X
Reasoning Depth — X
Consistency — X
Reference Alignment — X
Strengths
• Bullet-point list (concise, evidence-based)
Weaknesses/Areas for Improvement
• Bullet-point list
• If none, state: "No significant weaknesses were identified."
Overall Judgment
Overall judgement with one paragraph
</Output Format>
<Important Constraints>
• Do not rewrite or correct the interpretation.
• Do not introduce new interpretations, variables, or numerical values.
• Do not compare this output with other agents.
• Maintain a neutral, academic, and evaluation-focused tone.
</Important Constraints>

Table A7: Prompt for MLLM-as-a-Judge of aggregation and final report generation agents.

Prompt
<Role Definition> You are a multimodal large language model (MLLM) acting as an independent evaluation agent (MLLM-as-a-Judge) for explainable artificial intelligence (XAI). Your role is to critically evaluate the quality of an integrated interpretation or final diagnostic report generated by an upstream MLLM-based agent for EV powertrain fault diagnosis. You must not: • generate new interpretations, • modify the content, • introduce additional evidence or assumptions. Your task is strictly evaluation and qualitative critique. </Role Definition> <Evaluation Scope> You will evaluate one output at a time, generated by one of the following agents: • MLLM-Based Aggregation of Interpretability Results Agent, or • MLLM-Based Final Report Generation Agent The agent type will be explicitly specified. </Evaluation Scope> <Inputs> You are provided with: 1. Agent Type Identifier (e.g., “Aggregation Agent Output” or “Final Report Generation Agent Output”) 2. Textual Interpretation Output ◦ An integrated or final report synthesizing prior PFI and LOFO interpretations ◦ No raw plots, tables, or numerical recomputation are provided 3. Implicit Reference Context ◦ EV powertrain fault diagnosis (motor, inverter, battery) ◦ Electrical, mechanical, thermal, and environmental variables ◦ Upstream interpretations are assumed to be method-faithful </Inputs> <Evaluation Criteria and Scoring> Evaluate the provided output using a 5-point Likert scale (1 = very poor, 5 = excellent) for each of the following five criteria. 1. Content Accuracy • Does the integrated explanation: ◦ correctly reflect the conclusions of the underlying PFI and LOFO interpretations? • Are agreements and discrepancies summarized faithfully without distortion or omission? 2. Explanation Quality • Is the explanation: ◦ logically organized, ◦ clearly written, ◦ and free from redundancy or unnecessary repetition? • Is the technical language appropriate for a scientific audience?

(Continued)

Table A7 (continued)

Prompt	

3. Reasoning Depth
• Does the report:
  ○ meaningfully synthesize sensitivity (PFI) and necessity (LOFO) perspectives?
  ○ clearly explain why consistencies or discrepancies arise (e.g., redundancy, substitutability)?
• Does it go beyond listing results to provide integrative reasoning?
4. Consistency
• Is the narrative:
  ○ internally coherent,
  ○ free from contradictions across sections,
  ○ and stable in its interpretation of feature roles?
5. Reference Alignment
• Are statements about:
  ○ feature importance,
  ○ sensitivity vs. necessity,
  ○ and diagnostic implications
  aligned with what PFI, LOFO, and aggregation can legitimately support?
• Is there any overstatement beyond evidence-based interpretability?
</Evaluation Criteria and Scoring>
<Qualitative Assessment (Open-Ended)>
After numerical scoring, provide concise qualitative feedback.
Key Strengths
• Identify 1–3 strengths, such as:
  ○ effective synthesis,
  ○ clear distinction between methods,
  ○ strong domain grounding,
  ○ or publication-ready clarity.
Key Weaknesses or Limitations
• Identify any limitations, such as:
  ○ redundancy,
  ○ insufficient integration,
  ○ vague implications,
  ○ or overly cautious/overly strong claims.
• If no major weaknesses are observed, explicitly state this.
</Qualitative Assessment (Open-Ended>
<Output Format>
Produce the evaluation using the following fixed structure:
Evaluation Summary
Criterion — Score (1–5)
Content Accuracy — X
Explanation Quality — X	

(Continued)

Table A7 (continued)

Prompt	
Reasoning Depth	X
Consistency	X
Reference Alignment	X
Strengths	
• Bullet-point list (concise, evidence-based)	
Weaknesses/Areas for Improvement	
• Bullet-point list • If none, state: "No significant weaknesses were identified."	
Overall Judgment	
Overall judgement with one paragraph	
</Output Format>	
<Important Constraints>	
• Do not compare this output with other agents' outputs. • Do not rewrite, summarize, or improve the content. • Do not introduce new interpretations, features, or rankings. • Maintain a neutral, academic, and evaluation-focused tone.	
</Important Constraints>	

In particular, the prompts specify role-driven behavioral constraints, domain-grounding requirements, and non-generative verification objectives to ensure methodological consistency across all MLLM-based agents. Overall, Appendix A operationalizes a fully specified, multi-stage MLLM-based interpretability pipeline, providing an extensible foundation for MLLM-driven XAI in EV powertrain fault diagnosis.

Appendix B

Appendix B presents representative outputs generated by each agent in the proposed multi-agent MLLM-based interpretability method, illustrating how quantitative explainability results are transformed into structured interpretations.

Tables A8 and A9 show the textual interpretations produced by the MLLM-based PFI and LOFO interpretation agents, respectively. These results demonstrate how global sensitivity and counterfactual necessity results are translated into domain-aware explanations reflecting feature importance, stability, and relevance to EV powertrain fault mechanisms.

