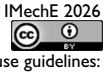
Check for updates

Original research article

Institution of MECHANICAL ENGINEERS
Journal of AUTOMOBILE ENGINEERING

Proc IMechE Part D:
J Automobile Engineering
1–15
© IMechE 2026

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09544070261463329
journals.sagepub.com/home/pid
Sage

Embedded fault detection for crewed and uncrewed defence vehicle suspension systems via explainable machine learning and four post rig testing

Irfan Ullah Khan[1], John Economou[1], Manik Preet Singh Sandhu[2], Adam Sandle[1], Robert Cooper[1] and David Galvão Wall[1]

Abstract

Suspension faults in defence vehicles (either crewed or uncrewed ground vehicles), such as damper leakage, loose mountings, and bushing degradation, can significantly impair stability, mobility, and control performance under severe off road terrain conditions. This paper presents an explainable machine learning framework for embedded suspension fault detection using multi domain vibration and acoustic sensing. The approach integrates eight physically interpretable time/frequency features with a Random Under Sampling Boosted (RUSBoost) ensemble classifier to discriminate four suspension health states. Full scale experimental validation was carried out on a Land Rover Defender 110 mounted on a four-post hydraulic vibration rig capable of reproducing broadband, defence-relevant excitation. The proposed method achieved 100% accuracy (AUC = 1) in a holdout evaluation and maintained a mean accuracy of 74% (AUC = 0.87) under tenfold cross validation, demonstrating strong generalisation. Feature importance analysis identified peak-to-peak amplitude, standard deviation, and RMS as the most influential indicators, confirming that suspension degradation is primarily characterised by increased vibration excursion and dispersion rather than spectral redistribution. Principal component analysis further showed partial overlap among fault classes, highlighting the necessity of nonlinear ensemble decision boundaries. Despite higher training cost, the ensemble classifier achieved prediction times that indicate computational feasibility for future real time embedded implementation. Overall, the framework provides a transparent, robust, and control compatible diagnostic layer, supporting future integration with predictive maintenance and fault tolerant suspension control in defence mobility systems.

Keywords

suspension fault detection, explainable machine learning, RUSBoost ensemble, vibration and acoustic sensing, four-post rig testing, defence vehicle diagnostics, feature importance analysis, fault tolerant control

Received: 23 February 2026; accepted: 11 June 2026

Introduction

Modern defence ground vehicles, encompassing both crewed and uncrewed platforms, are required to operate under highly demanding and unpredictable conditions while maintaining mobility, stability, and operational safety. These requirements impose significant stress on vehicle subsystems, particularly the suspension system, which plays a critical role in isolating the vehicle body from terrain-induced disturbances, maintaining tyre-ground contact, and ensuring controllability under dynamic loading conditions. Degradation of suspension components such as dampers, bushings, and mounting interfaces can significantly alter vehicle dynamics, leading to increased vibration transmission, degraded ride quality, reduced handling performance, and accelerated

[1]Centre for Defence Engineering and Physical Science, Cranfield University, Defence Academy of the United Kingdom, Shrivenham, UK
[2]Indian Army, Armoured Corps, New Delhi, India

Corresponding author:
Irfan Ullah Khan, Centre for Defence Engineering and Physical Science, Cranfield University, Defence Academy of the United Kingdom, Shrivenham, SN6 8LA, UK.
Email: irfan.khan2@cranfield.ac.uk

structural fatigue.[1,2] In military contexts, where vehicle reliability is closely tied to mission readiness and personnel safety, undetected suspension degradation can therefore have serious operational consequences.

Active and semi-active suspension systems have been extensively studied within the control engineering literature as a means of improving ride comfort, handling performance, and robustness under varying operating conditions. A wide range of control strategies has been proposed for automotive suspension systems, including frequency-shaped, robust, and gain-scheduled formulations tailored to semi-active actuators and real-world implementation constraints.[3–5] More recently, linear parameter-varying (LPV) control frameworks have proven effective in systematically addressing actuator limitations and nonlinear suspension behaviour while maintaining robust performance across operating regimes.[6]

While such control strategies improve performance, their effectiveness ultimately depends on the underlying health of suspension components, motivating reliable condition monitoring and diagnostic capabilities. Conventional suspension health assessment relies largely on periodic inspections and schedule-based maintenance strategies. While straightforward to implement, such approaches are poorly suited to defence vehicles operating across diverse terrains and variable mission profiles. Consequently, condition-based maintenance (CBM) and prognostics and health management (PHM) strategies have gained increasing attention within military ground vehicle fleets.[7,8] These strategies aim to infer system health directly from operational data, enabling maintenance decisions to be driven by actual degradation rather than fixed service intervals. However, compared with aerospace systems, the deployment of CBM and PHM in ground vehicles remains limited due to data uncertainty, environmental variability, and challenges associated with reliable fault labelling.[9,10]

A key challenge in implementing CBM for ground vehicles lies in acquiring reliable and informative health data under real operating conditions. Ground vehicles experience broadband and non-stationary excitation arising from terrain interaction, speed variation, and payload changes, which complicates the isolation of fault related signatures using simple threshold-based indicators. Nevertheless, vibration measurements have consistently been shown to contain valuable information regarding suspension health, as changes in damping and stiffness characteristics directly affect the vehicle dynamic response.[1,11] As a result, vibration-based condition monitoring has emerged as a promising approach for suspension diagnostics when supported by appropriate signal processing and analysis techniques.

Early vibration-based fault diagnosis studies relied on analytical and signal processing driven approaches, employing time domain statistics and frequency domain indicators to detect deviations from nominal system behaviour.[12] While effective under controlled conditions, such methods often struggle in the presence of stochastic excitation and operating variability typical of off-road environments. To address these limitations, recent research has increasingly adopted data driven and machine learning (ML) techniques capable of capturing nonlinear relationships between measured signals and system health. ML based approaches have demonstrated strong performance across a wide range of mechanical systems, including vibration-based fault diagnosis and remaining useful life prediction.[13–15]

Despite these advances, several challenges remain when applying ML based diagnostics to vehicle suspension systems. One persistent issue is class imbalance, as vehicles operate in healthy conditions for most of their service life, resulting in relatively few fault samples for training classifiers. This imbalance can bias learning algorithms toward the majority class and reduce sensitivity to early stage or rare fault modes.[16] In addition, many ML models operate as black boxes, offering limited insight into the physical factors driving their decisions. In safety and mission critical applications, this lack of transparency can hinder trust, validation, and integration with maintenance or supervisory control frameworks.[10]

Ensemble learning techniques provide a promising pathway to improving robustness under noisy and imbalanced conditions. By combining multiple base learners, ensemble models can reduce variance and improve generalisation compared with single classifiers. In particular, boosting-based ensemble methods such as RUSBoost have demonstrated strong performance in imbalanced fault diagnosis problems across a range of mechanical systems.[17,18] Within the context of vehicle suspensions, experimental studies have shown that damper and suspension faults can be effectively identified using vibration derived features when datasets are carefully labelled and features are selected with clear physical relevance.[19]

Alongside diagnostic accuracy, interpretability has become an increasingly important requirement for ML based diagnostics in safety-critical systems. This has motivated growing interest in explainable artificial intelligence (XAI) techniques, which aim to provide transparency by quantifying the contribution of individual input features to model outputs. Comprehensive surveys have highlighted the role of XAI in improving trust, validation, and regulatory acceptance of ML based industrial diagnostics.[20,21] In vibration-based diagnostics, explainability enables diagnostic decisions to be directly linked to physically meaningful indicators such as vibration amplitude, dispersion, and transmitted energy.

Building on these developments, recent studies have explored explainable machine learning frameworks for condition monitoring and predictive maintenance, particularly in vibration-based diagnostics, where interpretability enables diagnostic decisions to be linked to physically meaningful system behaviour.

Methods such as feature attribution, model agnostic explanation, and interpretable ensemble learning have been applied to improve transparency and trust in fault diagnosis systems.[22–24]

In parallel, machine learning based condition monitoring using vibration data has been widely investigated for mechanical systems and fault diagnosis, demonstrating strong capability in identifying degradation patterns under complex operating conditions.[25–27] These approaches provide a foundation for integrating data driven diagnostics with physically interpretable features in practical engineering applications.

