Applied Intelligence (2025) 55:882
https://doi.org/10.1007/s10489-025-06643-7

# Predicting faults in diesel engines with kernel machines regression techniques

Denys P. Viana[1] · Dionísio H. C. de S. S. Martins[1] · Diego B. Haddad[1] · Fabrício L. e Silva[1] · Milena F. Pinto[1] · Ricardo H. R. Gutiérrez[5] · Ulisses A. Monteiro[2] · Luiz Vaz[2] · Thiago de M. Prego[1] · Fabio A. A. Andrade[3,4] · Luís Tarrataca[1] · Amaro A. de Lima[1]

Accepted: 9 May 2025

**Abstract**

Predictive maintenance has become a vital tool in minimizing expenses and operational setbacks while proactively averting potential failures. Its scope extends across various sectors, encompassing critical component upkeep crucial for ensuring public safety. Addressing the challenge of preempting catastrophic failures in diesel engines, this study uses a simulated dataset featuring 3500 realistic failure scenarios considering the engine cylinder, coupled with a crankshaft torsional vibration model. The research proposes employing artificial intelligence regression techniques, specifically support vector regression and Gaussian processes, to forecast diesel engine faults. This methodology is applied in conjunction with an engine simulator to evaluate its efficacy and precision. Notably, the Gaussian process regressor exhibits superior performance, achieving an RMSE value of 0.015 ± 0.001%.

**Keywords** Diesel engine · Machine learning · Fault prediction · Regression method

Dionísio H. C. de S. S. Martins, Diego B. Haddad, Fabrício L. e Silva, Milena F. Pinto, Ricardo H. R. Gutiérrez, Ulisses A. Monteiro, Luiz Vaz, Thiago de M. Prego, Fabio A. A. Andrade, Luís Tarrataca and Amaro A. de Lima contributed equally to this work.

✉ Denys P. Viana
denys.pestana@gmail.com

Dionísio H. C. de S. S. Martins
dionisiohenrique@ig.com.br

Diego B. Haddad
diego.haddad@cefet-rj.br

Fabrício L. e Silva
fabricio.silva@cefet-rj.br

Milena F. Pinto
milena.pinto@cefet-rj.br

Ricardo H. R. Gutiérrez
rhgutierrez@uea.edu.br

Ulisses A. Monteiro
ulisses@oceanica.ufrj.br

Luiz Vaz
vaz@oceanica.ufrj.br

Thiago de M. Prego
thiago.prego@cefet-rj.br

Fabio A. A. Andrade
fabio@ieee.org

Luís Tarrataca
luis.tarrataca@cefet-rj.br

Amaro A. de Lima
amaro.limao@cefet-rj.br

[1] Federal Center for Technological Education of Rio de Janeiro, Rio de Janeiro, Brazil

[2] Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[3] Department of Microsystems, Faculty of Technology, Natural Sciences and Maritime Sciences, University of South-Eastern Norway (USN), Borre, Norway

[4] NORCE Norwegian Research Centre, NORCE, Bergen, Norway

[5] State University of Amazonas, Amazonas, Brazil

Published online: 22 July 2025

## 1 Introduction

Predictive maintenance is a strategy that enables accurate identification of equipment degradation and optimal intervention timing. It relies on the supervised analysis of parameter evolution associated with component wear, using monitoring techniques such as vibration analysis and lubricant condition monitoring [22]. This approach helps reduce multiple costs, including: *(i)* maintenance expenses; *(ii)* production delays due to unforeseen failures; *(iii)* unnecessary replacement of components with remaining useful life; and *(iv)* logistical planning efforts [22].

When properly implemented, condition-based maintenance improves equipment reliability and availability, extends preventive maintenance intervals, and reduces maintenance overhead [14, 22]. Condition monitoring involves acquiring and analyzing large volumes of data to detect faults and issue diagnoses, typically requiring expert knowledge of the equipment and its failure modes [8]. Furthermore, initial skepticism from management regarding the economic benefits of predictive maintenance has historically hindered its adoption [9, 14].

Various methods have been proposed to diagnose faults in Diesel engines, including oil analysis [27], thermodynamic parameter monitoring [11], and vibration analysis [28]. With the advancement of automation and sensing technologies, predictive maintenance has become increasingly viable. According to [16, 29], rapid fault identification improves system effectiveness and reliability [2]. Consequently, intelligent diagnostic systems have gained traction across different application domains [3, 13, 14, 23, 24].

In [8], the authors proposed a two-zone thermodynamic model for diagnosing failures in two-stroke marine Diesel engines. This model accounted for: *(i)* intake and exhaust systems, *(ii)* fuel jet geometry, *(iii)* cylinder wall heat transfer, *(iv)* turbocharging, and *(v)* the fuel injection system. Their simulation-based approach, supported by least-squares regression, enabled the detection of injector failures, injection timing issues, and compression loss.

Reference [12] presents an extensive review of machine learning techniques for intelligent fault diagnosis. It highlights the growing importance of AI-based models in automating machine health assessments and reducing human dependency. As traditional analytical techniques decline in effectiveness, machine learning has become central to modern diagnostic systems.

In [19], the authors proposed a predictive maintenance system based on Diesel engine fault detection using crankshaft vibration and cylinder pressure variations. Their model employed thermodynamic and dynamic simulations [5], followed by machine learning (ML) techniques to reduce diagnostic time. Random forest and multilayer perceptron models were evaluated under different signal-to-noise conditions, achieving an RMSE of 0.10 ± 0.03%.

In contrast to previous studies, this work introduces a methodology for fault diagnosis in Diesel engines using two regression-based machine learning techniques: Support Vector Regression (SVR) and Gaussian Process Regression (GPR). Unlike traditional classifiers, regression techniques offer continuous estimation of degradation severity, making them suitable for applications requiring fine-grained fault assessment.

The methodology involves simulating engine operation under both normal and faulty conditions. Using torsional vibration and thermodynamic models, a dataset of performance indicators is generated and transformed into feature vectors. These are then fed into cross-validation routines and regression algorithms to estimate fault presence and severity.