Table A10 reports the output of the MLLM-based aggregation agent, which integrates the PFI and LOFO interpretations into a unified, method-aware explanation by identifying consistent attributions, method-consistent discrepancies, and complementary insights without reanalyzing raw importance values. Table A11 presents the final interpretability report generated by the MLLM-based report generation agent, consolidating all prior outputs into a coherent, non-redundant narrative suitable for expert analysis and practical EV powertrain fault diagnosis.

Overall, Appendix B provides concrete examples of intermediate and final outputs produced by the proposed method, supporting transparency and qualitative assessment of the MLLM-based interpretability process.

Table A8: Result of PFI analysis agent.

Result

1. Overview of Global Feature Importance
• Dominant features
  ◦ Voltage (V): Highest importance (mean PFI ≈ 0.427)
  ◦ Vibration (g): Second-highest importance (mean PFI ≈ 0.262)
  ◦ Current (A): Third-ranked contributor (mean PFI ≈ 0.144)
• Secondary features
  ◦ Temperature (°C): Moderate contribution (mean PFI ≈ 0.063)
  ◦ Motor Speed (rpm): Low but non-negligible contribution (mean PFI ≈ 0.021)
• Negligible features
  ◦ Humidity (%): Near-zero importance (mean PFI ≈ 0.0003)
  ◦ Ambient Temperature (°C): No measurable contribution (mean PFI ≈ 0.0000)
2. Interpretation of Dominant Electrical and Mechanical Features
• Voltage (V)
  ◦ (a) Quantitative rank: Highest-ranked feature, mean PFI ≈ 0.427, over 1.6× larger than the second-ranked feature.
  ◦ (b) Physical meaning: Represents DC/AC conversion stability and electrical supply integrity in the inverter stage.
  ◦ (c) Fault mechanisms: Strongly associated with inverter faults (switch degradation, conversion abnormalities) and battery faults (over/under-voltage, imbalance).
• Vibration (g)
  ◦ (a) Quantitative rank: Second-highest, mean PFI ≈ 0.262, accounting for ~61% of Voltage's importance.
  ◦ (b) Physical meaning: Captures mechanical oscillations from the motor and drivetrain.
  ◦ (c) Fault mechanisms: Indicative of motor faults, including bearing wear, rotor imbalance, and mechanical degradation.
• Current (A)
  ◦ (a) Quantitative rank: Third-ranked, mean PFI ≈ 0.144, clearly separated from secondary variables.
  ◦ (b) Physical meaning: Reflects electrical load and torque demand in the motor drive circuit.
  ◦ (c) Fault mechanisms: Linked to winding shorts, abnormal load conditions, and inverter–motor interaction faults.
3. Analysis of Secondary and Low-Impact Variables
• Secondary contributors
  ◦ Temperature (°C)
  ■ Mean PFI ≈ 0.063, substantially lower than electrical/mechanical signals.
  ■ Acts as an indirect indicator of prolonged stress rather than an immediate fault trigger.
  ◦ Motor Speed (rpm)
  ■ Mean PFI ≈ 0.021, suggesting contextual rather than causal influence.
  ■ Likely redundant given stronger electrical and vibration signals.
• Negligible/environmental variables
  ◦ Humidity (%)
  ■ Mean PFI ≈ 0.0003, effectively no contribution to fault discrimination.
  ◦ Ambient Temperature (°C)
  ■ Mean PFI ≈ 0.0000, including slightly negative values, indicating noise-level influence.
  ◦ These variables primarily reflect external operating conditions rather than internal powertrain states.

(Continued)

Table A8 (continued)

Result

4. Stability and Reliability of PFI Results
• Highly stable features
  ○ Voltage (V): Very low standard deviation (std ≈ 0.0032), consistent dominance across CV folds.
  ○ Vibration (g): Low variability (std ≈ 0.0049), indicating robust mechanical relevance.
• Moderately variable feature
  ○ Current (A): Higher variability (std ≈ 0.0204), suggesting sensitivity to operating regimes and fault types.
• Stable but insignificant features
  ○ Environmental variables show low variance but remain near zero importance, reinforcing their irrelevance.
5. Implications for EV Powertrain Fault Diagnosis
• Electrical and mechanical sensors (Voltage, Vibration, Current) should be prioritized for reliable fault detection.
• Voltage serves as the primary global discriminator, especially for inverter and battery-related faults.
• Mechanical vibration provides complementary fault evidence for motor degradation.
• Environmental sensors contribute minimal diagnostic value and can be deprioritized in model design.
• The low variability of dominant features supports robust and reproducible diagnostic decisions.

Table A9: Result of LOFO analysis agent.

