A Two-Stage Machine Learning Approach for Quantitative Gear Crack Detection Using Vibration Signal Analysis

Paarth Singh Rathore[1], Arvind Keprate [2], L. Rajya Lakshmi[1], Omar D. Mohammed[3] and Debasish Ghose [4]

[1] *Birla Institute of Technology and Science Pilani, Rajasthan 333031, India*
h20240077@pilani.bits-pilani.ac.in
rajya99@gmail.com

[2] *Green Energy Lab, Department of Mechanical, Electrical and Chemical Engineering, Oslo Metropolitan University, Norway*
arvindke@oslomet.no

[3] *School of Science and Technology, Örebro University, Örebro, Sweden*
omar.mohammed@oru.se

[4] *School of Economics, Innovation and Technology, Kristiania University of Applied Sciences, Bergen, 5022, Norway*
Debasish.Ghose@kristiania.no

ABSTRACT

Early detection of gear tooth cracks is essential for preventing catastrophic failures in rotating machinery, yet existing approaches struggle to accurately detect small incipient cracks due to their distinct vibration characteristics compared to larger cracks. Current machine learning methods optimize for overall performance, sacrificing sensitivity to early-stage damage where preventive maintenance is most effective. This study presents a regime-aware feature-driven two-stage modeling framework employing separate polynomial Ridge regression models for small cracks and large cracks, selected through systematic correlation analysis and exhaustive grid search optimization using vibration features extracted from residual signals. The small crack model utilizes wavelet detail coefficients while the large crack model employs clearance factor, envelope peak, and wavelet d1_std features, both validated using Leave-One-Out cross-validation with limited training data. Simulation results demonstrate that the proposed approach achieves $R^2 = 0.9982$ under simulated, data-scarce conditions, with performance evaluated using leave-one-out cross-validation on 19 samples.

Keywords: Gear fault diagnosis, vibration analysis, two-stage modeling, Ridge regression, limited data, predictive maintenance

1. INTRODUCTION

Modern industrial machinery is becoming increasingly complex as automation and interconnected systems have become standard across manufacturing facilities. In such environments, gear transmissions play a critical role in power transfer and mechanical synchronization. Any failure in a gear system can propagate through the drivetrain, potentially halting entire production lines. Gear tooth cracks are particularly critical because they initiate locally but may evolve into catastrophic failure if undetected. (Mohammed, 2023) studied crack detection for high contact ratio gears with modelling different crack sizes. It was observed that the detection indicators are less sensitive compared to low contact ratio gears. Machinery faults, especially gear cracks, in industrial settings can lead to significant economic losses and serious safety risks (Naveen, SampathKumaran, Arvind, & Seetharamu, 2024). Therefore, early crack detection is essential for maintaining operational reliability and enabling effective predictive maintenance.

Despite extensive research in vibration-based gear fault diagnosis, quantitative crack size estimation remains a challenging problem. Recent comprehensive reviews of machine learning applications in structural health monitoring have identified several critical challenges that limit the practical deployment of crack detection systems (Omar, Khan, & Starr, 2022). First, data availability and quality remain fundamental obstacles. Obtaining sufficient high-quality labeled data representing all damage scenarios is often economically infeasible or poses safety concerns. In industrial

settings, collecting crack progression data typically requires deliberate damage introduction or waiting for natural failures to occur, both of which are impractical (Omar et al., 2022). Consequently, many models must operate under severely limited training samples, increasing the risk of overfitting and unstable generalization.

Second, feature selection significantly impacts model performance. The extraction of damage-sensitive features from raw vibration signals has frequently been performed without clear justification of their suitability across different crack severities (Omar et al., 2022). Conventional approaches generally apply a uniform feature set across the entire damage range, implicitly assuming a consistent relationship between vibration characteristics and crack size. However, gear crack propagation induces nonlinear changes in mesh stiffness and contact dynamics. Incipient cracks primarily generate subtle high-frequency perturbations, whereas advanced cracks produce stronger impact-dominated responses and amplitude modulation. Treating these stages uniformly may obscure regime-dependent vibration behavior.

Several works have employed statistical time-domain features such as RMS, kurtosis, and peak value for gear crack detection (Elyousfi, Soualhi, Medjaher, & Guillet, 2020). However, traditional time-domain features often exhibit minimal variation during gradual crack progression, limiting their sensitivity to early-stage damage (Li, Ma, Liu, Teng, & Jiang, 2015). To overcome this limitation, time–frequency methods such as wavelet decomposition have been introduced, with studies demonstrating that wavelet coefficient energy provides improved diagnostic information compared to purely time or frequency features (Li et al., 2015) (Cerrada, Sánchez, Cabrera, Zurita, & Li, 2015). Clearance factor has also been demonstrated as an effective indicator for crack depth estimation by capturing impact severity associated with progressive tooth root fracture (Rezaei, Poursina, & Rezaei, 2021). Additionally, envelope analysis is widely used to extract defect frequency components from vibration signals (Ho & Randall, 2000).

Although these features enhance damage sensitivity, they are typically embedded within single regression or classification models that assume a consistent feature–damage relationship across all crack stages. This unified modeling strategy may be suboptimal if the dominant vibration mechanisms differ between small and large cracks.