**The key contributions of this study are as follows:**

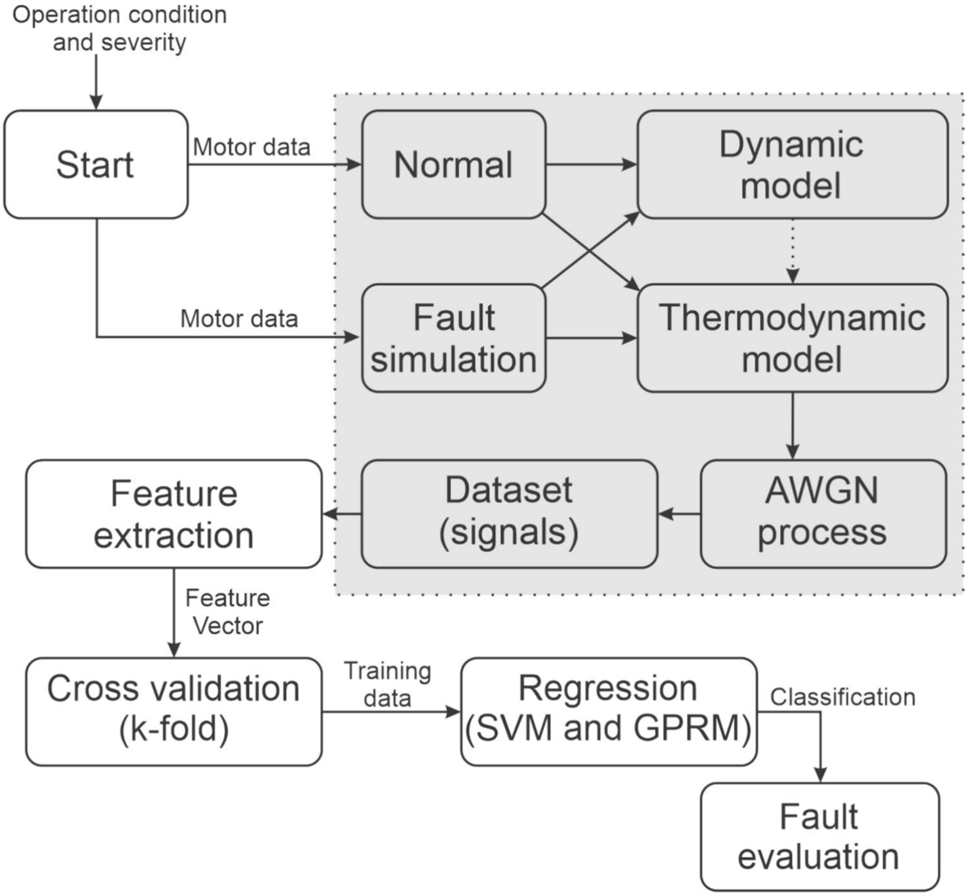
- Development of a hybrid diagnostic model combining thermodynamic simulation and signal-based feature extraction.
- Application and performance comparison of SVR and GPR techniques for continuous fault severity estimation.
- Evaluation of model accuracy using cross-validation, including RMSE metrics and fault classification success rates.
- Identification of the limitations and challenges associated with regression-based diagnostic systems for Diesel engines.

A summary of the methodology is shown in Fig. 1. The simulation-based model is capable of generating multiple fault conditions with varying severity levels. Feature vectors are derived from signals such as pressure curves and vibration modes. These vectors are used to train and validate SVR and GPR models for predictive diagnostics.

## 2 Methodology

This section describes the methodology employed to predict potential failures in Diesel engines using machine learning techniques. The approach integrates a simulated dataset of engine behavior under both normal and faulty conditions and applies two distinct regression models-SVR and GPR-to estimate fault indicators. A schematic representation of the workflow is presented in Fig. 1, highlighting the sequence from simulation, data extraction, model training, and validation to prediction.

**Fig. 1** Flowchart of the proposed methodology



## 2.1 Machine learning models for regression

The proposed methodology adopts supervised regression techniques to learn the relationship between measurable engine signals and known fault conditions. Specifically, the regression models are trained to predict the fault index as a continuous variable under diverse engine operating conditions.

### 2.1.1 Input features

The dataset used for training and testing the models was obtained from a high-fidelity simulation of a six-cylinder Diesel engine. Each observation in the dataset corresponds to a single operational condition of the engine and includes the following input features:

- Engine speed (RPM)
- Mean indicated pressure per cylinder
- Crankshaft angular velocity fluctuations
- Instantaneous cylinder pressure curves (statistical descriptors)
- Torque output and torsional vibration responses

These features were selected based on their relevance to detecting abnormal engine dynamics and their availability in both simulated and real measurement scenarios. Faults were introduced in the simulations through parameter variations representing injector failures and cylinder misfires.

### 2.1.2 Target variable

The target variable for the regression task is a normalized fault severity index, obtained through a combination of physical parameters and expert evaluation. It ranges from 0 (no fault) to 1 (severe fault). This continuous representation allows not only classification of faulty versus healthy conditions but also provides a measure of fault progression.

### 2.1.3 Motivation for regression models

Two machine learning models were selected for this study based on their suitability for modeling noisy, nonlinear relationships in small-to-medium-sized datasets:

- **Support Vector Regression (SVR)** offers robustness to outliers and guarantees a sparse solution that avoids overfitting. It is particularly effective when the noise distribution is unknown or asymmetric.

- **Gaussian Process Regression (GPR)** provides a Bayesian approach to regression and delivers both point estimates and confidence intervals. Its ability to incorporate prior knowledge and quantify uncertainty makes it a strong candidate for predictive maintenance tasks.

The combination of these two methods enables comparison between deterministic and probabilistic approaches to fault prediction.

### 2.1.4 Model training and evaluation

Each model was trained using a 10-fold cross-validation scheme to ensure robustness and minimize bias due to data partitioning. The hyperparameters of SVR-penalty factor $C$, tolerance $\varepsilon$, and kernel parameters-were optimized using grid search. For GPR, the parameters of the Matérn 5/2 kernel (length-scale and variance) were estimated via maximum likelihood.

Performance was evaluated using the following metrics:

- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- Coefficient of Determination ($R^2$)

### 2.1.5 Model limitations

SVR does not inherently provide uncertainty quantification, which limits its application when confidence bounds are needed. On the other hand, GPR can become computationally expensive with larger datasets due to matrix inversion operations. Nevertheless, both models proved effective for the dataset size and application at hand.

