Evaluating the Influence of Time Domain Feature Distributions on Estimating Rolling Bearing Flaking Size with Explainability

Osamu Yoshimatsu[1], Keiichirou Taguchi[2], Yoshihiro Sato[3], and Takehisa Yairi[4]

[1,2,3] *NSK Ltd., Kanagawa, 251-8501, Japan*
yoshimatsu-o@nsk.com
taguchi-kei@nsk.com
satou-yos@nsk.com

[1,4] *The University of Tokyo, Tokyo, 153-8904, Japan*
yairi@g.ecc.u-tokyo.ac.jp

ABSTRACT

To enhance the maintainability of rotating machines, such as wind turbines, where the response to bearing damage is both costly and time-consuming, it is essential to predict the progression of flaking, which is a common rolling bearing fault. Conventional rule-based methods estimate the magnitude of flaking by analyzing the time interval of feature vibrations. However, this method requires trial-and-error adjustments by experts, limiting its applicability to a wide range of rotating machines. To overcome this limitation, we developed a deep learning-based estimation model and demonstrated that its performance depends on the distribution of time-domain features in the training data, which are associated with flaking damage. We then analyzed the manner in which these feature distributions impose limitations on the estimation accuracy of the model. Additionally, we incorporated explainability using Grad-CAM to verify that the extracted features were aligned with the physical phenomena of flaking damage, thereby confirming the link between the feature vibrations and estimation results. Our experiments under various training–test split conditions indicate that time-domain shifts of these features affect the model's performance, providing insight into how feature distributions constrain the estimation of the flaking size.

1. INTRODUCTION

To optimize the performance of rotating machines, it is essential to regularly assess and diagnose its condition and maintain it at an appropriate time. One of the most critical targets for maintenance is the detection of rolling bearing faults, which are mechanical components subjected to loads in the rotating parts of a machine. Among the various types of rolling bearing faults, flaking on the raceways of the bearings is the most common. The diagnosis of flaking involves identifying periodic shocks in bearing vibrations (Randall & Antoni, 2011).

The development of a method for diagnosing rolling bearing flaking involves estimating the remaining useful life (RUL) by determining the size of flaking through vibrations. As the flaking size increases with continued operation after the onset of flaking, it can cause severe problems with the rotational accuracy, vibration, and acoustics of machines. Using this diagnosis method, severe damage can be avoided, which is particularly beneficial for machinery with high maintenance costs such as wind turbines. A common approach for estimating the flaking size is to measure the vibration of the rolling elements of the bearing as they enter and exit flaking and then calculate the interval between the two events (Sawalhi & Randall, 2011). Because the flaking progression rate accelerates sharply when the flaking size exceeds the rolling element pitch interval, estimating the flaking size within the range below this interval has been confirmed to be effective in reducing operational risks (Maekawa, Mizokuchi, Taguchi, Miyasaka, & Shibasaki, 2018).

However, estimating the flaking size from vibrations on a rule-based basis is time consuming and costly because it usually requires trial and error with high expertise. High expertise in installing appropriate vibration sensors and tuning the parameters of noise reduction methods in vibration improves the detectability of feature vibrations for flaking-size estimation. The difficulty of detecting vibrations, particularly those of rolling elements entering flaking, was mentioned in (Smith, Hu, Randall, & Peng, 2015). Moreover, (F. Zhang, Huang, Chu, & Cui, 2020) highlighted that when the flaking size expands beyond the rolling element pitch interval,

https://doi.org/10.36001/IJPHM.2025.v16i3.4163

the overlapping of the feature vibrations resulting from the simultaneous entry and exit of multiple rolling elements further complicates the size estimation process.