Another limitation of much of the existing suspension diagnostics literature is the reliance on simulations or simplified laboratory experiments that do not fully capture the coupled dynamics of full vehicle systems. Suspension behaviour is strongly influenced by vehicle mass distribution, suspension geometry, and multi-axial loading, which are difficult to replicate using small scale test rigs. Full scale experimental validation is therefore essential to assess the practical viability of diagnostic approaches. Four post hydraulic vibration rigs provide a controlled environment capable of reproducing representative excitation profiles while preserving realistic vehicle suspension dynamics, making them particularly suitable for full vehicle suspension assessment under repeatable conditions.

Despite the progress reported in vibration-based diagnosis and machine learning based condition monitoring, several gaps remain in the context of defence vehicle suspension systems. First, many existing studies rely on simulations, component level tests, or simplified laboratory conditions rather than full scale vehicle level validation. Second, the interpretability of diagnostic decisions is often limited, which reduces confidence in safety critical applications where engineering transparency is essential. Third, relatively few studies discuss whether the resulting diagnostic framework is practically compatible with embedded monitoring architectures suitable for tactical or autonomous ground vehicles.

Motivated by these gaps, this study proposes an explainable machine learning framework for suspension fault detection in crewed and uncrewed defence vehicles, validated using full scale four post rig testing. The approach integrates vibration measurements acquired through an embedded data acquisition architecture suitable for tactical vehicles[28] with physically meaningful feature extraction and an ensemble-based classification framework.

The main contributions of the work are threefold. First, a full-scale experimental dataset is generated on a Land Rover Defender 110 under controlled but defence relevant broadband excitation profiles. Second, an interpretable diagnostic pipeline is developed using physically meaningful vibration and acoustic features together with a RUSBoost ensemble classifier. Third, the framework is analysed not only in terms of classification performance but also in terms of feature level interpretability and computational feasibility for future embedded implementation. By combining full scale experimentation, interpretable learning, and deployment-oriented analysis, the study aims to advance practical suspension health monitoring for defence mobility systems.

The remainder of this paper is organised as follows. Section "Problem formulation" presents the problem formulation. Section "Experimental setup" describes the experimental setup and data acquisition process. Section "Methodology" details the feature extraction, classification, and explainability analysis. Section "Results and discussion" presents and discusses the results. Section "Conclusion" concludes the study and outlines future directions toward embedded implementation and health aware control integration.

Problem formulation

Based on the diagnostic requirements outlined in Section "Introduction", the suspension fault detection task is formulated as a supervised, data driven classification problem in feature space. Multi domain vibration and acoustic measurements are first transformed into a set of physically interpretable statistical and spectral features, which serve as inputs to the diagnostic model.

From each sensor channel, a feature vector is extracted as defined in equations (1) and (2), respectively:

$$f_v = \left[f_{\{v,1\}}, f_{\{v,2\}}, \ldots, f_{\{v,p\}}\right] \tag{1}$$

$$f_a = \left[f_{\{a,1\}}, f_{\{a,2\}}, \ldots, f_{\{a,q\}}\right] \tag{2}$$

Where $f_{v,i}$ and $f_{a,i}$ include both time domain descriptors (peak-to-peak amplitude, standard deviation, RMS, kurtosis, skewness, crest factor) and frequency domain parameters (spectral centroid, entropy).

The fused feature vector is given in equation (3):

$$f = [f_v, f_a] \tag{3}$$

and the labelled dataset is defined in equation (4) as:

$$D = \left\{\left(f_j, y_j\right) | j = 1, 2, \ldots, N\right\};\ y_j \in \{0, 1, 2, 3\} \tag{4}$$

where N denotes the total number of run level samples, and $y_j \in \{0, 1, 2, 3\}$ represents the class label corresponding to the healthy condition and the three fault types.

The diagnostic task is therefore formulated as the multiclass classification problem given in equation (5):

$$\hat{y} = \arg\max_k \mathrm{P}(y | f) \tag{5}$$

where $\mathrm{P}(y | f)$ denotes the posterior probability of each health state estimated by the classifier.

To address both class imbalance and nonlinear decision boundaries inherent in real world suspension fault data, a Random Under Sampling Boosted (RUSBoost) ensemble tree classifier is employed. Let the feature matrix be defined as in equation (6):

$$F \in \mathbb{R}^{n \times m} \tag{6}$$

where $F \in \mathbb{R}^{n \times m}$ contains n run level samples and m extracted features.

The learning objective is to determine a decision function $H(\mathrm{f}) \in \mathcal{H}$ that minimises the expected misclassification risk as expressed in equation (7):

$$\min_{\{H \in \mathcal{H}\}} E[L(y, H(f))] \tag{7}$$

where $L(\cdot)$ is a loss function penalizing false classifications. The RUSBoost algorithm integrates random under sampling with AdaBoost style weighting.

At iteration t, the algorithm forms a balanced dataset D_t by under sampling majority classes and trains a weak learner $h_t(f)$. The corresponding weighted classification error is computed as in equation (8):

$$\epsilon_t = \sum_{i=1}^{n} w_i^{(t)} \mathbb{I}(h_t(f_i) \neq y_i) \tag{8}$$

where $w_i^{(t)}$ denotes the sample weight at iteration t, and $\mathbb{I}(\cdot)$ is the indicator function. The learner weight is then given by equation (9):

$$\alpha_t = \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{9}$$

The sample weights are subsequently updated according to equation (10)

$$w_i^{(t+1)} = w_i^{(t)} \exp\left(\alpha_t \mathbb{I}(h_t(f_i) \neq y_i)\right) \tag{10}$$

and normalised to sum to one.

The final multiclass decision is then obtained through the weighted majority vote defined in equation (11):

$$H(\mathrm{f}) = \arg\max_k \sum_{t=1}^{T} \alpha_t \, \mathbb{I}(h_t(\mathrm{f}) = k) \tag{11}$$

where T denotes the total number of boosting iterations.

Random under sampling ensures adequate representation of minority fault classes in each iteration, while boosting progressively emphasises difficult to classify samples, thereby enhancing discrimination capability. This mathematical framework provides a transparent mapping from physical vibration features to diagnostic outputs while maintaining computational simplicity suitable for embedded controllers. The classifier output is subsequently analysed using an explainability procedure described in Section "Explainability analysis".

Although the classifier operates in feature space, the selected descriptors correspond to physically meaningful aspects of suspension behaviour, including vibration excursion, dispersion, impulsiveness, and spectral content. The resulting diagnostic outputs are therefore suitable for interpretation within supervisory monitoring and health aware control architectures.

Experimental setup

The experimental configuration combines a four-post hydraulic excitation platform with a distributed sensing architecture to capture synchronized dynamic responses from the suspension system under controlled operating conditions. The framework was designed to ensure repeatable excitation and reliable measurement acquisition.

Four post vibration rig

All experiments were conducted on a four-post hydraulic vibration rig designed for full vehicle suspension testing. Each post consists of a servo-hydraulic actuator (±50 mm stroke, 5 kN peak force) driven by a high response servo valve and PID displacement controller. The actuators operate in closed loop displacement control, ensuring precise replication of reference inputs regardless of dynamic loading. The system is synchronized via a real time controller, allowing independent or coordinated excitation of all four corners for broadband road profile reproduction. A Land Rover Defender 110 suspension assembly was selected as a representative defence vehicle platform.

Instrumentation and data acquisition

A distributed multi sensor architecture was deployed on the Land Rover 110 shown in Figure 1, to enable synchronized acquisition of vibration, acoustic and inertial data during four post rig testing.

Six microcontroller-based nodes were installed across the vehicle. Four nodes (T1–T4) were mounted adjacent to each shock absorber at the rear left, rear right, front right and front left suspension points. Each node recorded tri axial acceleration using an MPU6050 module and captured local acoustic signatures via a MAX9814 microphone. A fifth node (T5) installed at the vehicle centre of gravity provided reference inertial measurements for global motion correlation and a RX node, acted as the master unit responsible for aggregating all incoming sensor messages. All nodes communicated through a Controller Area Network (CAN) using MCP2515 interfaces. Each transmitter broadcast an 8-byte CAN frame at 100 Hz.