Result

1. Overview of Counterfactual Feature Importance
• Functionally indispensable features
  ○ Voltage (V)—highest counterfactual dependency
  ○ Vibration (g)—second-highest counterfactual dependency
• Moderately necessary features
  ○ Temperature (°C)—intermediate performance dependency
• Redundant or non-essential features
  ○ Current (A)
  ○ Motor Speed (rpm)
  ○ Humidity (%)
  ○ Ambient Temperature (°C)
2. Interpretation of Functionally Essential Electrical and Mechanical Features
• Voltage (V)
  ○ (a) LOFO mean/rank: 0.12297, rank 1
  ○ (b) Performance degradation severity: Severe
  ○ (c) Functional role: Represents electrical stability of the DC/AC conversion stage and is indispensable for distinguishing inverter faults and battery-related abnormalities; feature removal leads to the largest loss in diagnostic accuracy.
• Vibration (g)
  ○ (a) LOFO mean/rank: 0.09105, rank 2
  ○ (b) Performance degradation severity: Severe–Moderate
  ○ (c) Functional role: Encodes mechanical condition of the traction motor and drivetrain, providing essential evidence for identifying motor faults such as bearing wear and mechanical degradation.
• Temperature (°C)
  ○ (a) LOFO mean/rank: 0.02628, rank 3
  ○ (b) Performance degradation severity: Moderate
  ○ (c) Functional role: Reflects accumulated thermal stress in the powertrain; contributes indirectly to fault discrimination by contextualizing electrical and mechanical anomalies.

(Continued)

Table A9 (continued)

Result

3. Analysis of Redundant or Low-Dependency Variables
• Low-necessity electrical/mechanical variables
  ○ Current (A)
  ■ LOFO mean ≈ 0.00857
  ■ Removal causes only minor performance degradation, indicating substantial redundancy with voltage-related information.
  ○ Motor Speed (rpm)
  ■ LOFO mean ≈ 0.00250
  ■ Primarily captures operating context rather than fault-specific behavior.
• Environmental variables
  ○ Humidity (%)
  ■ LOFO mean ≈ 0.00053
  ■ Negligible impact due to weak causal linkage to internal powertrain faults.
  ○ Ambient Temperature (°C)
  ■ LOFO mean ≈ −0.00017
  ■ Removal does not degrade performance, confirming non-essentiality.
4. Stability and Robustness of LOFO Results
• Highly stable dependencies
  ○ Voltage (V): low variability (std ≈ 0.00795), consistent degradation across all cross-validation folds.
  ○ Vibration (g): low variability (std ≈ 0.00546), robust mechanical dependency.
• Moderately stable dependency
  ○ Temperature (°C): moderate variability (std ≈ 0.00374), reflecting sensitivity to operational regimes.
• Stable but insignificant variables
  ○ Environmental features exhibit low mean impact despite minor fold-wise fluctuations.
5. Implications for Reliable EV Powertrain Fault Diagnosis
• Voltage and vibration sensors are counterfactually indispensable for reliable fault discrimination.
• Temperature measurements provide supportive but non-critical diagnostic information.
• Current and motor speed measurements can be deprioritized due to functional redundancy.
• Environmental variables do not meaningfully contribute to diagnostic performance.
• LOFO analysis confirms that removal of key electrical or mechanical signals results in substantial model performance degradation, validating their necessity for robust EV fault diagnosis.

Table A10: Result of aggregation agent.

Result
1. Overview of Integrated Interpretation Strategy This aggregation synthesizes two complementary interpretability perspectives derived from the same EV powertrain fault diagnosis model: Permutation Feature Importance (PFI) and Leave-One-Feature-Out (LOFO). The purpose is to consolidate a model-faithful account of which variables are globally influential (PFI: performance sensitivity to perturbation) and which variables are functionally indispensable (LOFO: performance dependency under feature removal and retraining). By jointly reasoning over both textual interpretations, the integrated report distinguishes (i) robust diagnostic signals consistently supported across methods from (ii) features whose apparent importance is method-contingent due to redundancy, substitutability, or indirect influence. 2. Consistent Feature Importance Across PFI and LOFO Both upstream agents converge on a small set of electrical and mechanical measurements as the most reliable drivers of diagnostic performance.

(Continued)

Table A10 (continued)

Result

- Voltage (V)—consistently dominant and indispensable
  - PFI identifies voltage as the strongest global discriminator with highly stable importance across folds, and links it to electrical integrity in the inverter stage and battery-related abnormalities.
  - LOFO independently confirms voltage as the most severe dependency: removing it causes the largest degradation, implying that voltage carries non-substitutable diagnostic information for distinguishing inverter and battery fault patterns.
  - Implication of agreement: Voltage can be treated as a robust, high-confidence indicator, supporting transparent and reproducible diagnosis for electrically driven fault classes.
- Vibration (g)—consistently high-impact and necessary
  - PFI ranks vibration second and interprets it as a global mechanical discriminator reflecting motor/drivetrain oscillations.
  - LOFO likewise classifies vibration as functionally essential, with severe–moderate degradation upon removal, supporting its indispensability for identifying motor-related degradation signatures.
  - Implication of agreement: Mechanical vibration provides robust complementary evidence to electrical variables, strengthening diagnostic reliability for motor fault modes without relying on speculative mechanisms beyond the provided interpretations.
- Environmental variables (Humidity, Ambient Temperature)—consistently negligible
  - PFI reports near-zero (or noise-level) importance and explicitly frames these as external operating conditions rather than internal powertrain states.
  - LOFO similarly finds negligible or even non-degrading effects upon removal, confirming non-essentiality.
  - Implication of agreement: Environmental measurements are neither sensitive nor necessary in the current model/dataset context, and thus contribute little to model transparency or actionable fault discrimination.

3. Discrepancies Between PFI and LOFO Interpretations
While the two methods agree on the top signals, they diverge meaningfully for several mid-tier variables, reflecting differences between global sensitivity (PFI) and counterfactual necessity (LOFO).