We propose a deep learning-based model for estimating the flaking size in rolling bearing vibrations, with the aim of reducing diagnostic costs and leveraging the expertise of highly skilled professionals. Previous studies have mainly focused on using deep learning models to classify the flaking-damaged parts of the rolling bearings. (W. Zhang, Peng, Li, Chen, & Zhang, 2017) proposed deep convolutional neural networks with wide first-layer kernels (WDCNN) that utilized 1D vibration acceleration waveforms as input and employed large kernel sizes in shallow layers. Lu et al. (Lu et al., 2023) proposed a Pulse Induction Convolutional Neural Network (PICNN) that used the envelope spectrum of the vibration acceleration waveform as the input, with weights assigned based on the impact vibration period in the case of bearing flaking. We applied the CNN-LSTM model to a variety of flaking size vibration dataset in (Yoshimatsu, Taguchi, Yoshihiro, & Yairi, 2023) to estimate the flaking size. The previous study did not address the distribution of the feature vibrations of flaking sizes within the training data.

This study investigated the effect of feature vibration interval distribution on the performance of a model that estimates the flaking size. The variations in the distribution were caused by changes in the flaking size and operational parameters of the training data. The investigation focused on understanding how the time-domain feature distribution affects the model performance. A CNN-LSTM architecture was employed for the flaking size estimation model. The model's training data incorporated test data featuring artificially created defects on the inner ring of the cylindrical roller bearing as well as flaking that developed through ongoing operation. In addition, Grad-CAM (Selvaraju, Cogswell, Das, & others, n.d.) was used to confirm whether the feature vibrations were related to the estimated results, similar to rule-based methods that require high expertise.

The contributions of this study are as follows:

1. The relationship between the performance of the flaking size estimation model and the distribution of time-domain feature vibrations in the training data was verified.
2. The limitations imposed on the performance of the estimation model were analyzed using the distribution of time-domain features when such relationships were present.
3. The association between the feature vibrations related to the physical phenomena and estimation results were verified by incorporating explainability into the model.

The remainder of this paper is organized as follows. Section 2 provides an overview of the proposed method, evaluation process, and dataset. Section 3 presents the evaluation results, and Section 4 presents the conclusions.

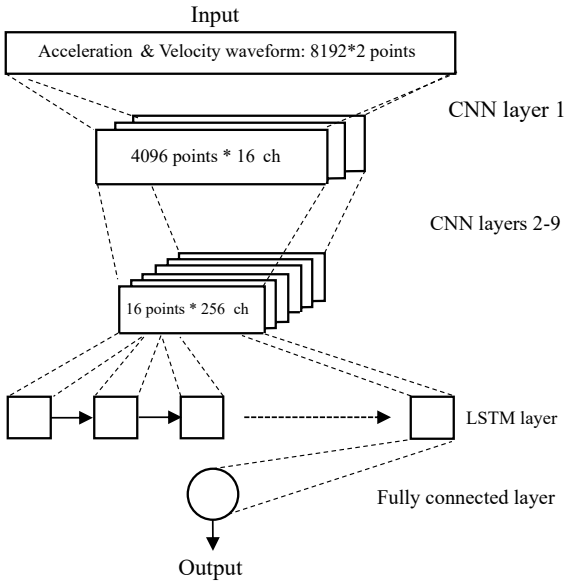


Figure 1. Overview of proposed flaking size estimation CNN-LSTM model.

2. METHODOLOGY

In this section, we present the proposed deep learning-based methodology for estimating the flaking size of rolling bearings. Section 2.1 discusses how multiple feature vibrations derived from 1D time-domain signals are captured through a combined CNN-LSTM framework. Section 2.2 explains how Grad-CAM to interpret the focus of the trained model on flaking-induced vibrations. Section 2.3 outlines the overall architecture of the proposed model. Section 2.4 describes the contents of the dataset used. Finally, section 2.5 and 2.6 describes the train-test procedure and data scenarios for changing the distribution of time-domain feature vibrations for evaluation.