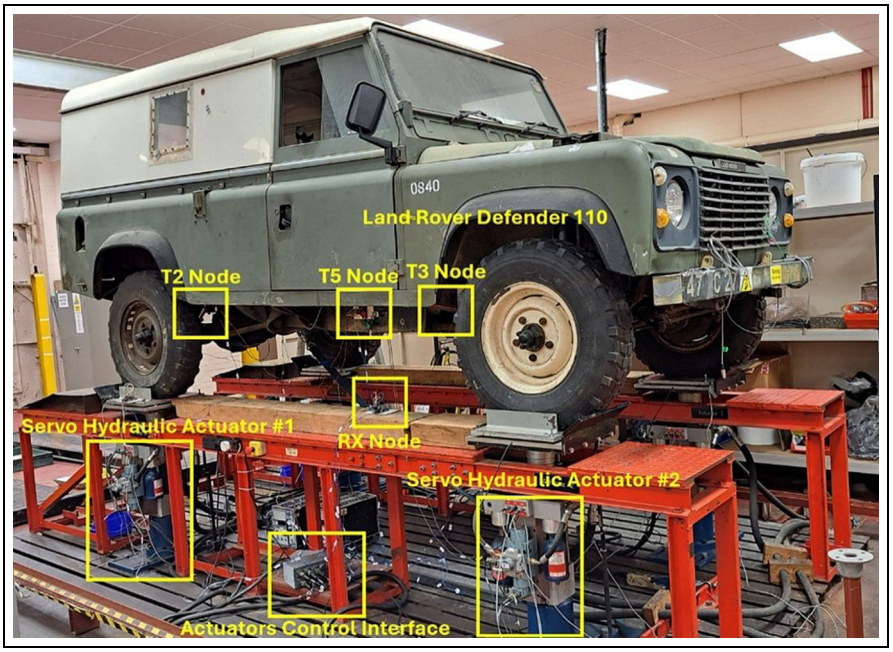


Figure 1. Experimental setup showing the distributed sensor architecture installed on the Land Rover Defender 110.

Table 1. Health state definitions and dataset structure.

Health state	Label code	Sampling rate (Hz)	Duration (s)	No. of runs
Healthy	0	100	300	12
Damper oil leakage fault	1	100	300	12
Suspension loose mounting	2	100	300	12
Damaged bushing	3	100	300	12

Excitation profiles

To reproduce representative defence terrain dynamics under controlled laboratory conditions, 12 excitation profiles were programmed on the four-post vibration rig. These profiles included sinusoidal excitation at two resonant frequencies (2.5 and 12.5 Hz) to characterise the fundamental dynamic response of the suspension system, as well as random broadband inputs of varying severity ranging from mild to extreme, to emulate stochastic off-road terrain excitation. In addition, composite manoeuvres combining pitch, roll, and twist components were applied to evaluate cross corner coupling effects and multi axial suspension behaviour. Each test was conducted for a duration of 300 seconds, allowing both transient and steady-state responses to be captured. Identical excitation profiles were applied across all four suspension health states to ensure consistency and enable direct comparative analysis.

Testing protocol

A total of 48 experiments were conducted shown in Table 1, grouped into four sets representing one healthy condition and three distinct suspension faults: shock absorber leakage, loose mounting hardware, and bushing degradation. Each set comprised 12 standardised tests including resonant frequency excitation, random broadband inputs and low-frequency pitch/roll/twist manoeuvres, ensuring comprehensive excitation of suspension dynamics. In the subsequent machine learning analysis, each experiment was treated as one independent run level sample after preprocessing and feature extraction.

Before each experiment, the vehicle and distributed sensor network were initialised, and all microcontroller nodes verified for synchronized CAN based data transmission. During each test, triaxial acceleration, acoustic responses and central rotational motion were recorded simultaneously across 19 sensor channels. Fault conditions were introduced manually between test sets, followed by mechanical inspection to ensure repeatability and safety. Following every run, data integrity checks were performed to identify communication losses, sensor anomalies or inconsistencies before proceeding to the next configuration.

This structured protocol enabled the generation of a high-quality labelled dataset under controlled yet operationally representative conditions, supporting robust training and validation of the machine learning classifiers for real time suspension fault detection.

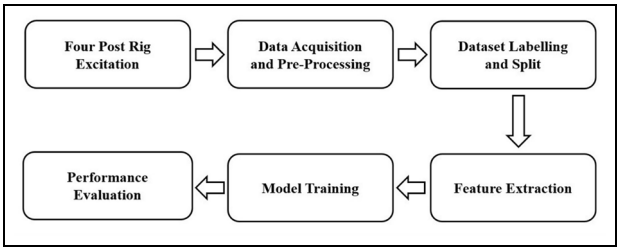


Figure 2. Experimentation process flowchart illustrating data acquisition, feature extraction, machine learning classification, and performance evaluation stages.

Figure 2 summarises the experimental workflow. Vibration and acoustic responses from the four-post rig are acquired, labelled by fault condition, and converted into engineered features in the time and frequency domains. These features serve as inputs to various machine learning models, which are subsequently evaluated using standard performance metrics.

Methodology

The proposed diagnostic framework integrates vibration and acoustic signal processing, feature engineering, explainable ensemble learning, and quantitative evaluation into a unified workflow suitable for embedded implementation within control-oriented architectures.

Feature extraction

A total of eight statistical features were extracted from each signal segment shown in Table 2. The eight selected features; six-time domain descriptors and two frequency domain measures were chosen because they provide a compact yet physically meaningful representation of suspension vibrational behaviour under broadband excitation. These features capture amplitude statistics, impulsiveness, waveform asymmetry, and spectral distribution. They were selected after preliminary exploration because they showed consistent, monotonic changes across all fault modes while remaining computationally feasible for embedded microcontrollers. Although the framework uses eight features for real time efficiency, it can be extended with additional spectral or nonlinear features to enhance discrimination in future work.

Each 300 seconds experimental run was first preprocessed by removing transient portions at the beginning and end of the recording. Feature extraction was then performed over the full duration of the remaining signal for each sensor channel, rather than on short, segmented windows. This produced one fused feature vector per experimental run. Since the dataset contains 48 independent runs, the final labelled dataset used for classification therefore comprised 48 run level samples.

Table 2. Extracted statistical features and their mathematical definitions.

Domain	Feature	Equation
Time domain	Standard deviation	$\sigma_x = \sqrt{\left[\left(\frac{1}{N-1}\right)\Sigma\left(x_i - \bar{x}\right)^2\right]}$
	Root mean square	$RMS = \sqrt{\left[\left(\frac{1}{N}\right)\Sigma x_i^2\right]}$
	Skewness	$S_x = \frac{\left[\left(\frac{1}{N}\right)\Sigma\left(x_i - \bar{x}\right)^3\right]}{\sigma_x{}^3}$
	Kurtosis	$K_x = \frac{\left[\left(\frac{1}{N}\right)\Sigma\left(x_i - \bar{x}\right)^4\right]}{\sigma_x{}^4}$
	Peak-to-peak value	$x_{p-p} = x_{max} - x_{min}$
	Crest factor	$CF = \frac{x_p e_a k}{RMS}$
Frequency domain	Spectral centroid	$f_c = \frac{(\Sigma f_k P_k)}{(\Sigma P_k)}$
	Spectral entropy	$H_s = -\Sigma P_k log P_k$

Classification models

Four supervised classification models were selected for comparative evaluation to represent a range of learning paradigms with differing assumptions, complexity and robustness characteristics. A Decision Tree was included as a baseline classifier due to its interpretability and ability to model nonlinear decision boundaries using simple rule-based structures. However, single tree models are known to be sensitive to data variability and class imbalance, making them useful reference points for assessing the benefits of more advanced methods.

Support Vector Machine (SVM) and k-Nearest Neighbour (kNN) classifiers were selected as widely used benchmark models in vibration-based fault diagnosis. The SVM represents a margin-based learning approach that performs well when class boundaries are well separated, while kNN provides a distance-based, non-parametric method that relies on local neighbourhood structure. Including these models enables comparison against commonly adopted classifiers that do not explicitly address class imbalance or ensemble learning.

The RUSBoost ensemble tree classifier was adopted as the primary diagnostic model because it offers several advantages for the present application. First, the ensemble structure is well suited to capturing nonlinear interactions among engineered vibration and acoustic features. Second, boosting improves stability and discrimination relative to a single decision tree when the dataset is limited in size and subject to variability across experimental runs. Third, the tree-based formulation remains compatible with feature importance analysis, which supports interpretability. Although RUSBoost is often motivated by imbalanced learning, in the present study its selection was based more broadly on robustness, nonlinear classification capability, and empirical performance relative to the benchmark models considered.

Table 3. Single-run (hold-out) classification performance of four models.