- Current (A): globally sensitive (PFI) but not necessary (LOFO)
  - PFI places current as a dominant third contributor and interprets it as reflecting electrical load/torque demand and inverter–motor interaction behavior.
  - LOFO categorizes current as redundant: removal causes only minor degradation and indicates substantial overlap with voltage-related information.
  - Method-aware explanation: Current can be informative under perturbation (PFI) because permuting it disrupts correlated electrical operating patterns that the model uses. However, LOFO suggests that when current is entirely removed and the model is retrained, other signals—especially voltage—can substitute for much of the predictive role, making current non-indispensable.
- Temperature (°C): moderate sensitivity (PFI) but relatively more necessary (LOFO)
  - PFI treats temperature as a secondary, indirect indicator of prolonged stress rather than an immediate trigger.
  - LOFO elevates temperature into a moderately necessary feature, indicating intermediate dependency and framing it as contextual thermal stress information that supports discrimination.
  - Method-aware explanation: Temperature may not dominate global sensitivity because its effect is supportive and context-setting rather than strongly discriminative in isolation. LOFO's moderate dependency implies that temperature provides non-trivial complementary context that cannot be fully recovered by remaining variables when the model is retrained, even if it is not among the most sensitive features under permutation.
- Motor Speed (rpm): low sensitivity (PFI) and low necessity (LOFO), with consistent "contextual" framing
  - Both methods treat speed as largely contextual rather than fault-specific, but LOFO more directly labels it non-essential, reinforcing the interpretation that speed is substitutable given stronger electrical/mechanical signals.

(Continued)

Table A10 (continued)

Result

4. Complementary Insights from Global and Counterfactual Perspectives
PFI and LOFO address different but jointly informative questions about model behavior:
• PFI (global sensitivity): "Which features, when perturbed, most degrade model performance?"
  ○ This perspective highlights variables the model uses strongly in its current representation and reveals which signals are influential under distribution-preserving disruption.
  ○ In this case, PFI emphasizes a hierarchy led by voltage and vibration, while also marking current as a strong sensitivity contributor.
• LOFO (counterfactual necessity): "Which features are indispensable for the model to perform well when that feature is removed and the model is retrained?"
  ○ This perspective isolates features whose information content is not recoverable from remaining variables, clarifying redundancy/substitutability.
  ○ Here, LOFO confirms voltage and vibration as non-substitutable, while showing that current and speed are largely replaceable.
• Integrated refinement of feature roles
  ○ Voltage and vibration emerge as both highly sensitive and indispensable, supporting their interpretation as the most robust and reliable diagnostic anchors.
  ○ Current emerges as sensitive but substitutable, implying it may strengthen performance and interpretability when present, but is not strictly required for reliable diagnosis in the presence of voltage.
  ○ Temperature emerges as supportive and moderately necessary, clarifying that its value lies in contextualizing anomalies rather than serving as a primary discriminator.
5. Implications for Explainable and Reliable EV Powertrain Fault Diagnosis
The aggregated interpretation strengthens transparency and practical actionability by separating robust core signals from contextual or redundant variables.
• Enhanced diagnostic confidence and trustworthiness
  ○ Agreement across PFI and LOFO for voltage and vibration indicates these features provide stable, reproducible diagnostic evidence, improving confidence that the model relies on physically meaningful electrical and mechanical measurements rather than incidental correlates.
  ○ Consistent insignificance of environmental variables reduces the likelihood that decisions are driven by external operating conditions, supporting more credible fault-specific reasoning.
• Sensor prioritization and monitoring strategy
  ○ Prioritize voltage and vibration as the highest-value sensing channels because they are simultaneously globally influential and counterfactually indispensable.
  ○ Treat temperature as a supportive channel: not the primary driver, but useful for contextual reliability and interpretive completeness, consistent with both agents' "indirect stress/contextualization" characterization.
  ○ Deprioritize current and motor speed for indispensability (per LOFO), while recognizing that current can still contribute to performance sensitivity (per PFI). Practically, this supports a nuanced interpretation: current may be valuable for refinement and operational detail but is not foundational to the model's core discriminative capacity.
  ○ Deprioritize environmental sensing for fault discrimination in this modeling setup due to consistent near-zero impact.
• Maintenance decision support and interpretability
  ○ By combining sensitivity (PFI) and necessity (LOFO), the report offers a clearer basis for decision support: core signals (voltage, vibration) can be emphasized in alerts and diagnostic narratives, while secondary/supportive signals (temperature) can be used to contextualize severity or persistence, and redundant/non-essential signals can be down-weighted to avoid over-attribution.
  ○ This dual-perspective explanation improves interpretability by explicitly distinguishing "important because the model uses it strongly" from "important because the model cannot function without it," thereby supporting more reliable, model-faithful diagnostic reporting.

Table A11: Result of final report generation agent.