2.1. CNN-LSTM Model for Multiple Feature Vibrations in 1D Time Waveforms

A convolutional neural network (CNN) (Lecun, Bottou, Bengio, & Haffner, 1998) is widely used in rolling bearing diagnostics, particularly because it can strongly extract frequency-domain features. CNN excel at capturing short-duration shock components, which often indicate the presence of flaking. However, CNN are constrained by their limited receptive field. This limitation makes it difficult for a CNN to learn the relationships among feature vibrations that occur far apart in the time domain. (J. Chen et al., 2021) applied a multiscale CNN to deal with changes in impact vibration intervals due to differences in the parts on which flaking occurred on rolling bearings. Few studies have considered the distant or diverse

event intervals.

To address this challenge, previous studies have explored the use of sequence models that retain temporal information, or methods that add positional information to the model during data processing. Sequence models include RNN (Elman, 1990), LSTM (Hochreiter & Schmidhuber, 1997), and others, such as phased-LSTM (Neil, Pfeiffer, & Liu, 2016), a variant of LSTM that considers periodic events. CoordConv was used to add a position channel to the input data (Liu et al., 2018). Moreover, transformer-based approaches leverage self-attention to capture the relationships among features at distant positions by incorporating position encoding to preserve the sequence information (Vaswani et al., 2017). Although these positional methods have not yet been widely adopted for rotating machinery diagnosis, there are studies that apply transformers with rotary position embedding (Su et al., 2024) to bearing-fault detection (Zhou & Farimani, 2023). However, few studies have directly addressed the impact of the time-domain position of the multiple-feature vibrations on the output of the trained model.

Given the goal of flaking size estimation and the known efficacy of CNNs in extracting critical frequency-domain signatures, we adopted a CNN-LSTM model. The CNN identifies localized shocks, whereas the LSTM layer captures longer time-domain dependencies, which are important for measuring the intervals between shocks. We intentionally omitted the integration of position-aware techniques such as CoordConv or rotary position embedding. By excluding these, we focus on examining how time-domain shifts in the input signals, caused by changes in flaking size, affect model performance in a simpler, more controlled setting.

The CNN-LSTM architecture is suitable for deployment in condition monitoring environments for rotating machines. Although it features a reduced parameter count compared to models with advanced time-domain processing capabilities, such as transformers, it incurs higher computational costs and relies on sequential processing relative to conventional machine learning approaches. Consequently, this is less appropriate for high-speed execution on memory-constrained edge devices. However, condition monitoring for systems such as wind turbines is typically performed at intervals of ten minutes or longer using Supervisory Control And Data Acquisition (SCADA) systems, which are adequate for tracking drivetrain damage that progresses over several hours to months (Shi, Liu, & Gao, 2021). Therefore, when diagnostics are executed on server-class hardware, the memory footprint and processing time of the proposed approach remain within the practical limits. These considerations substantiate the efficacy of the CNN-LSTM framework for flaking size estimation in real-world applications.

2.2. Explainability via Grad-CAM

We employ Gradient-weighted Class Activation Mapping (Grad-CAM) to investigate whether the intervals between flaking-induced shocks affect flaking-size estimation. Previous studies on deep learning-based rolling bearing diagnostics have mainly explored classification tasks, verifying whether model decisions correlate with periodic shocks that are also recognized by rule-based methods. For instance, (Li, Zhang, & Ding, 2019) introduced an attention mechanism to a one-dimensional vibration waveform classification model and confirmed that high attention weights coincide with periodic impacts. (B. Chen, Liu, He, Liu, & Zhang, 2022) proposed GS-CAM to visualize the relationship between periodic impacts and the classification outputs in more detail.

However, few studies have discussed how such periodic shock vibrations are directly related to flaking size estimation performance in deep learning models. In this study, we applied Grad-CAM to determine whether the CNN-LSTM model highlights the time-domain feature vibrations of the rolling element entry and exit into the flaking area when estimating the flaking size. By comparing the input waveforms with the Grad-CAM outputs, we verified whether the features learned by the model were associated with impact vibrations caused by rolling elements entering and exiting flaking. A quantitative evaluation based on the precise time-domain positions of the shock vibrations is desirable. Because obtaining such detailed positional data on a large scale requires high-cost systems, such as absolute angle sensors or high-speed cameras, which are not available in our dataset, we opted to manually analyze a subset of the data. This approach allowed us to identify the time-domain position of the feature vibrations and use this information for visualization. We hypothesize that this visualization step demonstrates that the flaking size estimation of the proposed model is based on physical events, similar to those targeted by rule-based methods.