Classifier	Accuracy (%)	Precision	Recall	F1 score	AUC
RUSBoost ensemble	100	0.75	0.75	0.75	1.00
Decision tree	92	0.75	0.67	0.70	1.00
SVM	30	0.50	0.23	0.32	0.78
kNN	23	0.22	0.18	0.20	0.65

Explainability analysis

Explainability was incorporated into the diagnostic framework using permutation-based feature importance computed from the trained ensemble model. This approach quantifies the contribution of each feature by measuring the reduction in predictive performance when the values of that feature are permuted while the remaining features are left unchanged. In the present study, this method was selected because it is simple to implement, directly compatible with the tree-based ensemble formulation, and suitable for identifying the global contribution of each engineered feature to suspension fault classification.

The resulting feature importance ranking provides a global interpretation of the model and helps relate classifier behaviour to physically meaningful descriptors such as vibration excursion, dispersion, impulsiveness, and spectral content. While this global analysis improves transparency, it does not capture local sample wise explanations. More advanced methods, such as SHAP analysis or partial dependence analysis, may provide complementary interpretability and are identified as useful directions for future work.

Model training and validation

Model training and validation were carried out at the level of independent experimental runs. Each recorded file corresponds to one complete run, and one fused feature vector was extracted from each run after preprocessing, as described in Section "Feature extraction".

In the hold-out evaluation, the run level dataset was randomly partitioned into 70% training data and 30% testing data. Because the partitioning was performed across complete runs, no data from a given experiment appeared simultaneously in both training and testing sets. This ensured independence between model development and evaluation.

To obtain a more statistically robust estimate of performance, K-fold cross-validation was additionally applied using the same run level samples. In each fold, the model was trained on the training subset and evaluated on the corresponding validation subset, with the partitions kept identical across all classifiers to ensure a fair comparison. Performance was quantified using accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC).

This run level evaluation strategy is important because it avoids the optimistic bias that can arise when correlated samples from the same experimental record are shared across training and testing partitions. The cross-validation results are therefore interpreted as the more representative estimate of generalisation capability.

Results and discussion

This section analyses the diagnostic performance of the proposed framework using both single-run hold out evaluation and ten-fold cross validation, followed by computational efficiency assessment. Together, these results provide insight into peak performance, generalisation capability, physical interpretability, and practical deployability.

Single-run (hold out) evaluation

In the initial evaluation, models were trained on 70% of the dataset and tested on the remaining 30%. The purpose was to estimate the maximum achievable classification accuracy without repeated sampling. Table 3 summarises the classification performance of the four models under a single run hold out evaluation, providing an estimate of peak achievable performance on a fixed training/testing split. Accuracy offers a high-level measure of correct classification, while precision, recall, and F1 score collectively describe the balance between false alarms and missed fault detections. The AUC complements these metrics by assessing threshold independent class separability.

Under this evaluation, the RUSBoost ensemble and the Decision Tree demonstrate strong peak performance, whereas SVM and kNN exhibit limited discriminatory capability. These results establish an upper bound on classifier performance but do not, by themselves, reflect robustness to data variability. Detailed class wise behaviour and decision stability are therefore examined using confusion matrices and ROC analysis.

Figure 3 presents the normalised confusion matrices for the four classifiers under the single-run hold-out evaluation. The RUSBoost ensemble achieves perfect classification across all four suspension health states, indicating that the ensemble effectively captures nonlinear decision boundaries and compensates for class imbalance. The Decision Tree also performs strongly but misclassifies 33.3% of Class 2 samples as Class 0, reflecting the sensitivity of single tree models to training data partitioning.

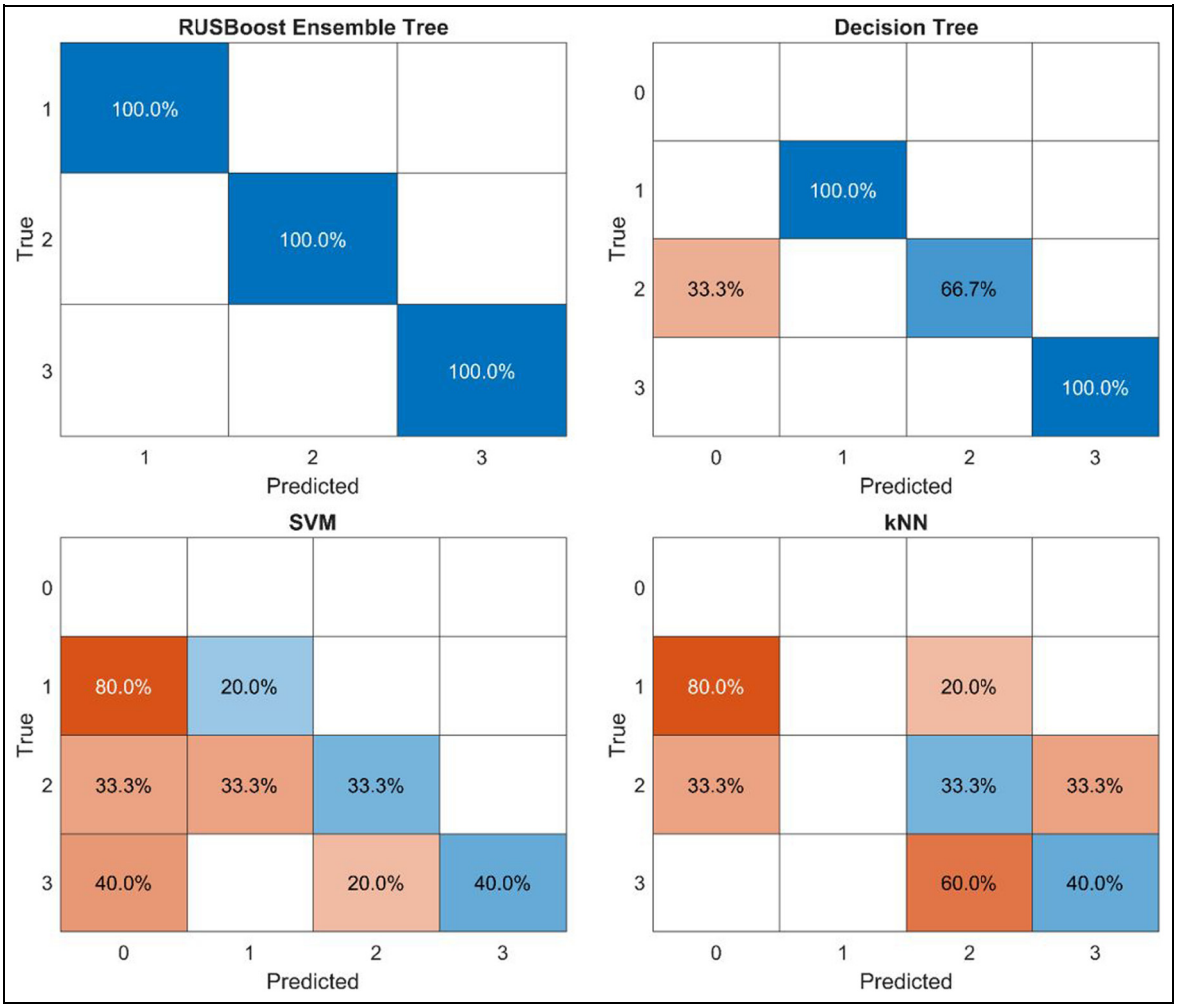


Figure 3. Normalised confusion matrices for the four classifiers under single-run (hold-out) evaluation.

In contrast, SVM and kNN exhibit substantial confusion across all fault categories. Both classifiers frequently misidentify faulty conditions as healthy and show strong cross class ambiguity. This behaviour reflects the difficulty of distance based and margin-based classifiers in resolving overlapping vibration feature distributions under broadband excitation.

Overall, the confusion matrices confirm that ensemble learning significantly enhances diagnostic reliability compared with individual classifiers.

Figure 4 presents the ROC curves obtained from the single run hold out evaluation. Both the RUSBoost ensemble and the Decision Tree achieve an AUC of 1, indicating perfect class separability for this specific data partition. However, identical ROC performance does not imply identical classification robustness in a multi-class setting. While the ensemble achieves perfect class assignment, the Decision Tree exhibits misclassification in the corresponding confusion matrix, demonstrating that ROC primarily reflects ranking capability rather than discrete class prediction stability.

The SVM and kNN classifiers exhibit ROC curves that lie above the diagonal reference line but remain only marginally separated from it, indicating limited discriminative capability. Their gradual rise and proximity to the diagonal over much of the false positive range explain the modest AUC values and poor classification performance observed in the confusion matrices. These observations highlight the limitation of relying solely on single run ROC analysis and reinforce the importance of complementary evaluation metrics and cross validation.