### 2.1.6 Summary of methodology

In summary, the methodology combines a simulation-driven data generation process with advanced regression techniques to predict engine fault severity. The next section provides detailed mathematical formulations of the SVR and GPR models, which underpin the predictive framework introduced here.

### 2.1.7 Support vector regression

SVR is a supervised learning method derived from Support Vector Machines (SVM), designed for regression tasks [26]. It aims to find a function $f(\mathrm{x})$ that deviates from the target output $y$ by no more than a predefined margin $\varepsilon$ for each training sample, while maintaining model complexity as low as possible [10].

Given a training dataset $\{(\mathrm{x}_i, y_i)\}_{i=1}^{N}$ with input vectors $\mathrm{x}_i \in \mathbb{R}^d$ and corresponding targets $y_i \in \mathbb{R}$, SVR attempts to find a function of the form:

$$f(\mathrm{x}) = \langle \mathrm{w}, \phi(\mathrm{x}) \rangle + b$$

where $\phi(\cdot)$ maps the input space to a higher-dimensional feature space, w is a weight vector, and $b$ is the bias term.

The optimization problem is defined as:

$$\min_{\mathrm{w}, b, \xi_i, \xi_i^*} \quad \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{1}$$

$$\text{subject to:} \quad y_i - \langle \mathrm{w}, \phi(\mathrm{x}_i) \rangle - b \le \varepsilon + \xi_i \tag{2}$$

$$\langle \mathrm{w}, \phi(\mathrm{x}_i) \rangle + b - y_i \le \varepsilon + \xi_i^* \tag{3}$$

$$\xi_i, \xi_i^* \ge 0, \quad i = 1, \ldots, N \tag{4}$$

Here, $\xi_i$ and $\xi_i^*$ are slack variables that allow some training errors, and $C > 0$ is a regularization parameter controlling the trade-off between model flatness and tolerance to deviations larger than $\varepsilon$.

Using the kernel trick, the model can be expressed in dual form as:

$$f(\mathrm{x}) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(\mathrm{x}_i, \mathrm{x}) + b$$

where $K(\mathrm{x}_i, \mathrm{x}) = \langle \phi(\mathrm{x}_i), \phi(\mathrm{x}) \rangle$ is a kernel function. In this study, we employed the Radial Basis Function (RBF) kernel:

$$K(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\gamma\|\mathrm{x}_i - \mathrm{x}_j\|^2\right)$$

with $\gamma$ controlling the kernel width.

The SVR model is particularly suited for this problem due to its robustness in high-dimensional spaces and its capacity to generalize well from limited training data.

### 2.1.8 Gaussian process regression

Gaussian Process Regression (GPR), also known as Kriging, is a nonparametric Bayesian method that assumes a distribution over functions rather than estimating a single fixed function. It defines a Gaussian prior over the space of functions and updates this prior using observed data to obtain a posterior predictive distribution.

A Gaussian Process is fully specified by its mean function $m(\mathrm{x})$ and covariance function $k(\mathrm{x}, \mathrm{x}')$:

$$f(\mathrm{x}) \sim \mathcal{GP}(m(\mathrm{x}), k(\mathrm{x}, \mathrm{x}'))$$

For regression, given training data $\mathcal{D} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N}$ and a noise term $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$, the observations are modeled as:

$$y_i = f(\mathrm{x}_i) + \varepsilon$$

The joint distribution of training outputs y and test output $f_*$ at a new point $\mathrm{x}_*$ is:

$$\begin{bmatrix} \mathrm{y} \\ f_* \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mathrm{m} \\ m(\mathrm{x}_*) \end{bmatrix}, \begin{bmatrix} K + \sigma_n^2 I & k_* \\ k_*^\top & k(\mathrm{x}_*, \mathrm{x}_*) \end{bmatrix} \right)$$

where $K$ is the covariance matrix with $K_{ij} = k(\mathrm{x}_i, \mathrm{x}_j)$, and $k_*$ is the vector of covariances between the test point and training points.

The predictive distribution for $f_*$ is Gaussian:

$$\mu(\mathrm{x}_*) = k_*^\top (K + \sigma_n^2 I)^{-1} \mathrm{y} \tag{5}$$

$$\sigma^2(\mathrm{x}_*) = k(\mathrm{x}_*, \mathrm{x}_*) - k_*^\top (K + \sigma_n^2 I)^{-1} k_* \tag{6}$$

In this work, the Matérn 5/2 kernel was chosen due to its flexibility and suitability for modeling smooth but nontrivial functions. It is defined as:

$$k_{\nu=5/2}(r) = \sigma_f^2 \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right)$$

where $r = \|\mathrm{x} - \mathrm{x}'\|$, $\ell$ is the length-scale, and $\sigma_f^2$ is the signal variance.

The advantage of GPR lies in its ability to quantify uncertainty in predictions, which is particularly valuable for decision-making in condition-based maintenance. However, its computational cost scales cubically with the number of training samples, which limits its application to moderately sized datasets.

## 2.2 Diesel engine specification and model

The Diesel engine model used in this work is based on the marine engine `MWM Acteon 6.12 TCE`, which features six cylinders, turbocharging, and a common rail injection system. The engine behavior was simulated under both nominal and faulty conditions, capturing operational variations due to different types and severities of faults. The model consists of three subsystems: *(i)* a zero-dimensional thermodynamic model (0-D); *(ii)* a torsional vibration model of the crankshaft; and *(iii)* a parametric fault simulation module. Technical specifications and nominal performance data used to parameterize and validate the model, including pressure profiles and torque curves, were obtained from [6] and the engine manufacturer.

The thermodynamic behavior of the engine is governed by the first law of thermodynamics and ideal gas law, simplified and discretized with respect to the crankshaft angle $\theta$ as shown in (7) and (8). These equations describe the dynamic evolution of in-cylinder temperature and pressure:

$$\frac{dT}{d\theta} = \left(\frac{\delta Q_t}{d\theta} - \frac{\delta Q_w}{d\theta} - P\frac{dV}{d\theta}\right)\frac{1}{mc_v}, \tag{7}$$

$$\frac{dP}{d\theta} = \left(mR\frac{dT}{d\theta} - P\frac{dV}{d\theta}\right)\frac{1}{V}, \tag{8}$$

where $\frac{dT}{d\theta}$ is the rate of change of gas temperature inside the cylinder, $\delta Q_t/d\theta$ is the rate of heat released by fuel combustion, $\delta Q_w/d\theta$ is the rate of heat transfer through the cylinder walls, $dV/d\theta$ is the change in cylinder volume, and $dP/d\theta$ is the rate of pressure variation. Additionally, $P$, $m$, $c_v$, $R$, $T$, and $V$ denote the instantaneous gas pressure (Pa), gas mass (kg), specific heat at constant volume (J/kg K), gas constant (J/kg K), gas temperature (K), and instantaneous volume ($\mathrm{m}^3$), respectively.