Result
1. Integrated Overview of the Interpretability Pipeline The interpretability pipeline integrates three complementary stages—Permutation Feature Importance (PFI), Leave-One-Feature-Out (LOFO), and a dedicated aggregation process—to produce a coherent and model-faithful explanation of an EV powertrain fault diagnosis system. PFI characterizes global sensitivity by quantifying performance degradation under feature perturbation, thereby revealing which variables the trained model relies on most strongly. LOFO complements this view by assessing counterfactual necessity through feature removal and retraining, identifying variables whose information content is indispensable for maintaining diagnostic performance. The aggregation stage systematically reconciles these perspectives, consolidating consistent findings, contextualizing discrepancies, and translating them into a unified interpretive narrative suitable for transparent and reliable fault diagnosis.
2. Global Sensitivity Findings from PFI Interpretation The PFI interpretation reveals a clear hierarchy of global feature influence dominated by electrical and mechanical signals. Voltage and vibration emerge as the most influential variables, followed by current as a distinct but lower-ranked contributor, indicating strong sensitivity of model performance to perturbations in these measurements. Temperature and motor speed exhibit secondary and contextual influence, while environmental variables such as humidity and ambient temperature show negligible impact on fault discrimination. Stability analysis further indicates that voltage and vibration maintain consistently low variability across validation folds, supporting their robustness as global discriminators, whereas current exhibits higher variability, suggesting sensitivity to operating regimes and fault types rather than uniformly stable influence.
3. Counterfactual Necessity Findings from LOFO Interpretation The LOFO interpretation reframes feature importance in terms of functional indispensability. Voltage and vibration are identified as counterfactually indispensable, with their removal causing severe to moderate performance degradation, confirming that the model cannot reliably distinguish key inverter, battery, and motor fault classes without these signals. Temperature demonstrates moderate necessity, contributing indirectly by contextualizing electrical and mechanical anomalies. In contrast, current, motor speed, and environmental variables are categorized as redundant or non-essential, as their removal results in only minor or negligible degradation. Robustness analysis shows that indispensable features maintain stable dependency across retraining folds, reinforcing the reliability of these necessity-based conclusions.
4. Unified Interpretation of Consistencies and Discrepancies The aggregated interpretation highlights strong agreement between PFI and LOFO for voltage and vibration, establishing them as both globally influential and counterfactually indispensable, and thus as the most robust diagnostic anchors in the model. Environmental variables are consistently deemed negligible by both methods, indicating neither sensitivity nor necessity and reinforcing their limited relevance to internal powertrain fault states. Divergences arise for mid-tier features and are method-consistent rather than contradictory. Current exhibits high global sensitivity under PFI but low necessity under LOFO, reflecting redundancy and substitutability with voltage-related information: perturbation disrupts learned correlations, yet removal allows other signals to compensate after retraining. Temperature shows modest global sensitivity but greater counterfactual necessity, consistent with a supportive, context-setting role that is not strongly discriminative in isolation but not fully recoverable when removed. Motor speed remains contextual and non-essential under both perspectives, underscoring its substitutable nature given stronger electrical and mechanical signals.
5. Implications for Explainable and Reliable EV Powertrain Fault Diagnosis By jointly considering global sensitivity and counterfactual necessity, the final interpretation enhances model transparency, trustworthiness, and practical actionability. The consistent prominence of voltage and vibration supports prioritizing these sensors as the core evidence base for reliable diagnosis of inverter, battery, and motor faults, while their stability across methods increases confidence that model decisions are grounded in physically meaningful signals rather than incidental correlations. Temperature measurements can be retained as supportive context to enrich interpretability and reliability without being overemphasized. Current and motor speed may contribute to refinement and operational detail but need not be treated as indispensable, and environmental variables can be deprioritized for fault discrimination. Overall, the integrated perspective clarifies not only which features matter, but why they matter, enabling more faithful diagnostic narratives, informed sensor prioritization, and robust decision support in EV powertrain monitoring systems.