Motivated by these observations, this paper addresses vibration-based gear crack quantification under limited data conditions through a physics-informed two-stage regression approach. Instead of increasing model complexity, the proposed framework exploits regime-dependent vibration behavior to guide model design. Crack progression is partitioned into incipient (≤ 0.5 mm) and advanced (> 0.5 mm) stages, and a deterministic feature-based routing mechanism using a monotonic envelope statistic assigns samples to regime-specific models. Independent feature selection and regularization optimization are then performed within each regime. The main contributions of this work are summarized as follows:

1. A physics-informed regime-dependent modeling strategy that partitions crack progression into incipient (≤ 0.5 mm) and advanced (> 0.5 mm) stages, reflecting distinct vibration mechanisms across crack severities based on `hr_envelope_std`
2. A deterministic feature-based routing mechanism using a monotonic envelope statistic to assign samples to independently optimized polynomial Ridge regression models without introducing an additional classifier.
3. Empirical demonstration, under Leave-One-Out Cross-Validation on a limited dataset, that regime-specific feature selection and regularization improve crack size estimation accuracy compared to a unified single-regime regression model.

The remainder of this paper is organized as follows. Section II describes the Methodology. Section III presents the proposed RAFTS approach. Section IV discusses the results and comparative analysis, Section V discusses limitations and generalizability, and Section VI concludes the paper.

2. METHODOLOGY

2.1. Dataset

The dataset used in this study was generated through a numerical simulation of a 6 DOF spur gear dynamic model implemented in MATLAB. The used model was presented by (Mohammed, 2023) and (Mohammed, Rantatalo, & Aidanpää, 2015). The equations of motion were solved using the ODE45 solver with a sampling frequency of 200 kHz. A constant rotational speed was maintained throughout all simulations to ensure consistency across different crack conditions and to isolate the effect of crack severity on the vibration response.

The simulation was first performed for a healthy gear condition, which served as the baseline reference. Subsequently, the simulation was repeated for multiple faulty cases corresponding to progressively increasing crack depths. Crack propagation was modeled explicitly by varying the crack size parameter while keeping all other system properties unchanged.

The resulting dataset consisted of two categories of vibration signals:

- **Healthy baseline signal (hb0):** Vibration response of the gear system without any crack damage.
- **Faulty signals with increasing crack depth (hb1–hb18):** Vibration responses corresponding to discrete

stages of crack propagation.

In total, 19 crack conditions were simulated, spanning crack depths from 0.0 mm (healthy baseline) to 1.8 mm (severe damage) in uniform increments of 0.1 mm. This range captures the full evolution of gear crack growth from incipient damage to advanced fault states. Each crack condition produced a single vibration signal, resulting in a dataset of 19 samples.

Although the dataset size is limited, this setup reflects realistic industrial constraints, where acquiring large quantities of labeled crack progression data is often impractical due to cost, safety, and operational considerations. The dataset was therefore intentionally designed to evaluate the effectiveness of physics-informed modeling and feature selection strategies under data-scarce conditions.

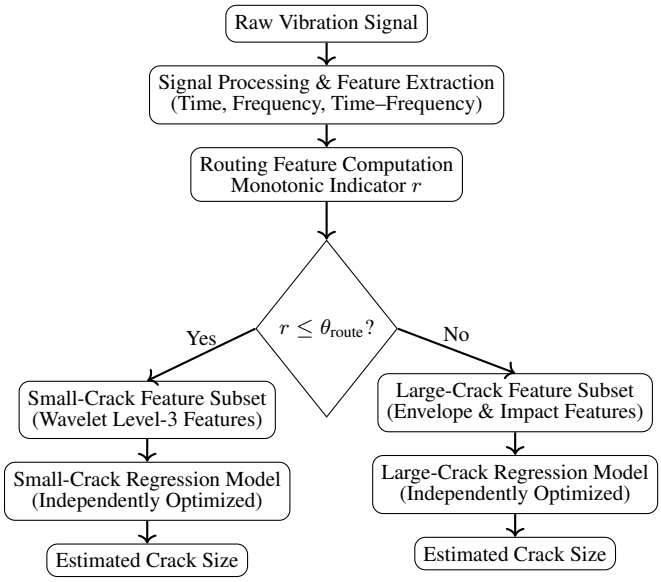

Figure 1. Flowchart of the proposed RAFTS.

2.2. Signal Processing and Residual Extraction

To enhance sensitivity to crack-induced vibration components and suppress nominal gear dynamics, a residual signal formulation was adopted (Mohammed & Rantatalo, 2015). For each crack condition, the residual vibration signal was computed as:

$$\text{hr}_i = \text{hb}_i - \text{hb}_0 \tag{1}$$

where hb_i denotes the vibration signal corresponding to crack size i, and hb_0 represents the healthy baseline signal obtained under identical operating conditions.

This residual formulation effectively removes repeatable components such as gear meshing harmonics, manufacturing tolerances, and baseline system dynamics that are common to both healthy and damaged states. As a result, vibration components directly attributable to crack-induced stiffness perturbations and impact phenomena are amplified.

Residual signal analysis is particularly beneficial for incipient crack detection, where damage-related signatures are subtle and may otherwise be masked by dominant meshing frequencies. Applying this preprocessing step to all simulated conditions yielded a total of 19 residual signals, corresponding to crack depths ranging from 0.0 mm to 1.8 mm.