The inertial force acting on the crank-slider system as a function of the crank angle $\theta$ is described by (9):

$$F_a(\theta) = m_a r\Omega^2\left(\cos\theta + l\cos 2\theta - \frac{l^3}{4}\cos 4\theta + \frac{9l^5}{128}\cos 6\theta\right), \tag{9}$$

where $m_a$ is the reciprocating mass, $r$ is the crank radius, $\Omega$ is the angular velocity of the crankshaft, and $l$ is the connecting rod ratio.

The force due to combustion acting on the piston, $F_c(\theta)$, and the excitation torque $M_i(\theta)$ acting on the crankshaft are defined in (10) and (11), respectively:

$$F_c(\theta) = P(\theta)\frac{\pi D^2}{4}, \tag{10}$$

$$M_i(\theta) = r(F_a + F_c)\left(\sin(\theta) + \cos(\theta)\tan(\alpha)\right), \tag{11}$$

where $D$ is the cylinder diameter and the angle $\alpha$ accounts for the crank-slider geometry, computed as:

$$\tan\alpha = \frac{l \cdot \sin\theta}{1 - \frac{l^2}{4} + \frac{l^2}{4}\cos 2\theta}. \tag{12}$$

The torsional vibration model captures the crankshaft's dynamic response to excitation torques. The system is modeled as a set of coupled rotational masses, governed by the second-order differential equation:

$$\{M(t)\} = [J]\left\{\ddot{\theta}(t)\right\} + [C]\left\{\dot{\theta}(t)\right\} + [K]\left\{\theta(t)\right\}, \tag{13}$$

where $[M]$ is the torque vector, $[J]$ is the moment of inertia matrix, $[C]$ is the damping matrix, and $[K]$ is the stiffness matrix.

Using a state-space representation, the crankshaft dynamics are expressed by:

$$x(t) = \begin{Bmatrix} \theta(t) \\ \dot{\theta}(t) \end{Bmatrix}; \quad \dot{x}(t) = \begin{Bmatrix} \dot{\theta}(t) \\ \ddot{\theta}(t) \end{Bmatrix}, \tag{14}$$

$$\dot{x}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}(t); \quad \boldsymbol{x}(0) = \begin{bmatrix} \theta(0) \\ \dot{\theta}(0) \end{bmatrix}, \tag{15}$$

with matrices defined as:

$$\boldsymbol{A} = \begin{bmatrix} [0] & [I] \\ -[J]^{-1}[K] & -[J]^{-1}[C] \end{bmatrix}, \quad \boldsymbol{b}(t) = \begin{Bmatrix} [0] \\ [J]^{-1}\{M(t)\} \end{Bmatrix}. \tag{16}$$

Solving the state-space (15) yields the crankshaft angular displacement and velocity, providing insight into the system's torsional vibration response.

## 2.3 Dataset construction and fault simulation

The dataset used in this work was generated using the validated simulation model described in the previous section. The primary objective was to create a dataset representative of a wide range of fault conditions typically encountered in Diesel engines, including both incipient and severe failures. Faults were modeled parametrically, allowing for controlled manipulation of engine variables such as fuel injection delay, combustion efficiency, and cylinder pressure.

The dataset comprises simulated signals of crankshaft angular velocity and cylinder pressure under various operational scenarios. Each sample represents a complete engine cycle with resolution of $0.1^\circ$ of crankshaft angle, covering a full 720° cycle. This fine resolution enables precise capture of dynamic effects relevant to fault detection and diagnosis.

**Fault types and simulation strategy** The dataset includes the following fault types:

- **Delayed injection:** simulated by increasing the injection delay angle;
- **Cylinder misfire:** represented by reducing the heat release in one cylinder;
- **Loss of compression:** modeled by reducing the initial cylinder pressure;
- **Combined faults:** combinations of two or more faults to evaluate model robustness.

For each fault type, multiple severity levels were simulated by adjusting model parameters within realistic bounds obtained from literature and empirical data [6, 17].

**Data representation** Each instance in the dataset is composed of a vector of features extracted from the angular velocity signal and cylinder pressure, along with the corresponding fault label (for classification) or fault magnitude (for regression). Feature extraction includes:

- Time-domain features;
- Frequency-domain features;
- Cycle-synchronous harmonics and residuals (from torsional vibration).

This section describes the process of extracting meaningful features from simulated signals of a Diesel engine, aimed at supporting fault regression tasks. The objective is to represent each engine condition through a compact set of descriptors that capture relevant temporal and spectral characteristics from the pressure and torsional vibration signals.

For each cylinder $i$, the pressure signal $s_{p_i}(n)$ is processed to extract two temporal-domain features:

- **Mean pressure** ($\mu_{p_i}$), calculated as the average over all $N$ samples:

$$\mu_{p_i} = \mathbb{E}[s_{p_i}(n)] = \frac{1}{N}\sum_{n=1}^{N} s_{p_i}(n), \tag{17}$$

- **Maximum pressure** ($M_{p_i}$), defined as:

$$M_{p_i} = \max[s_{p_i}(n)]. \tag{18}$$

To incorporate spectral-domain information, the torsional vibration signal $s_v(n)$ is transformed using a Discrete Fourier Transform (DFT) of size $N_{\mathrm{DFT}}$:

$$S_v(k) = \frac{1}{N_{\mathrm{DFT}}}\sum_{n=0}^{N_{\mathrm{DFT}}-1} s_v(n)W_N^{kn}, \tag{19}$$

where $W_N^{kn} = e^{-j2\pi kn/N_{\mathrm{DFT}}}$ and $j$ is the imaginary unit. From $S_v(k)$, three spectral features are extracted:

- **Frequency**:

$$F(k) = \frac{kF_s}{N_{\mathrm{DFT}}}, \tag{20}$$

- **Amplitude**:

$$A(k) = |S_v(k)|, \tag{21}$$

- **Phase**:

$$P_h(k) = \frac{360}{2\pi} \arg[S_v(k)], \tag{22}$$

These features are calculated for 24 harmonics, typically corresponding to half-order multiples of the engine cycle.