Figure 5 presents the out-of-bag permutation feature importance obtained from the ensemble classifier. Peak-to-peak amplitude and standard deviation dominate the ranking, followed by RMS. These amplitude domain features are physically linked to suspension degradation, as faults reduce damping effectiveness and increase vibration excursion, dispersion, and transmitted energy. Kurtosis and crest factor contribute moderately, indicating the presence of impulsive behaviour but confirming that suspension degradation in

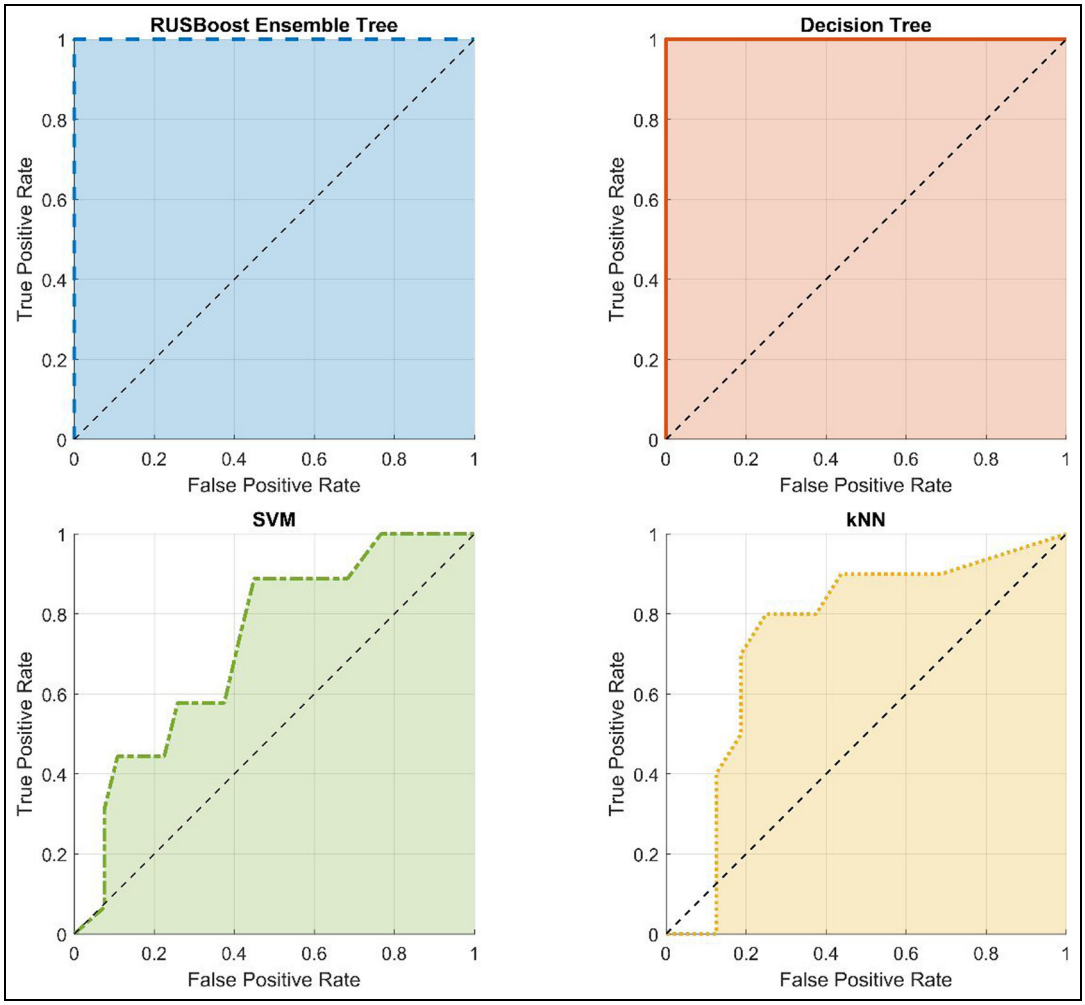


Figure 4. ROC curves with shaded AUC regions for the four classifiers under single-run (hold-out) evaluation.

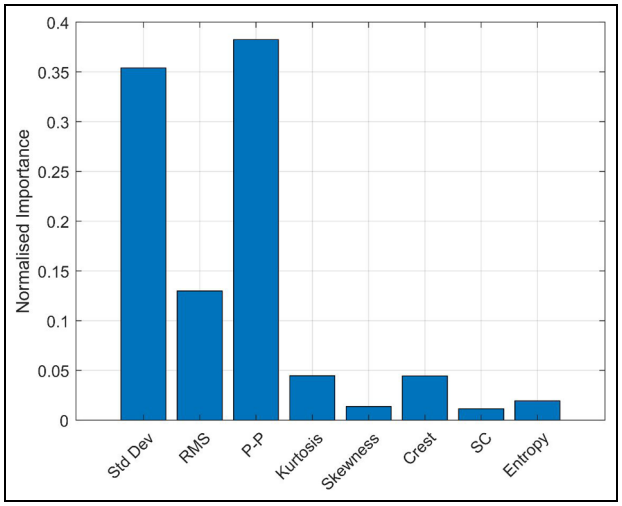


Figure 5. Normalised out-of-bag (OOB) permutation feature importance derived from the ensemble classifier.

this dataset is not dominated by impact type events. Spectral entropy, skewness, and spectral centroid exhibit minimal influence, showing that frequency redistribution and waveform asymmetry play a secondary role.

This ranking confirms that suspension fault signatures under broadband excitation are primarily amplitude driven rather than frequency complexity driven, which is fundamentally different from many rotating machinery fault scenarios. It should be noted that the presented feature importance analysis provides a global interpretation of model behaviour across the dataset and does not capture local sample-specific explanations. More advanced local explainability methods, such as SHAP analysis or partial dependence analysis, may provide complementary insight into individual predictions and are identified as useful directions for future work.

Figure 6 presents the PCA projection of the extracted feature space onto the first two principal components, which together account for 59% of the total variance. The first principal component (45.5%) is primarily associated with amplitude related characteristics, reflecting variations in vibration energy and excursion, while the second component (13.5%)

Table 4. Ten-fold cross validation performance metrics.

Classifier	Accuracy (%)	Precision	Recall	F1 score	AUC
RUSBoost ensemble	74	0.75	0.78	0.76	0.87
Decision tree	64	0.65	0.68	0.66	0.78
SVM	36	0.29	0.33	0.31	0.54
kNN	28	0.38	0.29	0.33	0.47

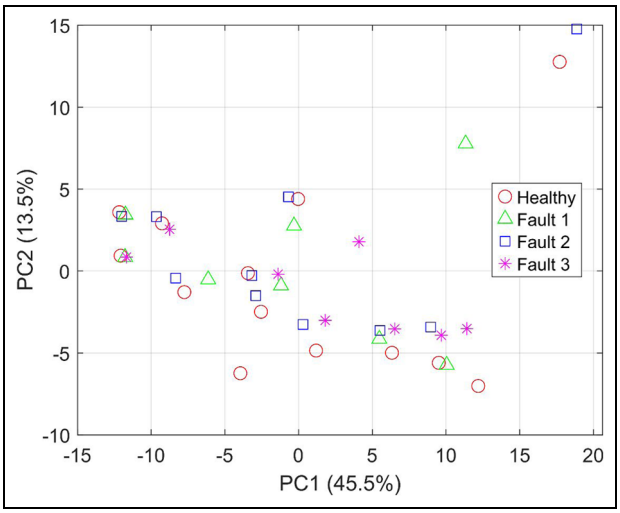


Figure 6. PCA projection of the feature space for single run evaluation, showing class distribution and overlap.

captures secondary variability linked to higher order statistical and spectral effects.

The projection shows a tendency for healthy samples to cluster separately from faulty conditions; however, significant overlap is observed among the different fault classes. This overlap reflects the inherent physical similarity of suspension degradation mechanisms, which often produce comparable vibration energy levels under broadband excitation. The absence of clear linear separability explains the limited performance of linear classifiers and motivates the use of ensemble learning to construct nonlinear decision boundaries.

Overall, the PCA confirms that the selected feature set preserves meaningful discriminatory structure while also highlighting the intrinsic complexity of suspension fault classification.