2.3. Feature Extraction

A comprehensive feature extraction framework was employed to convert each residual vibration signal into a set of quantitative descriptors sensitive to crack progression. Features were selected to capture complementary characteristics across time, frequency, time–frequency, and nonlinear dynamic domains, while maintaining physical interpretability.

In total, 138 features were computed from each residual signal. Table 1 summarizes the feature categories, definitions, and physical interpretations used in this study.

2.4. Regime-Aware Feature-Driven Two-Stage Modeling Framework (RAFTS)

The vibration response of a cracked gear evolves nonlinearly with crack severity due to regime-dependent changes in gear mesh dynamics (Guo, Guo, Sun, & Gao, 2014; Goel & Kumar, 2021; Nguyen, Pham, Do, Liang, & Nguyen, 2025). In the incipient damage stage, crack initiation and early propagation primarily induce localized perturbations in mesh stiffness. These perturbations generate weak, nonstationary, high-frequency vibration components embedded within the dominant gear meshing signal. As crack severity increases, the damage mechanism progressively transitions toward intermittent tooth contact, reduced load-carrying capacity, and repeated impact events. This transition results in strong amplitude modulation and impulsive vibration signatures.

Due to this regime-dependent physical behavior, vibration features do not exhibit uniform diagnostic sensitivity across the full crack progression range. Features that are effective for detecting subtle stiffness variations in early-stage cracks differ fundamentally from those that best characterize impact-dominated responses associated with advanced damage. Consequently, the use of a single globally optimized feature representation and regression model is inadequate for accurate crack size estimation over the entire severity range.

To address this limitation, a feature-driven two-stage modeling strategy is adopted. The RAFTS methodology consists of (i) deterministic regime separation using a monotonic vibration feature and (ii) regime-specific regression modeling.

In the first stage, the crack size threshold is fixed at 0.5mm based on domain knowledge of gear mesh dynamics, parti-

Table 1. Summary of extracted vibration features and their physical interpretation

Feature Domain	Computed Features	Physical Interpretation
Time domain	Mean, RMS, standard deviation, variance, peak, peak-to-peak, skewness, kurtosis, crest factor, clearance factor, impulse factor, shape factor	Captures amplitude dispersion, impulsiveness, and statistical deviation caused by crack-induced impacts
Frequency domain	Spectral mean, RMS, peak magnitude, skewness, kurtosis, spectral centroid, spectral energy	Characterizes redistribution of vibration energy and harmonic distortion due to stiffness variation
High-frequency band (> 50 kHz)	Energy, RMS, peak, mean, standard deviation, kurtosis	Enhances sensitivity to short-duration transients associated with crack initiation and propagation
Wavelet domain (db4, 5 levels)	Energy, RMS, peak, standard deviation, kurtosis for detail coefficients d1–d5	Provides localized time–frequency characterization of transient crack signatures across scales
Envelope analysis	Mean, RMS, standard deviation, peak, skewness, kurtosis, crest factor; envelope spectrum peak, mean, energy	Extracts amplitude modulation caused by intermittent tooth contact and impact events
Time–frequency (STFT)	Spectrogram peak, mean, standard deviation, total energy, spectral entropy	Captures nonstationary energy distribution and complexity of crack-induced vibration behavior
Phase-space domain	Delay-coordinate correlation, trajectory range metrics	Represents nonlinear dynamic changes in system behavior due to crack evolution

tioning the dataset into small-crack and large-crack regimes. For runtime regime classification, a monotonic feature is used as a routing indicator. The routing threshold is determined by evaluating two-stage model performance across a range of the said monotonic feature values spanning the boundary region between small and large crack regimes. The threshold yielding the best overall LOOCV performance is selected. A sensitivity analysis confirms that performance is moderately sensitive to threshold selection, emphasizing the importance of threshold calibration.

In the second stage, independent regression models are trained for each regime. For both the small- and large-crack regions, hyperparameter optimization is performed separately, considering the regression model type, regularization parameter, and feature subset selection. This regime-specific optimization enables each model to focus on vibration characteristics that are physically consistent with the dominant damage mechanism in the corresponding regime.

The resulting RAFTS framework explicitly incorporates regime-dependent vibration behavior by combining feature-based regime separation with regime-optimized regression models.

3. EXPERIMENTAL SETUP

3.1. Hardware Configuration and Software Environment

All experiments were conducted on a laptop workstation equipped with a 12th Gen Intel Core i5-12450H processor (2.00 GHz, 8 cores), 16.0 GB RAM (15.7 GB usable), running a 64-bit Windows 11 Home operating system (Build 26200.7623). No dedicated GPU was utilized; all computations were performed exclusively on CPU, reflecting the practical deployment constraints of industrial diagnostic systems with limited computational resources. All experiments were implemented in Python. The scientific computing stack comprised NumPy and SciPy for numerical operations, pandas for data management, and scikit-learn for machine learning model implementation, preprocessing, and cross-validation. Signal processing and feature extraction utilized PyWavelets (pywt) for wavelet decomposition and scipy.io for loading MATLAB (.mat) simulation files. Visualization was performed using matplotlib and seaborn. XGBoost was installed separately for gradient boosting baseline evaluation.