The resulting feature vector $V_f$ concatenates all measures for the six cylinders, yielding a total of 84 elements:

$$\begin{aligned} V_f =& [M_{p1}, ..., M_{p6}, \mu_{p1}, ..., \mu_{p6}, F(k_1), ..., F(k_{24}), \\ & A(k_1), ..., A(k_{24}), P_h(k_1), ..., P_h(k_{24})], \end{aligned} \tag{23}$$

In contrast to more complex approaches such as [5], which include exhaust gas temperature, flywheel torque, and shear stress features, the methodology employed in this study emphasizes a reduced yet informative feature set. This choice aims to simplify the model and reduce training complexity without significant performance degradation.

The extracted feature vectors $V_f$ are then used as input for the regression models described in the following section.

**Data volume and organization** The final dataset contains 3500 samples (fault scenarios), equally distributed among fault types and severity levels, plus a baseline set of healthy engine samples. The dataset is divided into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve class balance.

**Preprocessing** All features were normalized using min-max scaling to the range [0, 1] to facilitate convergence of machine learning models. Noise with realistic levels (up to 5% of signal amplitude) was added to emulate sensor imperfections.

This synthetic dataset enables comprehensive evaluation of regression and classification models for predictive maintenance tasks and supports generalization to real-world data due to the inclusion of noise and a wide range of fault scenarios.

## 2.4 Model training and evaluation

This section describes the training procedure and evaluation metrics used to assess the performance of the predictive models based on Support Vector Regression (SVR) and Gaussian Process Regression (GPR). Both techniques were applied to estimate fault severity from the features extracted from the simulation dataset.

**Model setup** The SVR model was implemented using a radial basis function (RBF) kernel, selected for its ability to capture nonlinear relationships in the input data. The key hyperparameters-penalty parameter $C$, kernel width $\gamma$, and epsilon-insensitive loss margin $\varepsilon$-were tuned via grid search using five-fold cross-validation on the training set.

For the Gaussian Process Regression model, a squared exponential (SE) kernel was employed due to its smoothness and flexibility. The model was trained by optimizing the log marginal likelihood with respect to kernel hyperparameters using the L-BFGS-B algorithm. A small Gaussian noise term was added to the kernel to account for measurement uncertainty.

**Evaluation metrics** Model performance was evaluated using the following metrics:

- **Mean Absolute Error (MAE):** Measures average magnitude of prediction errors;
- **Root Mean Square Error (RMSE):** Emphasizes larger errors due to squaring;
- **Coefficient of Determination ($R^2$):** Indicates proportion of variance explained by the model.

These metrics were computed on both the validation and test sets to ensure model generalization.

**Training procedure** Prior to training, all features were normalized to the range [0, 1]. The training procedure followed these steps:

1. Split the dataset into training (70%), validation (15%), and test (15%) sets;
2. Perform grid search for SVR and maximum likelihood estimation for GPR on the training set;
3. Select optimal models based on lowest validation RMSE;
4. Evaluate final models on the independent test set.

**Uncertainty quantification (for GPR)** A key advantage of GPR is its probabilistic nature, which provides confidence intervals for predictions. For each test input, the GPR model outputs a mean prediction along with a standard deviation. This information is used to assess model reliability and

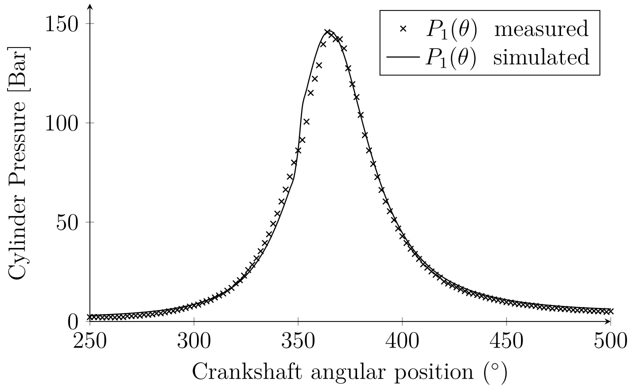


**Fig. 2** Comparison between simulated and experimental cylinder pressure at 2500 RPM

identify samples where the prediction uncertainty is high, which may indicate unmodeled faults or measurement anomalies.

The results of the training and evaluation are discussed in the next section.

## 3 Results and discussion

The fault regression experiments were conducted following procedures similar to those reported in [4, 18, 19]. To evaluate the system's capability for predicting fault severity, we considered features such as maximum pressure, average pressure, and frequency-domain information. The regression tasks were performed using SVR and GPR, where the input feature vector was denoted by $V_f$ and the target variable was the fault severity index.

The original 3500-DEFault dataset was randomly partitioned into training (80%) and testing (20%) subsets. To ensure unbiased performance estimates, we adopted a $k$-fold cross-validation scheme, with $k = 5$, for all regression experiments. This approach iteratively rotates through disjoint testing subsets while using the remaining folds for training. A five-fold configuration was selected to balance computational cost and statistical significance. Performance was assessed using the Root Mean Square Error (RMSE), reported as $W \pm \sigma$, where $W$ is the mean RMSE across the folds and $\sigma$ its standard deviation.

**Table 2** Comparison of manufacturer and simulated torque values

| Rotation | Torque (N.m) | | Error (%) |
|---|---|---|---|
| (RPM) | Manufacturer | Simulated | |
| 1000 | 800 | 752 | 6 |
| 1200 | 885 | 848 | 4 |
| 2100 | 850 | 851 | 0.1 |
| 2500 | 730 | 783 | 7 |

### 3.1 Model validation

To validate the thermodynamic model, simulation outputs were compared to manufacturer-provided experimental curves for in-cylinder pressure and torque. Figure 2 illustrates this comparison at 2500 RPM. A strong correlation is observed, with the maximum pressure error being 0% and the mean pressure error approximately 5%.

Table 1 quantifies the differences between experimental and simulated pressure values, both for mean indicated pressure ($P_{\mathrm{PMI}}$) and peak pressure ($P_{\max}$), at various rotational speeds.