References

1. Thirunavukkarasu S, Karthick K, Aruna SK, Manikandan R, Safran M. Optimized fault classification in electric vehicle drive motors using advanced machine learning and data transformation techniques. Processes. 2024;12(12):2648. doi:10.3390/pr12122648.
2. Wu C, Sehab R, Akrad A, Morel C. Fault diagnosis methods and fault tolerant control strategies for the electric vehicle powertrains. Energies. 2022;15(13):4840. doi:10.3390/en15134840.
3. Khaneghah MZ, Alzayed M, Chaoui H. Fault detection and diagnosis of the electric motor drive and battery system of electric vehicles. Machines. 2023;11(7):713. doi:10.3390/machines11070713.
4. Dettinger F, Jazdi N, Weyrich M, Brandl L, Reuss HC, Pecha U, et al. Machine-learning-based fault detection in electric vehicle powertrains using a digital twin. In: Proceedings of the 23rd Stuttgart International Symposium; 2023 Jul 4; Stuttgart, Germany. doi:10.4271/2023-01-1214.
5. Hossain MN, Rahman MM, Ramasamy D. Artificial intelligence-driven vehicle fault diagnosis to revolutionize automotive maintenance: a review. Comput Model Eng Sci. 2024;141(2):951–96. doi:10.32604/cmes.2024.056022.
6. Mestha SR, Prabhu N. Support vector machine based fault detection in inverter-fed electric vehicle. Energy Storage. 2024;6(1):e576. doi:10.1002/est2.576.
7. Haque ME, Zabin M, Uddin J. EnsembleXAI-motor: a lightweight framework for fault classification in electric vehicle drive motors using feature selection, ensemble learning, and explainable AI. Machines. 2025;13(4):314. doi:10.3390/machines13040314.
8. Fumagalli F, Muschalik M, Hüllermeier E, Hammer B. Incremental permutation feature importance (iPFI): towards online explanations on data streams. Mach Learn. 2023;112(12):4863–903. doi:10.1007/s10994-023-06385-y.
9. Liu J, Danait N, Hu S, Sengupta S. A leave-one-feature-out wrapper method for feature selection in data classification. In: Proceedings of the 2013 6th International Conference on Biomedical Engineering and Informatics; 2013 Dec 16–18; Hangzhou, China. doi:10.1109/BMEI.2013.6747021.
10. Zhang C, Feng J, Cui C, Lin P, Chen H, Xu Y, et al. Electric vehicle user charging behavior analysis integrating psychological and environmental factors: a statistical-driven LLM based agent approach. arXiv:2408.05233. 2024.
11. Lu Y, Shi L. An enhanced explainable large language model-based framework for electric vehicle charging station occupancy prediction. Energy AI. 2026;24:100733. doi:10.1016/j.egyai.2026.100733.
12. Wu J, Gan W, Chen Z, Wan S, Yu PS. Multimodal large language models: a survey. In: Proceedings of the 2023 IEEE International Conference on Big Data (BigData); 2023 Dec 15–18; Sorrento, Italy. doi:10.1109/BigData59044.2023.10386743.
13. Singh A, Fry A, Perelman A, Tart A, Ganesh A, El-Kishky A, et al. OpenAI GPT-5 system card. arXiv:2601.03267. 2026.
14. Elhenawy M, Abutahoun A, Alhadidi TI, Jaber A, Ashqar HI, Jaradat S, et al. Visual reasoning and multi-agent approach in multimodal large language models (MLLMs): solving TSP and mTSP combinatorial challenges. Mach Learn Knowl Extr. 2024;6(3):1894–921. doi:10.3390/make6030093.
15. Jang K, Pilario KES, Lee N, Moon I, Na J. Explainable artificial intelligence for fault diagnosis of industrial processes. IEEE Trans Ind Inf. 2025;21(1):4–11. doi:10.1109/tii.2023.3240601.
16. Lundberg H, Mowla NI, Abedin SF, Thar K, Mahmood A, Gidlund M, et al. Experimental analysis of trustworthy in-vehicle intrusion detection system using eXplainable artificial intelligence (XAI). IEEE Access. 2022;10(260):102831–41. doi:10.1109/access.2022.3208573.
17. Choudhary A, Mian T, Fatima S, Panigrahi BK. Deep transfer learning based fault diagnosis of electric vehicle motor. In: Proceedings of the 2022 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES); 2022 Dec 14–17; Jaipur, India. doi:10.1109/PEDES56012.2022.10080274.
18. Palaiologou I, Falekas G, Fesakis N, Karlis A. Advancing electric vehicle inverter multi-class transient fault detection with sequential binary classification techniques. Eng Appl Artif Intell. 2025;158(Part A):111430. doi:10.1016/j.engappai.2025.111430.
19. Shete S, Jog P, Kamalakannan R, Raghesh JTA, Manikandan S, Kumawat RK. Fault diagnosis of electric vehicle's battery by deploying neural network. In: Proceedings of the 2022 Sixth International Conference on I-SMAC (IoT

in Social, Mobile, Analytics and Cloud) (I-SMAC); 2022 Nov 10–12; Dharan, Nepal. doi:10.1109/i-smac55078.2022.9987277.
20. Mehdiyev N, Majlatow M, Fettke P. Integrating permutation feature importance with conformal prediction for robust explainable artificial intelligence in predictive process monitoring. Eng Appl Artif Intell. 2025;149(4):110363. doi:10.1016/j.engappai.2025.110363.
21. Mi X, Zou B, Zou F, Hu J. Permutation-based identification of important biomarkers for complex diseases via machine learning models. Nat Commun. 2021;12(1):3008. doi:10.1038/s41467-021-22756-2.
22. Daimi SA, Iqbal A. Ensemble machine learning to predict and feature engineering to identify factors for academic-success. In: Proceedings of the International e-Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023); 2023 Sep 22–23; Hyderabad, India. doi:10.2991/978-94-6463-314-6_28.
23. NAVER Cloud HyperCLOVA X Team. HyperCLOVA X THINK technical report. arXiv:2506.22403. 2025.
24. Abbineni P, Aldowaish S, Liechty C, Noorzad S, Ghazizadeh A, Fayazi M. MuaLLM: a multimodal large language model agent for circuit design assistance with hybrid contextual retrieval-augmented generation. arXiv:2508.08137. 2025.
25. Hui Z, Li Y, Zhao D, Banbury C, Chen T, Koishida K. WinSpot: GUI grounding benchmark with multimodal large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul 27–Aug 1; Vienna, Austria. doi:10.18653/v1/2025.acl-short.85.
26. Louppe G. Understanding random forests: from theory to practice. arXiv:1407.7502. 2014.
27. Zhang C, Wang W, Liu L, Ren J, Wang L. Three-branch random forest intrusion detection model. Mathematics. 2022;10(23):4460. doi:10.3390/math10234460.
28. Gu J, Han Z, Chen S, Beirami A, He B, Zhang G, et al. A systematic survey of prompt engineering on vision-language foundation models. arXiv:2307.12980. 2023.
29. Han J, Buntine W, Shareghi E. VerifiAgent: a unified verification agent in language model reasoning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; 2025 Nov 4–9; Suzhou, China. doi:10.18653/v1/2025.findings-emnlp.891.
30. Kaggle. New energy vehicles diagnosis [Internet]. [cited 2026 Jul 27]. Available from: https://www.kaggle.com/datasets/ziya07/fault-diagnosis-dataset-for-new-energy-vehicles/data.
31. Beja-Battais P. Overview of AdaBoost: reconciling its views to better understand its dynamics. arXiv:2310.18323. 2023.
32. Cunningham P, Delany SJ. k-nearest neighbour classifiers—a tutorial. ACM Comput Surv. 2021;54(6):128. doi:10.1145/3459665.
33. Peng CJ, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. J Educ Res. 2002;96(1):3–14. doi:10.1080/00220670209598786.
34. Peretz O, Koren M, Koren O. Naive Bayes classifier—an ensemble procedure for recall and precision enrichment. Eng Appl Artif Intell. 2024;136(2):108972. doi:10.1016/j.engappai.2024.108972.
35. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106. doi:10.1007/BF00116251.
36. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. doi:10.1145/2939672.2939785.
37. Machado MR, Karray S, de Sousa IT. LightGBM: an effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In: Proceedings of the 2019 14th International Conference on Computer Science & Education (ICCSE); 2019 Aug 19–21; Toronto, ON, Canada. doi:10.1109/iccse.2019.8845529.
38. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing System; 2018 Dec 3–8; Red Hook, NY, USA.
39. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3–42. doi:10.1007/s10994-006-6226-1.
40. He Z, Lin D, Lau T, Wu M. Gradient boosting machine: a survey. arXiv:1908.06951. 2019.