3.2. Set up for RAFTS

The feature `hr_envelope_std` was selected as the routing discriminator because the envelope standard deviation provides a monotonically increasing signal with crack size that cleanly separates the two regimes without requiring feature combinations or learned classifiers. Ten candidate routing thresholds were evaluated over the range $[4.42 \times 10^{-8},\ 6.75 \times 10^{-8}]$ of `hr_envelope_std`. For each candidate, the dataset was partitioned into small and large-crack subsets, regime-specific Ridge polynomial models were trained, and Leave-One-Out Cross-Validation overall R^2 and MAE were computed as the evaluation metric. The optimal routing threshold was found at 5.004×10^{-8}.

To assess whether the selected features reflect genuine regime-dependent vibration behaviour or artifacts of the specific training samples, a leave-one-out feature stability analysis was conducted for the small-crack regime. In each of the 6 iterations, one sample was withheld and the complete feature selection procedure, correlation screening followed by exhaustive grid search over all 129 candidate combinations, was re-run on the remaining 5 samples. Wavelet level-3 features appeared in 4 of 6 iterations (67%), consistently emerging whenever sufficient sample diversity existed in the training subsample. While the exact feature triplet varied with sample composition, the consistent selection of the wavelet level-3 family suggests these features capture a physically meaningful signal related to the high-frequency stiffness perturbations of incipient cracks, rather than being artifacts of the specific 6 training samples.

The grid search was done across three axes: (i) feature subset drawn from filtered candidates, (ii) polynomial degree $\in \{1, 2\}$, and (iii) Ridge regularisation strength $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Each combination was evaluated under LOOCV within its respective regime, and the configuration minimising CV MAE (subject to maximising R^2) was selected. Candidate features were obtained by selecting those features having correlation as one in at least one of the correlation methods (Pearson, Spearman, and Kendall).

The optimal alpha for the small-crack model was 10^{-3}. The large-crack and all-cracks models both selected $\alpha = 10^{-4}$. The slightly higher optimal alpha for small cracks reflects the greater regularisation needed to prevent overfitting with only 6 training samples.

The optimal small-crack configuration uses features `hr_wavelet_d3_energy`, `hr_wavelet_d3_rms`, and `hr_wavelet_d3_std`, with polynomial degree 2 and $\alpha = 10^{-3}$. All three features derive from wavelet decomposition detail level 3. The large-crack model selects `hr_clearance_factor`, `hr_envelope_peak`, and `hr_wavelet_d1_std`, with polynomial degree 2 and $\alpha = 10^{-4}$. This configuration achieves near-perfect prediction, consistent with the stronger and more structured vibration signatures produced by larger cracks.

The single-regime model selects `hr_envelope_std`, `hr_spectral_mean`, and `hr_stft_std`, with polynomial degree 2 and $\alpha = 10^{-4}$. While competitive overall, it underperforms on small cracks, supporting the value of regime separation under the simulated conditions studied.

The proposed model operates as follows: given an input vibration sample, `hr_envelope_std` is computed and compared against the routing threshold (5.004×10^{-8}). Samples below the threshold are forwarded to the small-crack Ridge polynomial model; samples above are forwarded to the large-crack model. Each regime-specific model then produces a crack size prediction using its optimised feature set and scaling strategy.

$$\text{Model} = \begin{cases} f_{\text{small}}(\mathbf{x}) & \text{if hr\_envelope\_std}(\mathbf{x}) \leq \theta_{\text{route}} \\ f_{\text{large}}(\mathbf{x}) & \text{otherwise} \end{cases} \quad (2)$$

where f_{small} and f_{large} denote the regression models for the small- and large-crack regimes, respectively.

3.3. Baseline Methods

Six baseline regression methods were evaluated under identical LOOCV conditions using the same three-feature input set (hr_envelope_std, hr_spectral_mean, hr_stft_std) identified via grid search for the complete crack range. All methods em-

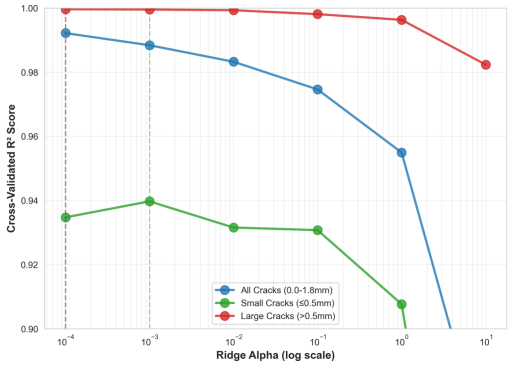

Figure 2. R^2 score vs Ridge regularization parameter. Large-crack models exhibit exceptional stability ($R^2 > 0.98$) across all α values, suggesting robust feature-target relationships, while all-cracks models degrade significantly with strong regularization.

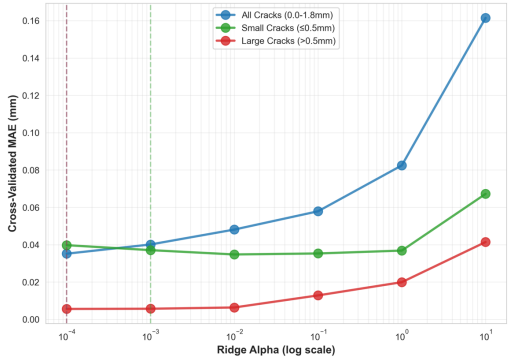

Figure 3. MAE vs Ridge regularization parameter across crack regimes. The divergent sensitivity patterns motivate the use of specialized models tailored to each crack size range.

ploy StandardScaler normalization prior to training to ensure fair comparison.