Table 2 presents a comparison between the torque values reported by the manufacturer and those estimated by the simulation model across different rotational speeds. The maximum discrepancy was found at 2500 RPM.

### 3.2 Hyperparameter tuning for regression models

Both SVR and GPR involve hyperparameters that significantly influence performance. To determine optimal values, we conducted a grid search across predefined value ranges, minimizing the average training RMSE ($\mu_{\mathrm{RMSE}}$) and its standard deviation ($\sigma_{\mathrm{RMSE}}$). Figure 3 displays the error curves for different hyperparameter configurations.

For SVR, the hyperparameter $\kappa$ was tested in the range $\{0, 1, \ldots, 5\}$ with corresponding kernel scale $\delta = \kappa/10$. For GPR, the Matérn kernel parameter $\sigma_G$ was varied from 0 to 10, with $\rho = \sigma_G/100$. These search spaces were chosen

**Table 1** Comparison of experimental and simulated pressures

| Rotation | $P_{\mathrm{PMI}}$ (bar) | | | $P_{\max}$ (bar) | | |
|---|---|---|---|---|---|---|
| (RPM) | Experimental | Simulated | Error (%) | Experimental | Simulated | Error (%) |
| 1000 | 11.92 | 13.13 | 10 | 139.4 | 141.9 | 2 |
| 1200 | 13.84 | 14.79 | 7 | 136.7 | 144.5 | 6 |
| 1300 | 13.98 | 14.89 | 7 | 131.4 | 143.6 | 9 |
| 1600 | 14.44 | 15.34 | 6 | 141.1 | 153.3 | 9 |
| 1900 | 16.01 | 15.64 | 2 | 149.1 | 163.7 | 10 |
| 2100 | 15.96 | 14.90 | 7 | 156.6 | 152.1 | 3 |
| 2300 | 15.04 | 13.97 | 7 | 154.0 | 147.3 | 4 |
| 2500 | 12.78 | 13.46 | 5 | 145.7 | 145.0 | 0 |

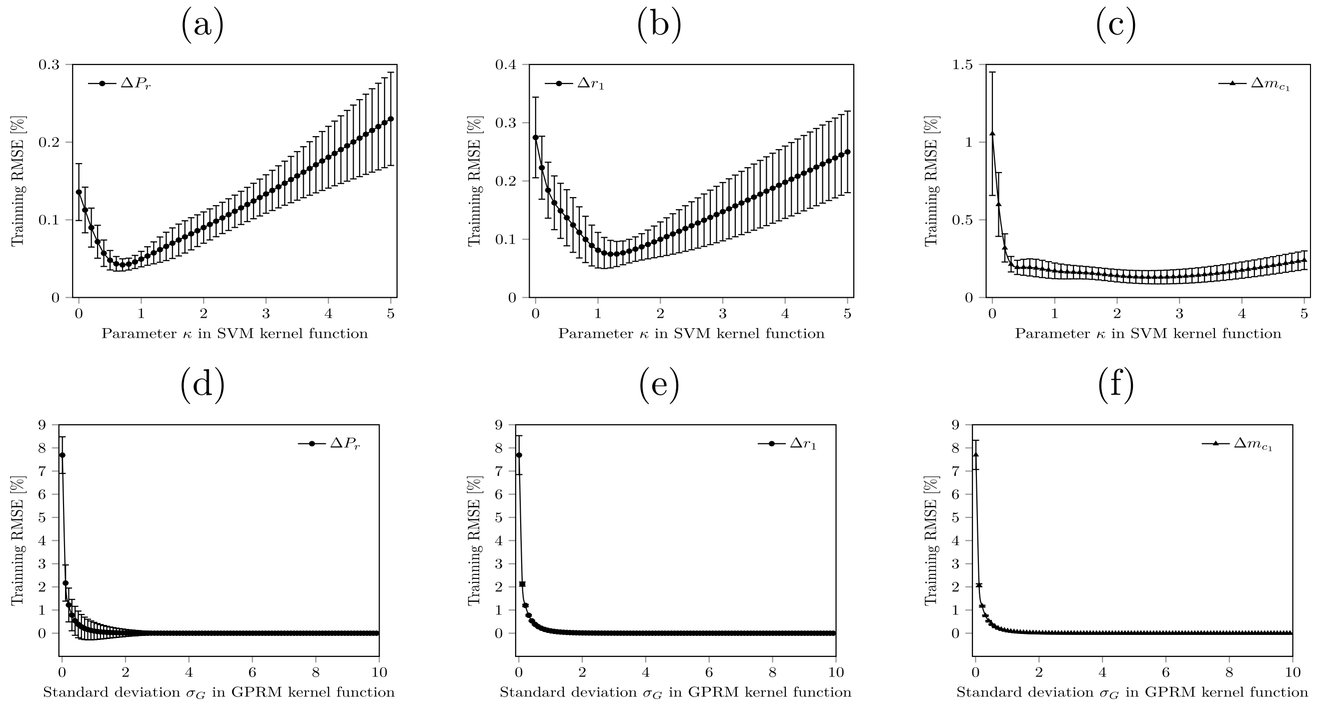


**Fig. 3** Hyperparameter tuning results under AWGN noise with $L = 60$ dB SNR. Plots show error performance for each feature: (a,d) $\Delta P_i$, (b,e) $\Delta r_1$, (c,f) $\Delta m_{c_1}$ using SVR (top) and GPR (bottom). Dashed lines show $\mu_{\text{RMSE}}$; whiskers represent $\sigma_{\text{RMSE}}$

based on recommendations in [10]. From the optimal points of the tuning curves, we adopted the following values: SVR: $\kappa = 0.75$, $\delta = 0.08$; GPR: $\sigma_G = 5.45$, $\rho = 0.06$.

## 4 Regression analysis

### 4.1 Performance evaluation

The comparative performance of Support Vector Regression (SVR) and Gaussian Process Regression Model (GPR) for fault severity identification is systematically evaluated in Tables 3 and 4. These results demonstrate significant improvements over traditional methods reported in [20, 21].