2.3. Proposed Model Architecture

Figure 1 illustrates the overall structure of the proposed CNN-LSTM model. The model is designed to estimate the flaking size from rolling bearing vibration signals and comprises three main components: Feature Extractor (CNN), LSTM layer, and regressor. Two channels of 1D signals were input into the model: the original vibration acceleration waveform, and a velocity-equivalent waveform obtained by integrating the acceleration signal. This dual-channel input is motivated by physical principles from rule-based flaking-size estimation, where low-frequency velocity-dominated signals are associated with flaking entry, and high-frequency acceleration-dominated signals are tied to the flaking exit (Maekawa et al., 2018).

The feature extractor was composed of nine convolutional blocks, each containing the following layers: a convolutional

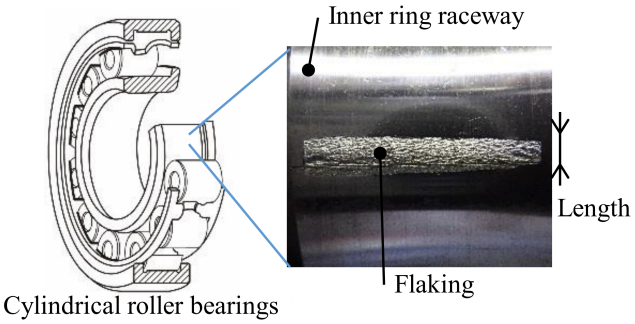


Figure 2. Example of progressed flaking resulting from artificial defect.

layer with a wide-to-narrow kernel (WDCNN), swish activation, an average pooling layer, batch normalization (Ioffe & Szegedy, 2015), and a squeeze-and-excitation (SE) block (Hu, Shen, & Sun, 2018). This architecture employs progressively narrower kernels to capture both broad- and fine-scale vibration characteristics. The model effectively isolates the low-frequency and high-frequency components associated with flaking entry and exit events using large kernels in the initial layers and smaller kernels in the deeper layers. The applied activation function swish is expressed by the following equation,

$$Swish(x) = x \cdot \sigma(x) \tag{1}$$

where σ denotes a sigmoidal function. Swish was used instead of ReLU because its smooth, non-monotonic shape empirically yields a better performance and facilitates richer feature extraction. Average pooling reduces the temporal dimension of the feature maps and mitigates overfitting by providing coarse-level features to subsequent layers. Batch normalization stabilizes and accelerates the training by normalizing the outputs of the convolutional layers, thereby reducing the internal covariate shift. The SE block adaptively recalibrates channel-wise feature responses, emphasizing the signal components that are highly relevant to the flaking damage.

Following the feature extraction, concatenated features are fed into the LSTM layer. LSTM integrates the time-domain context by retaining pertinent information over multiple time steps. This property is crucial for capturing the sequential nature of flaking events, where the time interval between shocks can indicate flaking size. The hidden states of the LSTM track how these feature intensities change over time, allowing the network to learn the relationship between repeated shocks and flaking progression.

The output of the fourth convolutional block was targeted for the Grad-CAM. At this stage, the original 8,192-point waveform is downsampled to 512 points, thereby producing a similarly reduced resolution in the Grad-CAM result.