Linear Ridge Regression applies L_2-regularized linear regression with regularization parameter $\alpha = 0.0001$ and a maximum of 10,000 solver iterations (Hoerl & Kennard, 1970). This serves as the simplest baseline, establishing a lower bound on performance by assuming a purely linear relationship between vibration features and crack size.

Polynomial Ridge Regression extends the linear baseline by applying degree-2 polynomial feature expansion prior to Ridge regression ($\alpha = 0.0001$), generating quadratic interaction terms and squared features (Hoerl & Kennard, 1970). This is the primary single-regime baseline.

Support Vector Regression (SVR) employs a radial basis function (RBF) kernel with regularization parameter $C = 100$, insensitive loss margin $\epsilon = 0.01$, and automatic kernel coefficient ($\gamma =$ scale) (Smola & Schölkopf, 2004). SVR is well-suited for small-sample regression due to its margin-based formulation and kernel-induced non-linearity, representing a fundamentally different algorithmic approach to Ridge regression.

Random Forest Regression constructs an ensemble of 100 decision trees with maximum depth 5 and minimum samples per split of 2, using a fixed random seed of 42 for reproducibility (Breiman, 2001). As a bagging ensemble method, Random Forest provides strong non-linear regression capability but is susceptible to overfitting under severe data scarcity, making it an important baseline for assessing data efficiency requirements.

Gradient Boosting Regression trains a sequential ensemble of 100 shallow decision trees (maximum depth 3) with learning rate 0.1 and fixed random seed 42 (Friedman, 2001). Unlike bagging methods, gradient boosting iteratively corrects residual errors, potentially achieving higher accuracy but with greater overfitting risk under limited data.

XGBoost implements extreme gradient boosting with 100 estimators, maximum tree depth of 3, learning rate of 0.1, and fixed random seed 42 (Chen & Guestrin, 2016). XGBoost represents the state-of-the-art in gradient boosting with additional regularization and computational optimizations, serving as the strongest ensemble baseline for comparison.

3.4. Hyperparameter Optimization

Grid search with leave-one-out cross-validation (LOOCV) identified optimal feature sets, polynomial degrees, and Ridge regularization parameters (α) for three scenarios: single-regime, small cracks only, and large cracks only. The single-regime model utilized envelope standard deviation, spectral mean, and STFT standard deviation with polynomial degree 2 and $\alpha = 0.0001$. Small-crack model employed wavelet level-3 features with $\alpha = 0.001$, while large-crack model used clearance factor, envelope peak, and wavelet d1 std features with $\alpha = 0.0001$.

Figure 2 demonstrates differences in regularization sensitivity across crack regimes. Large-crack models maintained $R^2 > 0.98$ across the entire regularization spectrum, while single-regime model showed significant degradation with increasing α, and small-crack models exhibited intermediate sensitivity. This divergent behavior motivated regime-specific modeling. Figure 3 confirms these trends through RMSE analysis, with large-crack models showing stability.

4. RESULTS AND DISCUSSION

4.1. Performance Comparison of RAFTS with Baseline

To rigorously evaluate the proposed RAFTS architecture, six baseline regression algorithms were assessed on both crack regimes using regime-optimized features identified via grid search. Each baseline was trained and evaluated independently on small cracks and large cracks under LOOCV conditions.

Small Crack Performance: The RAFTS small-crack model achieved $R^2 = 0.9434$ and MAE $= 0.0361$ mm, outperforming all baselines under the simulated, data-scarce conditions studied. Polynomial Ridge, using identical features, achieved only $R^2 = 0.5370$ (MAE $= 0.0815$ mm), while ensemble methods performed comparably (Gradient Boosting: $R^2 = 0.6571$, Random Forest: $R^2 = 0.5719$). Linear Ridge exhibited negative $R^2 = -0.2766$, suggesting an inability to capture non-linear crack progression dynamics in the incipient regime.

Large Crack Performance: The RAFTS large-crack model achieved near-perfect accuracy ($R^2 = 0.9997$, MAE $= 0.0055$ mm), matching polynomial Ridge baseline performance ($R^2 = 0.9996$, MAE $= 0.0055$ mm) on this regime. However, ensemble methods degraded substantially: XGBoost ($R^2 = 0.9283$, MAE $= 0.1002$ mm), Gradient Boosting ($R^2 = 0.9597$), and Random Forest ($R^2 = 0.9616$) demonstrated 10–18$\times$ higher error rates, suggesting these ensemble methods are more susceptible to overfitting under limited data conditions.

These results suggest three key findings under the simulated conditions studied: (1) regime-specific optimization is beneficial for small crack detection, as single-regime models sacrifice small-crack sensitivity for global performance; (2) polynomial Ridge with L_2 regularization outperforms complex ensemble methods under data scarcity; and (3) RAFTS achieves superior small-crack performance while maintaining large-crack accuracy through physics-informed architecture rather than algorithmic complexity. These findings should be interpreted as preliminary evidence given the limited dataset size.