Cross validation results

To obtain a statistically reliable estimate of model robustness, a 10-fold cross validation procedure was applied to the same dataset. Each model was trained and validated 10 times on different data partitions, and mean metrics were computed. Table 4 reports the averaged performance metrics obtained from 10-fold cross-validation, providing a statistically robust assessment of generalisation under repeated data resampling. Unlike the single run evaluation, these results reflect variability in training data composition and are therefore more representative of real-world diagnostic performance.

Although all classifiers exhibit reduced accuracy and AUC compared with the hold-out case, the RUSBoost ensemble maintains the strongest and most consistent performance across all metrics. The Decision Tree shows moderate degradation, while SVM and kNN remain close to random discrimination when all classes are considered collectively. These trends are further examined through cross validation confusion matrices and ROC curves to assess stability and class wise robustness.

Figure 7 presents the confusion matrices for all classifiers under 10-fold cross validation. The RUSBoost ensemble shows the most balanced performance across the four classes, correctly identifying the dominant patterns in each fault state, although some confusion remains between Classes 0, 1, and 2 due to overlapping vibration characteristics. The Decision Tree exhibits similar but slightly less stable behaviour, with stronger misclassification for Class 2. In contrast, the SVM and kNN classifiers show substantial confusion across all classes, particularly for classes with similar energy signatures, indicating weak generalisation. Overall, the matrices confirm that RUSBoost provides the most reliable multi-class discrimination under resampled conditions.

Figure 8 presents the ROC curves obtained under 10-fold cross validation. As expected, all classifiers exhibit reduced performance compared to the hold out evaluation due to variability introduced by fold-to-fold resampling. The RUSBoost ensemble consistently achieves the highest AUC and maintains superior class separability across the false-positive rate range, highlighting its robustness to changes in training subsets.

The Decision Tree classifier performs competitively and remains above the random baseline, but generally underperforms the ensemble, reflecting the benefits of boosting and resampling in mitigating overfitting. Although Decision Trees are known to be high variance models, the ensemble formulation in RUSBoost provides improved stability and discrimination. In contrast, SVM and kNN classifiers operate close to the diagonal, indicating limited discriminative capability

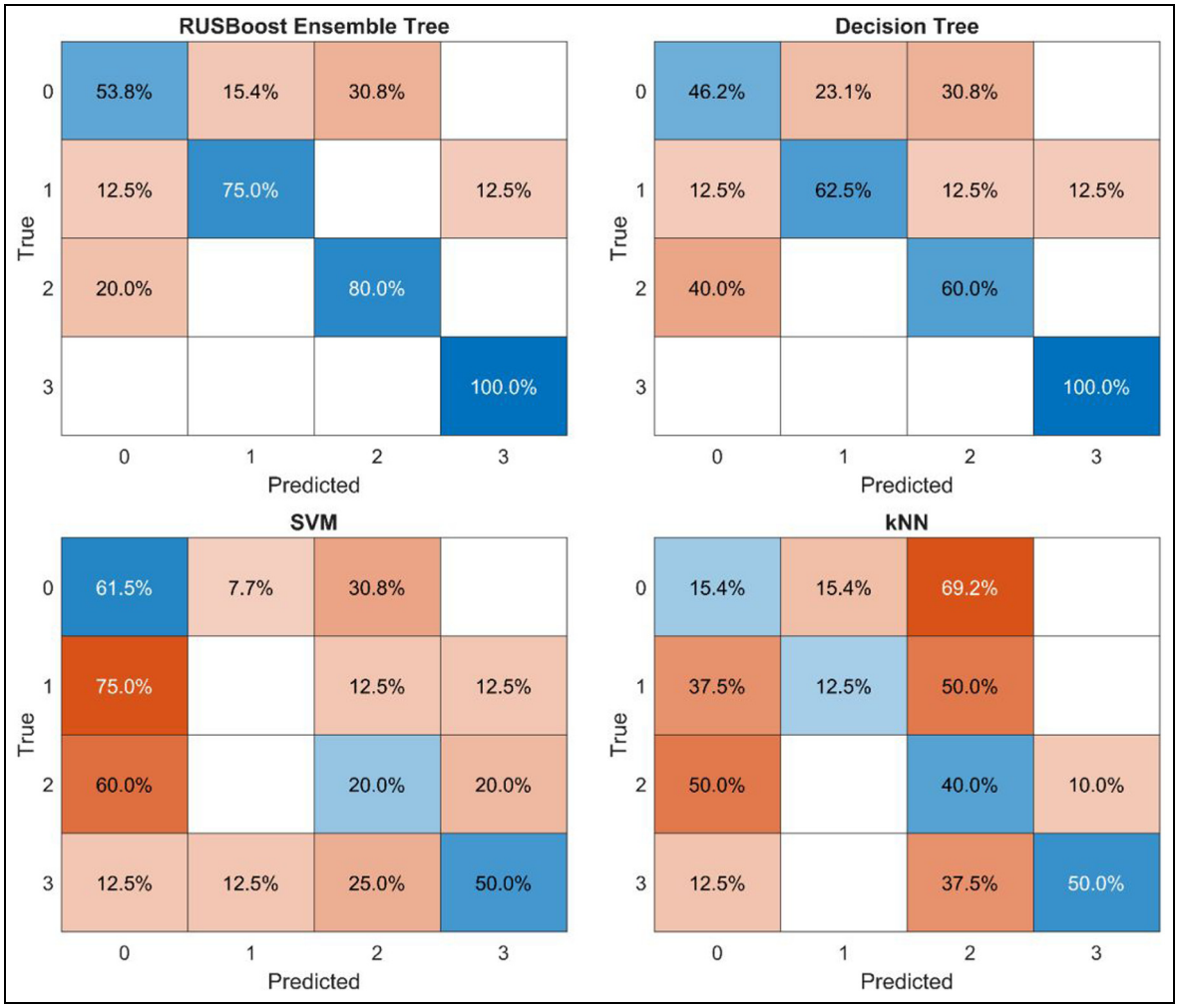


Figure 7. Confusion matrices for the four classifiers under 10-fold cross validation.

under cross validation. Overall, the results confirm that the RUSBoost ensemble generalises more reliably under realistic data variability.

Comparative analysis: Hold-out versus cross-validation

Table 5 compares the performance of each classifier under the single run hold out evaluation and the 10-fold cross validation procedure. As expected, the hold out accuracies are generally higher, since the model benefits from a fixed and relatively larger training set. In contrast, cross validation exposes the classifier to greater variability by repeatedly training on different subsets of the data, which naturally decreases accuracy but provides a more realistic estimate of generalisation capability.

The ensemble model achieved the strongest and most consistent performance across both evaluation strategies. However, the difference between the hold-out accuracy and the cross-validation accuracy is itself informative and should be interpreted carefully. The 100% hold-out result reflects performance on one fixed partition of a relatively small run level dataset and therefore represents a favourable single split outcome rather than a definitive estimate of real-world generalisation. By contrast, the cross-validation result is more conservative because it evaluates the classifier repeatedly under varying training validation partitions.

Several factors likely contribute to the observed reduction in performance under cross-validation. First, the number of independent experimental runs is modest, so performance estimates are sensitive to partition composition. Second, the PCA projection indicates partial overlap among the fault classes, suggesting that some degradation modes are not cleanly separable in the selected feature space. Third, the multiclass setting amplifies the effect of small shifts in decision boundaries when training subsets vary. For these reasons, the cross-validation result is considered the more representative estimate of model robustness, while the hold-out result is interpreted as an upper bound performance indication under one favourable partition. This behaviour highlights the possibility of optimistic performance estimates when evaluation is based on a single partition of a relatively small dataset, further