2.4. Dataset

To train and evaluate the proposed model, a vibration dataset was acquired using a cylindrical roller bearing, specifically the NU2228EM (NSK) model. This bearing was intentionally machined to include artificial defects in the inner ring to simulate flaking, and the size of the flaking progressed during the long-term operation (Figure 2). The dimensions of the artificial defects were 45 mm axial length, 0.3 mm circumferential length, and 0.2 mm groove depth. The test conditions were a radial load of 205 kN (P/C = 0.35), a rotational speed of 1,500 min^{-1}, and circulating lubrication. The dataset consisted of the vibration data acquired from the housing of the test rig. The bearings in which the flaking occurred were operated under various conditions. The runs were designed to replicate real-life scenarios in which flaking had progressed, thereby ensuring a comprehensive dataset.

During the experiments, the test rig was stopped several times to measure the size of the flaking and to acquire vibration data corresponding from 0.03 up to 1.44 times the rolling element spacing pitch. After each measurement, the test rig was restarted under a range of conditions, including variations in load, rotational speed, and sensor placement. This systematic variation in the operational conditions enabled the acquisition of vibration data corresponding to a wide array of flaking sizes and configurations. Importantly, for each measured flaking size, multiple datasets were recorded under different operational parameters, which enhanced the variability of the dataset.

The key operating conditions included three rotational speeds (1,200, 1,500, and 1,750 min^{-1}) and seven radial load levels ranging from 29.3 kN to 205k N. Vibration data were acquired in both radial and axial directions using sensors configured at a sampling frequency of 96 kHz over a duration of 30 s per test. The dataset includes measurements from two test pieces, each subject to different operational protocols, resulting in varying numbers of measurements for each test piece. Table 1 summarizes the test conditions and measurement parameters used for data acquisition.

This dataset provides a critical resource for understanding the relationship between the flaking size and time-domain feature vibrations. In particular, (Maekawa et al., 2018) confirmed using a rule-based method that the interval of time-domain feature vibrations expands with the progression of the flaking size during testing in this dataset. Consequently, variations in the train–test split scenarios allow control over the distribution of these time-domain features in the training data. The influence of time-domain feature distribution on the flaking-size estimation performance of the proposed model was then compared across multiple split scenarios.

Table 1. Operating and measurement conditions of test for the various flaking size dataset

Bearing number	NU2228EM(NSK)
Bearing type	Cylindrical roller
Rotational speed	3 speeds(1,200, 1,500, 1,750 min^{-1})
Load	7 radial loads(29.3 k - 205 kN)
Test pieces num.	2
All condition num.	220 (Test piece A: 52, B: 168)
Sensor type	Accelerometer
Sensor direction	Radial, Axial (On housing)
Sampling frequency	96,000 Hz
sampling time	30 s

Table 2. Training conditions

Epoch num.	50
Learning rate	2.5e^{-4} to 1.0e^{-3} (WarmUp)
Optimizer	Adam
Loss function	Mean squared error (MSE)
Train frames num.	61,776 (or 18,200)
Length of frame	8,192
Mini-batch frames num.	32

2.5. Train-Test Procedure

To evaluate the effect of the time-domain feature distribution within the training data on the performance of the proposed model in estimating the flaking size, K-fold cross validation was conducted under three distinct scenarios. In each scenario, the dataset was divided into several folds. Although cross validation typically involves separate test data, this study utilized split-fold data for test purposes to maintain the quality and quantity of flaking size. The remaining folds were used to train the model.

- Scenario A: The dataset was randomly split into five folds such that all data corresponding to the same flaking size were assigned to one fold to avoid data leakage. In this scenario, the domain shift of the time-domain features between the folds is expected to be small. This scenario replicates the methodology of a previous study by the authors, which confirmed a high flaking size estimation performance and the extraction of time-domain features owing to the physical phenomena in (Yoshimatsu et al., 2023).
- Scenario B: The dataset was sorted in ascending order of the flaking size and divided into five folds. This resulted in similar values of the flaking size within each fold, indicating that there was a domain shift in the time-domain features between the folds.
- Scenario C: The dataset was partitioned into two folds based on the two test pieces employed in the experiment, with data from each test piece allocated to distinct folds comprising 52 and 168 operating conditions, respectively. Although the domain shift in the time-domain feature distributions between the folds was minimal, there was an imbalance in the number of operating conditions represented in the training and test data.