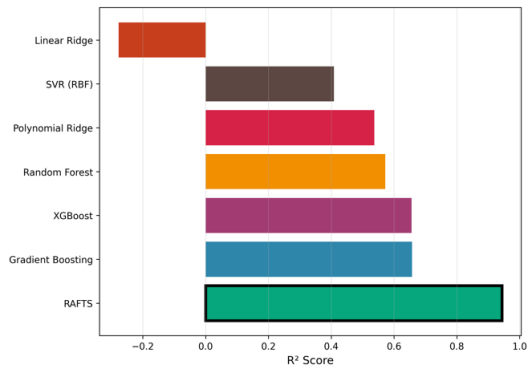

Figure 4. R^2 score comparison across baseline models for small cracks. The RAFTS architecture achieves the highest predictive accuracy under the simulated conditions studied, suggesting the potential effectiveness of regime-specific modeling over conventional algorithms.

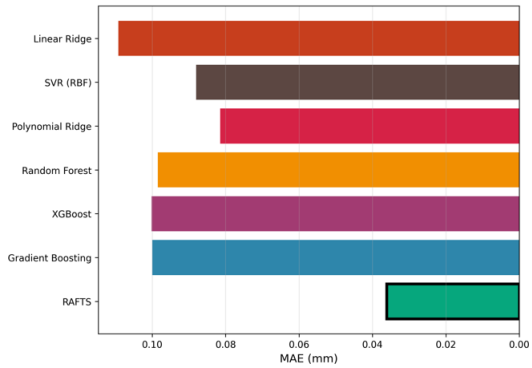

Figure 5. MAE comparison across baseline models for small cracks. The proposed RAFTS model achieves lower error than polynomial Ridge and outperforms ensemble methods under the simulated data-scarce conditions studied.

4.2. Impact of Routing Threshold in RAFTS

The RAFTS model employs a feature-based router that computes envelope standard deviation and directs samples to regime-specific models. Routing threshold optimization (Figures 9 and 10) tested 10 threshold values ranging from 4.42×10^{-8} to 6.75×10^{-8}. The optimal threshold of 5.00×10^{-8} achieved $R^2 = 0.9982$ and MAE$= 0.0151$ mm.

The proposed RAFTS model achieved overall $R^2 = 0.9982$ and MAE$= 0.0151$ mm, small-crack performance improved substantially ($\Delta R^2 = +0.1171$, ΔMAE$= -0.0205$ mm, -36.2%), while large-crack accuracy gained $\Delta R^2 = +0.0074$ and ΔMAE$= -0.0198$ mm (-78.4%).

4.3. Robustness Analysis Under Additive Noise

To partially address the gap between controlled simulation and real industrial signals, a noise robustness analysis was conducted by adding white Gaussian noise at three signal-to-noise ratio (SNR) levels to the residual signals prior to feature extraction. RAFTS and the single-regime polynomial Ridge baseline were re-evaluated under identical LOOCV conditions at each noise level, using the feature configurations and

Table 2. Model performance comparison (LOOCV)

Model	**Subset**	R^2	**MAE (mm)**
Baseline Model	Small	0.8264	0.0566
	Large	0.9923	0.0253
RAFTS Model	Small	0.9434	0.0361
	Large	0.9997	0.0055
Improvement			
	Small	+0.1171	-36.2%
	Large	+0.0074	-78.4%

Table 3. Bootstrap 95% confidence intervals (1,000 resamples). The large-crack regime shows tight, stable intervals confirming robust performance. The small-crack regime ($n = 6$) exhibits very wide intervals, quantifying the high sensitivity of model fitting to sample composition at this regime size and confirming that small-crack results should be treated as preliminary.

Regime	n	**LOOCV** R^2	**LOOCV MAE**	**95% CI** (R^2)
Large	13	0.9997	0.0056 mm	$(0.9919,\ 0.9999)$
Small	6	0.9434	0.0361 mm	$(-12.25,\ 0.9898)$

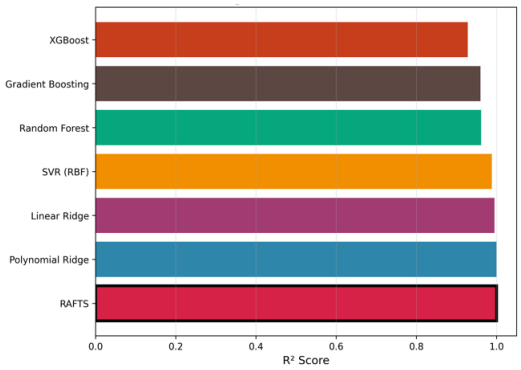

Figure 6. R^2 score comparison across baseline models for large cracks.

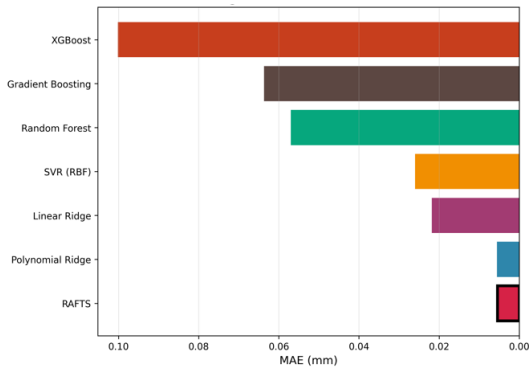

Figure 7. MAE comparison across baseline models for large cracks. The proposed RAFTS model achieves similar error compared to polynomial Ridge and outperforms ensemble methods under the simulated data-scarce conditions studied.