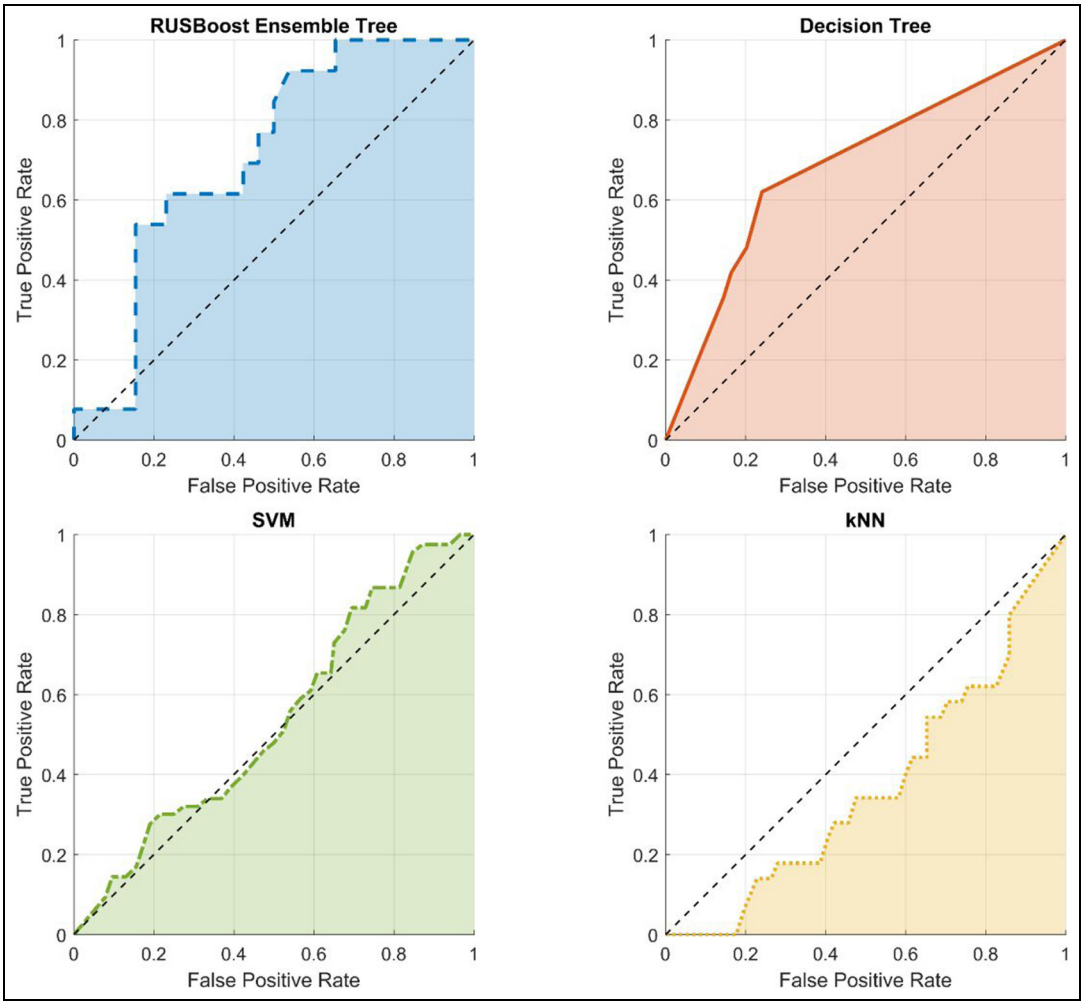


Figure 8. ROC curves with shaded AUC regions for the four classifiers under 10-fold cross validation.

Table 5. Comparison between hold out and cross validation performance.

Model	Hold out accuracies (%)	CV accuracies (%)	Hold out AUC	CV AUC
Ensemble tree	100	74	1.00	0.87
Decision tree	92	64	1.00	0.78
SVM	30	36	0.78	0.54
kNN	23	28	0.65	0.47

reinforcing the importance of repeated cross-validation for assessing model robustness and generalisation capability. In addition, the relatively small number of independent run-level samples may affect the statistical robustness of the reported performance estimates. Since the dataset was obtained from a single vehicle platform operating under controlled laboratory conditions, caution should be exercised when generalising the present results to different vehicle architectures, suspension geometries, payload conditions, terrain environments, or operational scenarios.

Direct comparison with existing studies is challenging because many published suspensions fault diagnosis frameworks rely on simulated datasets or component level experiments, that may produce more optimistic classification results. In contrast, the present study considers a substantially more challenging scenario involving full scale vehicle experiments, multi-class suspension fault classification, broadband excitation profiles, and independent run level validation. Therefore, the obtained 74% cross-validation accuracy represents competitive and practically meaningful performance while maintaining interpretability and deployment-oriented analysis.

Beyond classification accuracy, computational efficiency is a critical requirement for real time and

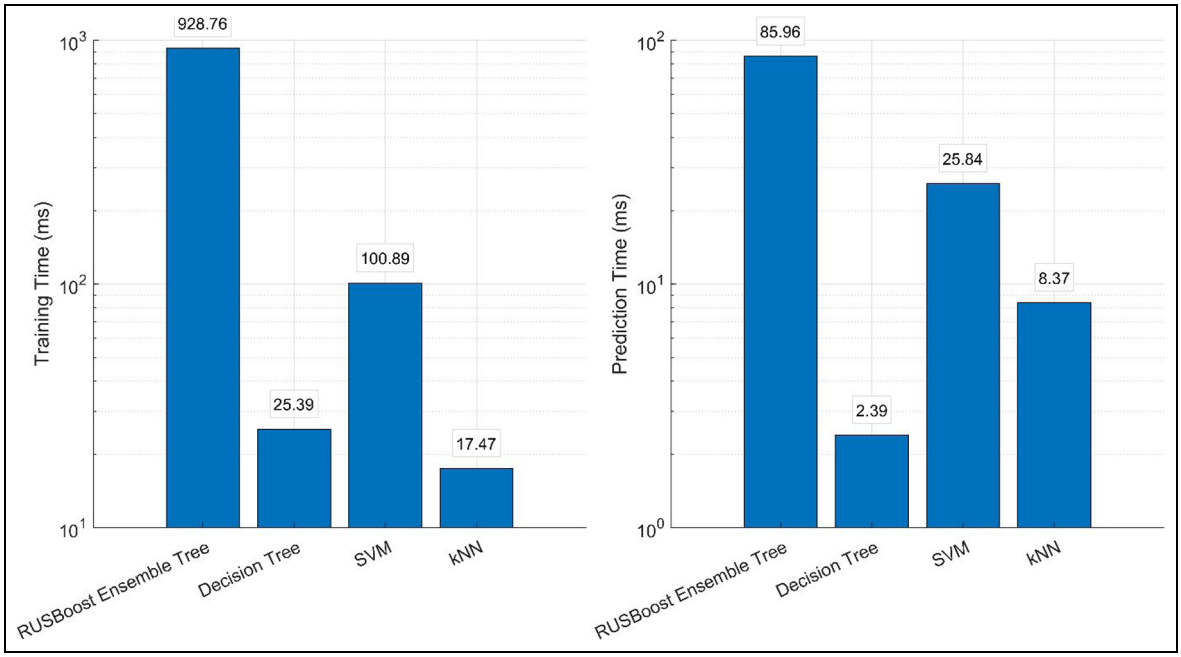


Figure 9. Comparison of training and prediction times for the evaluated classifiers.

onboard diagnostic applications. Figure 9 presents a comparative analysis of the training and prediction times of the four evaluated classifiers. The RUSBoost ensemble tree model exhibits the highest training time due to the iterative construction of multiple weak learners inherent to the boosting process. In contrast, the decision tree and kNN classifiers demonstrate significantly lower training times, reflecting their simpler model structures, while the SVM shows moderate computational demand associated with kernel optimization.

Despite its higher training cost, the ensemble model achieves competitive prediction efficiency in the present offline evaluation. Since training is performed offline, the higher training time does not directly constrain operational use. The measured inference time suggests that the framework is computationally promising for future embedded implementation; however, a full embedded deployment study, including processor specific timing, memory usage, and hardware level validation, remains outside the scope of the present work. The current results should therefore be interpreted as evidence of computational feasibility rather than as a complete embedded deployment demonstration.

Conclusion

This paper presented an embedded, explainable machine learning framework for fault detection in both defence crewed and uncrewed vehicle suspension systems, combining multichannel feature extraction with an ensemble-based classifier. The approach was experimentally validated using a full-scale Land Rover Defender 110 suspension assembly on a four-post hydraulic vibration rig operating under broadband excitation representative of defence mobility environments.

A single run hold out evaluation demonstrated the peak performance of the method, while the 10-fold cross validation assessment provided a more realistic estimate of generalisation under varying training conditions. The ensemble model delivered the most stable behaviour across both evaluations, maintaining consistent AUC values and demonstrating resilience to the high variance nature of suspension vibration data.

Feature importance analysis showed that peak-to-peak amplitude, standard deviation and RMS dominate diagnostic decisions, indicating that suspension degradation is primarily manifested through increased vibration excursion, dispersion, and transmitted energy. This aligns directly with physical degradation mechanisms such as reduced damping effectiveness, looseness, and elastomeric compliance. PCA analysis further confirmed partial class overlap, justifying the need for nonlinear ensemble-based decision boundaries.