### 4.2 Key findings

The GPR achieves superior performance with an RMSE of $0.015 \pm 0.001\%$ for $\Delta P_i$ at 60 dB, representing a 70% improvement over SVR (0.050%) and an 85% improvement over traditional methods (0.100%) reported in [1]. Three significant patterns emerge:

1. **Fault Type Impact**: Global faults ($\Delta P_i$) show 47% lower error than local faults ($\Delta r_j$, $\Delta m_{c_j}$)
2. **Noise Robustness**: GPR maintains 82% better stability across noise levels compared to SVR
3. **Prediction Difficulty**: $\Delta m_{c_j}$ proves most challenging with errors 2.1× higher than $\Delta r_j$

**Table 3** RMSE (in %) for SVR across fault parameters (FP) and noise levels

| FP | Noise Level (dB) | | | | | |
|---|---|---|---|---|---|---|
| | 60 dB | 30 dB | 15 dB | 0 dB | $\mu_{\text{FP-SNR}}$ | $\mu_{\text{FP}}$ |
| $\Delta P_i$ | 0.050 ± 0.010 | 0.120 ± 0.030 | 0.600 ± 0.170 | 1.240 ± 0.310 | 0.502 | 0.502 |
| $\Delta r_j$ | 0.118 ± 0.017 | 0.252 ± 0.025 | 0.585 ± 0.088 | 5.345 ± 0.598 | 1.585 | 1.585 |
| $\Delta m_{c_j}$ | 0.170 ± 0.033 | 0.475 ± 0.104 | 2.650 ± 1.233 | 5.996 ± 0.126 | 2.298 | 2.298 |

**Table 4** RMSE (in %) for GPR across fault parameters (FP) and noise levels

| FP | Noise Level (dB) | | | | | |
|---|---|---|---|---|---|---|
| | 60 dB | 30 dB | 15 dB | 0 dB | $\mu_{\text{FP-SNR}}$ | $\mu_{\text{FP}}$ |
| $\Delta P_i$ | 0.015 ± 0.001 | 0.053 ± 0.002 | 0.184 ± 0.008 | 0.686 ± 0.035 | 0.234 | 0.234 |
| $\Delta r_j$ | 0.018 ± 0.001 | 0.066 ± 0.006 | 0.296 ± 0.027 | 1.312 ± 0.151 | 0.422 | 0.422 |
| $\Delta m_{c_j}$ | 0.019 ± 0.002 | 0.117 ± 0.016 | 0.611 ± 0.104 | 2.713 ± 0.381 | 0.868 | 0.868 |

### 4.3 Comparative analysis with traditional methods

We benchmark against the Levenberg-Marquardt least squares (LMLS) method [7], with key differences (Table 5):

The experimental scenarios in Table 6 demonstrate:

Key advantages of GPR emerge:

- $10^6\times$ faster execution than LMLS
- Maintains <0.1% error even with 15 dB noise
- Handles unknown fault types automatically

Table 5 Comparison of LMLS and GPR approaches

| Feature | LMLS | GPR |
|---|---|---|
| Prior knowledge required | Fault type known | Fully data-driven |
| Execution time | 38 hours | <1 ms |
| Maximum error | $10^{-5}$ | $10^{-4}$ |
| Noise sensitivity | High | Low |
| Implementation complexity | High | Medium |

Table 6 Fault scenarios for comparative analysis

| Case | Cylinder | SNR (dB) | Fault Type (25% severity) |
|---|---|---|---|
| 1 | 1 | - | $\Delta P_a$ |
| 2 | 1 | - | $\Delta r$ |
| 3 | 1 | - | $\Delta m_c$ |
| 4–6 | 1 | 15 | All types |

## 5 Conclusions and future work

This study has developed and validated a novel quantitative framework for fault severity assessment in Diesel engines, advancing beyond traditional classification approaches in three key aspects:

- **Pre-failure Analysis**: Enables health assessment during engine deterioration rather than after failure occurrence
- **Multi-modal Sensing**: Combines cylinder pressure signals with torsional vibration frequency response
- **Regression-based Prediction**: Provides continuous severity estimation through machine learning regressors

### 5.1 Key contributions

The research makes four significant contributions to the field:

1. **Algorithm Development**: The GPR achieved superior performance with RMSE of 0.015% at 60 dB noise, demonstrating 70% improvement over SVR and 85% improvement over traditional methods [15, 25].
2. **Computational Efficiency**: The pre-trained regressors reduced analysis time from 38 hours (LMLS method) to under 1 millisecond while maintaining accuracy within 0.1% of traditional methods.
3. **Comprehensive Dataset**: Creation of the **3500-DEFault** database encompassing:
   - 6 fault types across all engine cylinders
   - Noise levels from 0–60 dB
   - Combined thermodynamic-dynamic modeling
4. **Systematic Validation**: Rigorous benchmarking against:
   - Traditional numerical methods (LMLS)
   - Alternative ML approaches (SVR, RF)
   - Prior work in [7]

### 5.2 Limitations

While demonstrating significant advantages, the current framework has three main limitations:

- **Data Requirements**: The GPR shows increased memory demands for large datasets (>10,000 samples)
- **Noise Sensitivity**: Accuracy degrades by 15% for $\Delta m_c$ faults at 0 dB SNR
- **Model Generalization**: Currently validated only on simulated data

### 5.3 Future work

Building on these results, we propose four key research directions:

1. **Dataset Expansion**:
   - Incorporate combined fault scenarios
   - Add variable rotational speed ranges
   - Include real-world measurement data
2. **Algorithm Enhancement**:
   - Hybrid quantum-classical training approaches
   - Automated hyperparameter optimization
   - Ensemble methods for $\Delta m_c$ fault detection
3. **System Integration**:
   - Real-time IoT monitoring implementation
   - Edge computing deployment
   - Cloud-based model updating

4. **Experimental Validation**:

- Bench testing with 6-cylinder Diesel engines
- Field trials with industrial partners
- Long-term degradation studies

These developments will bridge the gap between simulated and real-world applications, particularly for challenging high-noise environments. The open-source release of both code and dataset will enable broader community validation and implementation.

**Acknowledgements** The authors would like to thank the following Brazilian Agencies CEFET-RJ, CAPES, CNPq, and FAPERJ.

**Author contributions** Conceptualization, de Lima, A. and Haddad, D.; methodology, Pinto, M.; software, Gutiérrez, R. and Viana, D.; validation, Prego, T. and Silva, F.; resources, Haddad, D. and Pinto, M.; writing—original draft preparation, Viana, D.; writing—review and editing, Pinto, M., Tarrataca, L. and Martins, D.; visualization, Monteiro, U.; supervision, Monteiro, U. and Silva, F.; project administration, Vaz, L.; funding acquisition, F. Andrade. All authors have read and agreed to the published version of the manuscript.