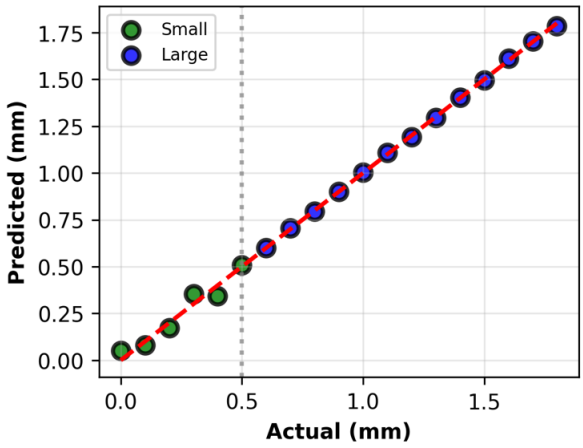

Figure 8. RAFTS model predictions ($R^2 = 0.9982$, MAE$= 0.0151$ mm) under simulated data-scarce conditions. Regime-specific modeling achieves improved alignment with the identity line across all crack depths, with a preliminary 57.0% error reduction observed over the single-regime baseline.

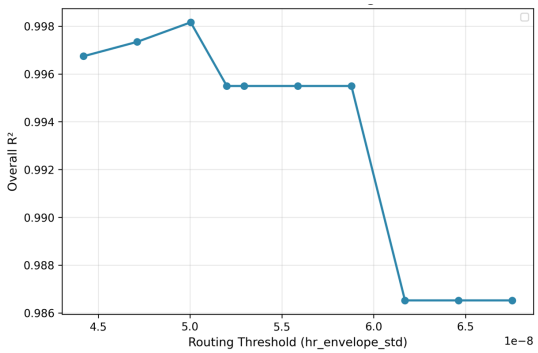

Figure 9. R^2 vs routing threshold. Performance peaks at 5.00×10^{-8} with $R^2 = 0.9982$.

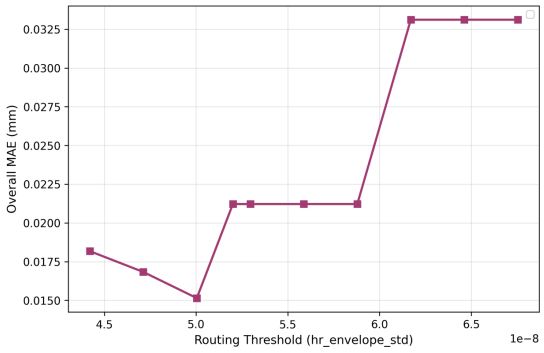

Figure 10. MAE vs routing threshold. Optimal threshold achieves MAE$= 0.0151$ mm.

hyperparameters identified on clean data without re-tuning. In the RAFTS evaluation, regime routing was also performed on noisy features, reflecting realistic deployment conditions.

Table 4. RAFTS and single-regime baseline performance under additive Gaussian noise (LOOCV, $n = 19$). Feature configurations and hyperparameters identified on clean data are applied without re-tuning, providing a conservative estimate of noise robustness.

Condition	**RAFTS**		**Single-Regime Baseline**	
	R^2	MAE (mm)	R^2	MAE (mm)
Clean (original)	0.9982	0.0151	0.9922	0.0352
SNR = 30 dB	0.9975	0.0178	0.9919	0.0351
SNR = 20 dB	0.9967	0.0227	0.9910	0.0369
SNR = 10 dB	0.9914	0.0367	0.9894	0.0429

These results suggest that the regime-separation principle underlying RAFTS is robust to moderate levels of additive noise, with RAFTS consistently outperforming the single-regime baseline across all tested noise conditions. Regime routing, performed on noisy features using the threshold calibrated on clean data, achieved 94.7% accuracy on the clean dataset and remained stable at 94.6% even at SNR = 10 dB, indicating that additive noise has negligible effect on regime assignment decisions. Performance degradation with increasing noise confirms that validation on real industrial signals — with structured noise, speed variation, and load fluctuation — remains necessary before deployment.

5. LIMITATIONS AND GENERALIZABILITY

Several limitations of the present study should be acknowledged. First, the dataset consists of 19 numerically simulated samples from a single operating condition. While this reflects realistic industrial constraints where labeled crack progression data is scarce, model design decisions—including feature selection, routing threshold, and regularization strength—were made using information from the full dataset. Although LOOCV was employed, held-out samples in each fold were not truly unseen with respect to these tuning decisions, which introduces optimistic bias into the reported performance metrics. The near-perfect $R^2 = 0.9982$ should therefore be interpreted as an upper-bound estimate of performance under the specific simulated conditions, rather than a definitive measure of real-world generalization capability.

Second, a bootstrap analysis (1,000 resamples) confirmed that the large-crack model is statistically stable, with a 95% confidence interval of $R^2 \in (0.9919,\ 0.9999)$. The small-crack model ($n = 6$) exhibited very wide bootstrap intervals (R^2 95% CI: -12.25 to 0.9898), consistent with the high sensitivity of model fitting to sample composition at this regime size. Such extreme lower bounds are a known artefact of bootstrap resampling on very small datasets, where a single influential sample can dominate the fitted model in some resamples; they indicate high variance in the fitted model rather than systematic failure in deployment. This reinforces that small-crack results should be treated as preliminary.