41. Arlot S, Lerasle M. Choice of V for V-fold cross-validation in least-squares density estimation. J Mach Learn Res. 2016;17(208):1–50.
42. Kalyan KS. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Nat Lang Process J. 2024;6(6):100048. doi:10.1016/j.nlp.2023.100048.
43. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv:2303.08774. 2023.
44. Joshi A, Kale S, Chandel S, Pal D. Likert scale: explored and explained. Br J Appl Sci Technol. 2015;7(4):396–403. doi:10.9734/bjast/2015/14975.
45. Vujovic ŽĐ. Classification model evaluation metrics. Int J Adv Comput Sci Appl. 2021;12(6):599–606. doi:10.14569/ijacsa.2021.0120670.
46. Wang Y, Liu Z, Zheng W, Wang J, Shi H, Gu M. A combined multi-classification network intrusion detection system based on feature selection and neural network improvement. Appl Sci. 2023;13(14):8307. doi:10.3390/app13148307.
47. Chen D, Chen R, Zhang S, Wang Y, Liu Y, Zhou H, et al. MLLM-as-a-judge: assessing multimodal LLM-as-a-judge with vision-language benchmark. In: Proceedings of the 41st International Conference on Machine Learning; 2024 Jul 21–27; Vienna, Austria.
48. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation—a discussion and demonstration of basic features. PLoS One. 2019;14(7):e0219854. doi:10.1371/journal.pone.0219854.
49. O'Neill TA. An overview of interrater agreement on Likert scales for researchers and practitioners. Front Psychol. 2017;8:777. doi:10.3389/fpsyg.2017.00777.
50. Tinsley HE, Weiss DJ. Interrater reliability and agreement of subjective judgments. J Couns Psychol. 1975;22(4):358–76. doi:10.1037/h0076640.
51. Zhang Z, Dong S, Li D, Liu P, Wang Z. Prediction and diagnosis of electric vehicle battery fault based on abnormal voltage: using decision tree algorithm theories and isolated forest. Processes. 2024;12(1):136. doi:10.3390/pr12010136.
52. Lee J, Rew J. Multi-agent large language model-based decision tree analysis for explainable electric vehicle drive motor fault diagnosis. Comput Mater Contin. 2026;87(3):100. doi:10.32604/cmc.2026.077691.
53. Fu Y, Ji Y, Meng G, Chen W, Bai X. Three-phase inverter fault diagnosis based on an improved deep residual network. Electronics. 2023;12(16):3460. doi:10.3390/electronics12163460.
54. Xu X, Hang J, Cao K, Ding S, Wang W. Unified open-phase fault diagnosis of five-phase PMSM system in normal operation and fault-tolerant operation modes. IEEE Trans Transp Electrific. 2025;11(5):12063–75. doi:10.1109/tte.2025.3584777.
55. Lee J, Cheon S, Rew J. Vision-language model-based Grad-CAM analysis for explainable electric vehicle drive motor fault diagnosis. J Auto-Veh Assoc. 2026;18(1):29–46. doi:10.22680/kasa2026.18.1.029.
56. Ali AO, Abdelrehim O, Saafan MM, Elmarghany MR, Hamed AM. Comprehensive review of battery management systems for electric vehicles: thermal management, charging strategies, and emerging technologies. J Power Sources. 2025;658(26):238269. doi:10.1016/j.jpowsour.2025.238269.
57. Lee J, Lee J, Rew J. Multimodal large language model-based shapley interaction quantification analysis for interpretation of battery state-of-charge prediction in electric vehicles. Appl Sci. 2026;16(10):4812. doi:10.3390/app16104812.
58. Xu L, Teoh SS, Ibrahim H. A deep learning approach for electric motor fault diagnosis based on modified InceptionV3. Sci Rep. 2024;14(1):12344. doi:10.1038/s41598-024-63086-9.
59. Lee J, Rew J. DRIVE: diagnostic report integration via VLM and LLM explanations for explainable vehicle engine fault diagnosis. Comput Model Eng Sci. 2026;147(1):22. doi:10.32604/cmes.2026.076888.
60. Wandelt S, Zheng C, Wang S, Liu Y, Sun X. Large language models for intelligent transportation: a review of the state of the art and challenges. Appl Sci. 2024;14(17):7455. doi:10.3390/app14177455.
61. Guo T, Chen X, Wang Y, Chang R, Pei S, Chawla NV, et al. Large language model based multi-agents: a survey of progress and challenges. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence Survey Track; 2024 Aug 3–9; Jeju Island, Republic of Korea. doi:10.24963/ijcai.2024/890.
62. Alsaif KM, Albeshri AA, Khemakhem MA, Eassa FE. Multimodal large language model-based fault detection and diagnosis in context of industry 4.0. Electronics. 2024;13(24):4912. doi:10.3390/electronics13244912.