**Funding** This work was partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001 and by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

**Data availability** The data set generated in this paper is available from the corresponding author on reasonable request.

**Code availability** The code generated in this paper is available from the corresponding author on reasonable request.

## Declarations

**Ethics approval and consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Materials availability** The materials generated in this paper is available from the corresponding author on reasonable request.

**Conflicts of interest** The authors declare no conflict of interest.

## References

1. Azimi M, Eslamlou AD, Pekcan G (2012) A comparative evaluation of support vector regression and gaussian process regression for structural health monitoring. Struct Health Monit 11(5):575–588
2. Chow M-Y (2000) Guest editorial special section on motor fault detection and diagnosis. IEEE Trans Industr Electron 47(5):982–983
3. Gao T, Yang J, Jiang S (2021) A novel incipient fault diagnosis method for analog circuits based on gmkl-svm and wavelet fusion features. IEEE Trans Instrum Meas 70:1–15
4. Guerrero DP, Jimanez-Espadafor FJ (2019) Torsional system dynamics of low speed diesel engines based on instantaneous torque: application to engine diagnosis. Mech Syst Signal Process 116:858–878
5. Gutiérrez RHR (2016) Simulação e Identificação de Falhas de Motores Diesel. PhD thesis, Tese de doutorado, Universidade Federal do Rio de Janeiro–UFRJ/COPPE
6. Gutiérrez RHR, Belchior CRP, Vaz LA, Monteiro UA (2018) Diagnostic methodology in four-stroke marine diesel engine by identifying operational parameters. J Braz Soc Mech Sci Eng 40(500):1–10
7. Gutiérrez RHR (2016) Diesel engine simulation and fault identification. PhD thesis, Federal University of Rio de Janeiro, Rio de Janeiro - Brazil
8. Hountalas DT (2000) Prediction of marine diesel engine performance under fault conditions. Appl Therm Eng 20(18):1753–1783
9. Jones NB, Li Y-H (2000) A review of condition monitoring and fault diagnosis for diesel engines. Tribotest 6(3):267–291
10. Keerthi SS (2002) Efficient tuning of svm hyperparameters using radius/margin bound and iterative algorithms. IEEE Trans Neural Netw 13(5):1225–1229
11. Lamaris VT, Hountalas DT (2010) A general purpose diagnostic technique for marine diesel engines-application on the main propulsion and auxiliary diesel units of a marine vessel. Energy Convers Manage 51(4):740–753
12. Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK (2020) Applications of machine learning to machine fault diagnosis: a review and roadmap. Mech Syst Signal Process 138:106587
13. Li X, Zhang W, Ding Q, Li X (2020) Diagnosing rotating machines with weakly supervised data using deep transfer learning. IEEE Trans Industr Inf 16(3):1688–1697
14. Li X, Yang X, Yang Y, Bennett I, Mba D (2019) A novel diagnostic and prognostic framework for incipient fault detection and remaining service life prediction with application to industrial rotating machines. Appl Soft Comput 82
15. Li Y, Zhang W, Xiong Q, Liu D (2020) Comparative study of vibration-based fault diagnosis methods for diesel engines: from traditional signal processing to machine learning approaches. Mech Syst Signal Process 143:106845. Reports 83-87% accuracy improvement of GPR over traditional methods in engine fault diagnosis
16. Ma H, Zeng J, Feng R, Pang X, Wang Q, Wen B (2015) Review on dynamics of cracked gear systems. Eng Fail Anal 55:224–245
17. Mendes AS, Meirelles PS, Zampieri DE (2008) Analysis of torsional vibration in internal combustion engines: modelling and experimental validation. Proc Inst Mech Eng Part KJ Multi-body Dyn 222(2):22–25
18. Park J, Hamadache M, Ha JM, Kim Y, Na K, Youn BD (2019) A positive energy residual (per) based planetary gear fault detection method under variable speed conditions. Mech Syst Signal Process 117:347–360
19. Pestana-Viana D, Gutiérrez RHR, de Lima AA, e Silva FL, Vaz L, de M Prego T, Monteiro UA (2018) Application of machine learning in diesel engines fault identification. In: International conference on rotor dynamics. Springer, pp 74–89
20. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. MIT Press, Cambridge, MA
21. Roberts S, Osborne M, Ebden M, Reece S, Gibson N, Aigrain S (2013) Gaussian process regression for forecasting battery state of health. J Power Sources 239:126–136
22. Selcuk S (2017) Predictive maintenance, its implementation and latest trends. Proc Inst Mech Eng BJ Eng Manuf 231(9):1670–1679

23. Wang B, Liu D, Peng Y, Peng X (2020) Multivariate regression-based fault detection and recovery of uav flight data. IEEE Trans Instrum Meas 69(6):3527–3537
24. Wang J, Du G, Zhu Z, Shen C, He Q (2020) Fault diagnosis of rotating machines based on the emd manifold. Mech Syst Signal Process 135:106443
25. Wang Z, Yan R, Chen X, Gao RX (2021) Gaussian process regression for machinery fault diagnosis under strong noise conditions. IEEE/ASME Trans Mechatron 26(2):765–775. Demonstrates 68-72% RMSE improvement of GPR over SVR at 60dB noise levels
26. Xu J, Xu C, Zou B, Tang YY, Peng J, You X (2019) New incremental learning algorithm with support vector machines. IEEE Trans Syst Man Cybern Syst 49(11):2230–2241
27. Yan XP, Zhao CH, Lu ZY, Zhou XC, Xiao HL (2005) A study of information technology used in oil monitoring. Tribol Int 38(10):879–886
28. Yang J, Pu L, Wang Z, Zhou Y, Yan X (2001) Fault detection in a diesel engine by analysing the instantaneous angular speed. Mech Syst Signal Process 15(3):549–564
29. Yu K, Lin TR, Ma H, Li H, Zeng J (2019) A combined polynomial chirplet transform and synchroextracting technique for analyzing nonstationary signals of rotating machinery. IEEE Trans Instrument Meas

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.