Third, real industrial vibration signals differ substantially from numerically simulated data. Factors such as sensor noise, manufacturing variability, load fluctuations, and varying rotational speeds introduce signal complexity absent in the controlled simulation environment (Omar et al., 2022). The gap between simulated and real-world signals remains an important open challenge for deployment of the proposed approach. A preliminary noise robustness analysis using additive Gaussian noise at SNR levels of 30, 20, and 10 dB showed that RAFTS maintained $R^2 \geq 0.991$ across all tested conditions, suggesting graceful degradation under controlled noise. However, real industrial noise has structured characteristics beyond white Gaussian noise that this analysis does not capture.

Fourth, the feature selection procedure, based on perfect correlation ($\rho = 1.0$) across 19 samples, may be susceptible to selection bias. The leave-one-out feature stability analysis showed that wavelet level-3 features appeared in 4 of 6 small-crack subsamples (67%), suggesting family-level consistency grounded in the known physical sensitivity of wavelet level-3 decomposition to high-frequency incipient crack signatures. However, exact feature identity remained sensitive to sample composition, and a larger dataset would be required to fully validate feature stability in this regime.

These limitations suggest that the results of this study are best interpreted as proof-of-concept for physics-informed regime separation under data-scarce conditions. Validation on experimental gear fault data under varying operating conditions remains an essential direction for future work.

6. CONCLUSION

Estimating gear crack size from vibration signals is challenging because small and large cracks induce fundamentally different frequency-domain responses, making a single regression model suboptimal across the full damage range. This work introduced the RAFTS framework, which uses an `hr_envelope_std`-based routing threshold to separate samples into two physically distinct regimes, each modelled with an independently optimised Ridge polynomial regressor. Evaluated using LOOCV on 19 samples (0.0–1.8 mm), RAFTS achieved $R^2 = 0.9982$ and MAE = 0.0151 mm on the simulated dataset, with preliminary results suggesting a 57.0% error reduction compared to a unified model and improved performance over all baselines. The primary contribution is showing that physically motivated regime separation, rather than increased model complexity, is a promising strategy for high-accuracy regression on small engineering datasets under simulated conditions. The observed dominance of wavelet level-3 energy features for small cracks and envelope-based statistics for larger cracks provides interpretable and transferable guidance for feature selection in gear health monitoring. Future work will focus on validation under broader operating conditions, including varying speeds, loads, and gear geometries, and on experimental gear fault datasets to assess real-world generalizability. Introducing probabilistic routing near the regime boundary and conducting in-situ industrial trials will be essential steps toward robust deployment in practical condition-monitoring systems.

REFERENCES

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Cerrada, M., Sánchez, R. V., Cabrera, D., Zurita, G., & Li, C. (2015). Multi-stage feature selection by using genetic algorithms for fault diagnosis in gearboxes based on vibration signal. *Sensors*, *15*(9), 23903–23926.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Elyousfi, B., Soualhi, A., Medjaher, K., & Guillet, F. (2020). A model-based analysis of crack fault in a two-stage spur gear system. In *Proceedings of the 2020 prognostics and health management conference (phm-besançon), besancon, france* (pp. 4–7).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Goel, V., & Kumar, N. (2021). Vibration response analysis of healthy and cracked gears through different signal processing techniques. *Vibroengineering Procedia*, *39*, 43–47.

Guo, J.-L., Guo, Y., Sun, S.-B., & Gao, Y. (2014). Simulation study of phase demodulation in resonance band for gear tooth cracking fault detection. In *2015 international conference on material science and applications (icmsa-15)* (pp. 795–798).

Ho, D., & Randall, R. (2000). Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals. *Mechanical systems and signal processing*, *14*(5), 763–788.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Li, Z., Ma, Z., Liu, Y., Teng, W., & Jiang, R. (2015). Crack fault detection for a gearbox using discrete wavelet transform and an adaptive resonance theory neural network. *Journal of Mechanical Engineering*, *61*(1), 63–73.

Mohammed, O. D. (2023). Crack detection of a modified high contact ratio gear. *Engineering Failure Analysis*, *152*, 107493.

Mohammed, O. D., & Rantatalo, M. (2015). Residual signal techniques used for gear fault detection. In *Cur-*

rent trends in reliability, availability, maintainability and safety: An industry perspective (pp. 157–163). Springer.

Mohammed, O. D., Rantatalo, M., & Aidanpää, J.-O. (2015). Dynamic modelling of gear system with gyroscopic effect and crack detection analysis. In *Proceedings of the 9th iftomm international conference on rotor dynamics* (pp. 1303–1314).

Naveen, G., SampathKumaran, P., Arvind, P., & Seetharamu, S. (2024). Analyzing gear blank failure: A comprehensive industrial case study. *Journal of Failure Analysis and Prevention*, *24*(1), 83–96.

Nguyen, P.-D., Pham, T.-D., Do, D.-T.-B., Liang, J.-W., & Nguyen, T.-D. (2025). Strain energy-based gear mesh stiffness modeling and synthetic data generation for ai-driven fault diagnosis in smart manufacturing. *Frontiers in Mechanical Engineering*, *11*, 1682102.

Omar, I., Khan, M., & Starr, A. (2022). Compatibility and challenges in machine learning approach for structural crack assessment. *Structural Health Monitoring*, *21*(5), 2481–2502.

Rezaei, M., Poursina, M., & Rezaei, E. (2021). Experimental investigation of helical gear tooth crack location and depth detection using moving average method on transmission error. *Proceedings of the Institution of Mechanical Engineers, Part L: Journal of Materials: Design and Applications*, *235*(10), 2266–2275.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*(3), 199–222.