Overall, the results indicate that the proposed ensemble learning framework provides an effective balance of accuracy, robustness, and interpretability, while remaining computationally efficient for embedded implementation. The method offers a reliable diagnostic layer that can be integrated with control architectures for defence mobility suspension systems.

Several limitations of the present study should be acknowledged. The experiments were conducted under controlled four post rig conditions rather than under fully operational field conditions, and the study was performed on a single vehicle platform. Although the rig environment provides repeatability and full-scale dynamic realism, broader validation across additional

platforms, fault severities, payload conditions, and terrain profiles will be necessary before general deployment can be claimed. In addition, relatively limited dataset comprising 48 independent run-level samples obtained from a single vehicle platform under controlled four-post rig laboratory conditions, may affect the statistical robustness of the reported performance estimates, particularly in the multiclass setting where decision boundaries are sensitive to partition variability. Consequently, broader validation using larger multi-vehicle datasets collected under diverse operational and terrain conditions will be necessary before operational generalisation can be fully established. Future work will therefore focus on extending the framework to a wider range of vehicles and operating conditions, incorporating richer explainability tools, and evaluating implementation on embedded hardware for in-vehicle deployment.

ORCID iDs

Irfan Ullah Khan https://orcid.org/0000-0001-8287-0365
John Economou https://orcid.org/0000-0002-3627-4949
David Galvão Wall https://orcid.org/0000-0001-9359-9822

Ethical considerations

This study involved experimental testing of mechanical systems only.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability statement

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1. Azadi S and Soltani A. Fault detection of vehicle suspension system using wavelet analysis. *Veh Syst Dyn* 2009; 47(4): 403–418. https://doi.org/10.1080/00423110802094298
2. Sert E and Boyraz P. Optimization of suspension system and sensitivity analysis for improvement of stability in a midsize heavy vehicle. *Eng Sci Technol Int J* 2017; 20(3): 997–1012. https://doi.org/10.1016/j.jestch.2017.03.007
3. Kawabe T, Isobe O, Watanabe Y, et al. New semi-active suspension controller design using quasi-linearization and frequency shaping. *Control Eng Pract* 1998; 6: 1183–1191. https://doi.org/10.1016/S0967-0661(98)00070-7
4. Fischer D and Isermann R. Mechatronic semi-active and active vehicle suspensions. *Control Eng Pract* 2003; 12: 1353–1367. https://doi.org/10.1016/j.conengprac.2003.08.003
5. Ieluzzi M, Turco P and Montiglio M. Development of a heavy truck semi-active suspension control. *Control Eng Pract* 2005; 14: 305–312. https://doi.org/10.1016/j.conengprac.2005.03.019
6. Poussot-Vassal C, Sename O, Dugard L, et al. A new semi-active suspension control strategy through LPV technique. *Control Eng Pract* 2008; 16(12): 1519–1534. https://doi.org/10.1016/j.conengprac.2008.05.002
7. Rabeno E and Bounds M. Condition-based maintenance of military ground vehicles. In: *Proceedings of the IEEE aerospace conference*, Big Sky, MT, USA, 7–14 March 2009, pp.1–6. IEEE. https://doi.org/10.1109/AERO.2009.4839683
8. Esperon-Miguez M, John P and Jennions IK. The effect of current military maintenance practices and regulations on the implementation of integrated vehicle health management technology. *IFAC Proc Vol* 2012; 45(31): 205–210. https://doi.org/10.3182/20121122-2-ES-4026.00011
9. Bond WG, Dozier H, Arnold TL, et al. A hybrid learning approach to prognostics and health management applied to military ground vehicles using time-series and maintenance event data. *Proc Annu Conf PHM Soc* 2020; 12(1): 1146. https://doi.org/10.36001/phmconf.2020.v12i1.1146
10. Zeng J and Liang Z. A dynamic predictive maintenance approach using probabilistic deep learning for a fleet of multi-component systems. *Reliab Eng Syst Saf* 2023; 238: 109456. https://doi.org/10.1016/j.ress.2023.109456
11. Burdzik R. A comprehensive diagnostic system for vehicle suspensions based on a neural classifier and wavelet resonance estimators. *Measurement* 2022; 200: 111602. https://doi.org/10.1016/j.measurement.2022.111602
12. Yan R, Gao RX and Chen X. Wavelets for fault diagnosis of rotary machines: a review with applications. *Signal Process* 2014; 96: 1–15. https://doi.org/10.1016/j.sigpro.2013.04.015
13. Jardine AKS, Lin D and Banjevic D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech Syst Signal Process* 2006; 20(7): 1483–1510. https://doi.org/10.1016/j.ymssp.2005.09.012
14. Lei Y, Jia F, Lin J, et al. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. *IEEE Trans Ind Electron* 2016; 63(5): 3137–3147. https://doi.org/10.1109/TIE.2016.2519325
15. Zhao R, Yan R, Chen Z, et al. Deep learning and its applications to machine health monitoring. *Mech Syst Signal Process* 2019; 115: 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050
16. Sokolova M and Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 2009; 45(4): 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

17. Seiffert C, Khoshgoftaar TM, Van Hulse J, et al. RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Trans Syst Man Cybern A* 2010; 40(1): 185–197. https://doi.org/10.1109/TSMCA.2009.2029559
18. Galar M, Fernandez A, Barrenechea E, et al. A review on ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches. *IEEE Trans Syst Man Cybern C Appl Rev* 2012; 42(4): 463–484. https://doi.org/10.1109/TSMCC.2011.2161285
19. Ata WG and Salem AM. Experimental characterisation and nonparametric modeling of armoured vehicle hydraulic shock absorber under periodic excitations. *J Vib Eng Technol* 2025; 13: 311–315. https://doi.org/10.1007/s42417-025-01830-9
20. Barredo Arrieta A, Diaz-Rodriguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 2020; 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
21. Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Comput Surv* 2018; 51(5): 93. https://doi.org/10.1145/3236009
22. Brito LC, Susto GA, Brito JN, et al. Fault diagnosis using eXplainable AI: a transfer learning-based approach for rotating machinery exploiting augmented synthetic data. *Expert Syst Appl* 2023; 232: 120860. https://doi.org/10.48550/arXiv.2210.02974.
23. Maged A, Haridy S and Shen H. Explainable artificial intelligence techniques for accurate fault detection and diagnosis: a review. *Expert Syst Appl* 2024; 2404: 1–44. https://doi.org/10.48550/arXiv.2404.11597
24. Becker M, Niemietz P and Bergs T. Explainable neural network for time series-based condition monitoring in sheet metal shearing. *J Intell Manuf* 2025; 37: 1247–1263. https://doi.org/10.1007/s10845-025-02596-3
25. Sharma J, Mittal ML and Soni G. Condition-based maintenance using machine learning and role of interpretability: a review. *Int J Syst Assur Eng Manage* 2022; 15(4): 1345–1360. https://doi.org/10.1007/s13198-022-01843-7
26. Koutsoupakis J, Giagopoulos D and Chatziparasidis I. AI-based condition monitoring on mechanical systems using multibody dynamics models. *Eng Appl Artif Intell* 2023; 123: 106467. https://doi.org/10.1016/j.engappai.2023.106467
27. Gangsar P, Bajpei AR and Porwal R. A review on deep learning based condition monitoring and fault diagnosis of rotating machinery. *Noise Vibr Worldw* 2022; 53(11): 550–578. https://doi.org/10.1177/09574565221139638
28. Ochando FJ, Cantero A, Guerrero JI, et al. Data acquisition for condition monitoring in tactical vehicles: onboard computer development. *Sensors* 2023; 23(12): 5645. https://doi.org/10.3390/s23125645



https://dspace.org/

Embedded fault detection for crewed and uncrewed defence vehicle suspension systems via explainable machine learning and four post rig testing

Khan, Irfan; Economou, John T.; Sandhu, Manik Preet Singh; Sandle, Adam; Cooper, Robert ...

2026-12-31
Sage
https://doi.org/10.1177/09544070261463329
Khan IU, Economou J, Sandhu MPS, et al., (2026) Embedded fault detection for crewed and uncrewed defence vehicle suspension systems via explainable machine learning and four post rig testing. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, Available online 6 July 2026

Downloaded from a DSpace repository. https://dspace.org/