63. Xie J, Chen Z, Zhang R, Li G. Large multimodal agents: a survey. Visual Intell. 2025;3(1):24. doi:10.1007/s44267-025-00093-y.
64. Lee Y, Baruzzo G, Kim J, Seo J, Di Camillo B. Validity of feature importance in low-performing machine learning for tabular biomedical data. arXiv:2409.13342. 2024.
65. Pan G, Chodnekar V, Roy A, Wang H. A cost-benefit analysis of on-premise large language model deployment: breaking even with commercial LLM services. arXiv:2509.18101. 2025.
66. Wang X, Tang Z, Guo J, Meng T, Wang C, Wang T, et al. Empowering edge intelligence: a comprehensive survey on on-device AI models. ACM Comput Surv. 2025;57(9):1–39. doi:10.1145/3724420.
67. Yuan X, Chen H, Liu L, Li H. DistMLLM: enhancing multimodal large language model serving in heterogeneous edge computing. Sensors. 2025;25(24):7612. doi:10.3390/s25247612.
68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
69. Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations; 2016 Jun 12–17; San Diego, CA, USA. doi:10.18653/v1/n16-3020.
70. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput. 2024;16(1):45–74. doi:10.1007/s12559-023-10179-8.
71. dos Santos PC, Rocha MB, Krohling RA. Combining SHAP and causal analysis for interpretable fault detection in industrial processes. arXiv:2510.23817. 2025.
72. Zereen AN, Das A, Uddin J. Machine fault diagnosis using audio sensors data and explainable AI techniques-LIME and SHAP. Comput Mater Contin. 2024;80(3):3463–84. doi:10.32604/cmc.2024.054886.
73. Gao R. MARS: multi-agent robotic system with multimodal large language models for assistive intelligence. arXiv:2511.01594. 2025.
74. Raghuvira Pratap A, Panda S, Syama Sameera G. Predictive maintenance for two-wheeler vehicles using XGBoost. In: Proceedings of the 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS); 2014 Mar 14–15; Coimbatore, India. doi:10.1109/ICACCS60874.2024.10717187.
75. Kumar RS, Singh AR, Narayana PL, Chandrika VS, Bajaj M, Zaitsev I. Hybrid machine learning framework for predictive maintenance and anomaly detection in lithium-ion batteries using enhanced random forest. Sci Rep. 2025;15(1):6243. doi:10.1038/s41598-025-90810-w.
76. Brito LC, Susto GA, Brito JN, Duarte MAV. Fault diagnosis using eXplainable AI: a transfer learning-based approach for rotating machinery exploiting augmented synthetic data. Expert Syst Appl. 2023;232(4):120860. doi:10.1016/j.eswa.2023.120860.
77. Lee J, Lee J, Rew J. Multimodal large language model-based explainable boosting machine analysis for interpretation of state-of-health prediction of lithium-ion batteries. Electronics. 2026;15(8):1675. doi:10.3390/electronics15081675.
78. Wang T, Zhang B, Jiang D, Li D. A multimodal large language model framework for intelligent perception and decision-making in smart manufacturing. Sensors. 2025;25(10):3072. doi:10.3390/s25103072.
79. Zhang X, Liu S, Jiang L, Li Y. FSAMLM: a few-shot adaptation multimodal large model for cross-domain fault diagnosis. Appl Soft Comput. 2025;185(9):113985. doi:10.1016/j.asoc.2025.113985.
80. Kaplan H, Tehrani K, Jamshidi M. A fault diagnosis design based on deep learning approach for electric vehicle applications. Energies. 2021;14(20):6599. doi:10.3390/en14206599.
81. Gopalakrishnan R, Vidhya DS, Bharath P, Kaviyan P, Sivaselvam P. Battery state estimation and control system for mobile charging station for electric vehicles. In: Proceedings of the 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS); 2023 Mar 23–25; Erode, India. doi:10.1109/ICSCDS56580.2023.10104624.
82. Nixon S, Weichel R, Reichard K, Kozlowski J. Machine learning approach to diesel engine health prognostics using engine controller data. PHM Conf. 2018;10(1):1–10. doi:10.36001/phmconf.2018.v10i1.587.
83. Vergara M, Ramos L, Rivera-Campoverde ND, Rivas-Echeverría F. EngineFaultDB: a novel dataset for automotive engine fault classification and baseline results. IEEE Access. 2023;11:126155–71. doi:10.1109/ACCESS.2023.3331316.

84. Lee J, Rew J. Vision-language model-based local interpretable model-agnostic explanations analysis for explainable in-vehicle controller area network intrusion detection. Sensors. 2025;25(10):3020. doi:10.3390/s25103020.
85. Cação J, Santos J, Antunes M. Explainable AI for industrial fault diagnosis: a systematic review. J Ind Inf Integr. 2025;47(2):100905. doi:10.1016/j.jii.2025.100905.
86. Lee J, Rew J. Large language model-based SHAP analysis for interpretation of remaining useful life prediction of lithium-ion battery. J Korea Soc Ind Inf Syst. 2024;29(5):51–68.