For each operating condition, the vibration data were segmented into frames consisting of 8,192 points. An equal number of frames were randomly selected from each file corresponding to a given condition, resulting in a maximum of 61,772 frames used for training. For training runs that involved extensive hyperparameter searches, such as those in subsequent ablation and comparative studies, a subset of 18,200 frames was used. Even with different frame counts across conditions, the proportion of data corresponding to each test condition in the training set was maintained so that the time-domain feature distribution remained unchanged. The output of the flaking size estimation model was defined as the logarithmic ratio of the feature vibration interval corresponding to the flaking size to the frame length, as expressed by the following equation:

$$Y = -\ln\left(\frac{P_{\text{flaking}}}{P_{\text{frame}}}\right) \tag{2}$$

where Y is the true output value. P_{flaking} denotes the number of data points equivalent to the ground-truth flaking size. P_{frame} denotes the total number of data points in a frame. This output formulation was chosen to mitigate the learning bias caused by a large amount of data with extremely short feature vibration intervals relative to the frame length, thereby suppressing biased output values for the model.

In each scenario, the data from each fold were used as the test data, and the remaining folds were used as the training data. The train-test was repeated for the number of folds. The performance of the model under different scenarios was systematically analyzed to assess the influence of the data-splitting strategies on estimation accuracy and generalizability. This comprehensive approach enabled a robust evaluation of the adaptability of the proposed methodology to variations in time-domain feature distributions. The training conditions, including the number of epochs, learning rate scheduling, and batch size, are listed in Table 2.

2.6. Ablation and Comparative Evaluation

An ablation study was conducted to elucidate the impact of the individual technical elements incorporated into the proposed method, as well as the overfitting prevention strategies, on the performance of the model. In this study, we evaluated the contributions of the additional speed channel and SE

block to investigate how these components enhance flaking size estimation. Moreover, we examined the effects of overfitting mitigation techniques and data augmentation by analyzing variations in estimation performance when batch normalization, dropout, L2 weight decay, average pooling and window sliding were added or removed. In all cases, cross validation was employed for training and evaluation to clearly delineate the influence of each component. Ultimately, this evaluation aimed to derive insights for selecting the optimal technical elements and for effective model improvement in the proposed framework.

Furthermore, to evaluate the suitability of the proposed method, we performed a comparative evaluation against alternative architectures, the details of which are presented in Table 3. For comparison, we selected a transformer encoder capable of handling time-domain features through an attention mechanism and position encoding and a CNN-based WDCNN, which is widely utilized in rolling bearing diagnostics. In the transformer encoder model, a CNN-based embedding layer was employed to extract local features (Zhou & Farimani, 2023), and the resulting sequential data were fed into the transformer encoder. In contrast, the CNN-based model adopts a WDCNN architecture with an enlarged kernel size in the convolution layer near the input. Both models incorporate a structure that combines CNN-based local feature extraction with an expanded receptive field in the time domain. We conducted cross validation on these architectures and, after optimizing the hyperparameters, compared their flake size estimation performance to determine the suitability of each model for evaluating the influence of time-domain feature distributions.

Table 3. Comparative Models List

Model name	Local feature extractor	Time-domain feature extractor
WDCNN	CNN	Wide kernel on 1st CNN layer
Transformer	CNN	Transformer encoder
CNN-LSTM	CNN	LSTM

3. RESULTS

3.1. Results in Scenario A

Figure 3 shows the estimated and truth flaking lengths for each test dataset under five-fold cross validation performed in scenario A. This figure shows the flaking sizes (both predicted and ground truth) as a ratio of the rolling-element pitch to compare the results across all operational conditions. The estimation error, defined as the difference between the estimated and true values, has a mean average error of 0.039 pitch for all 220 conditions, confirming its useful performance in facility operational decision making. However, the estimation performance tended to decline for flaking sizes larger than 1.4 pitches. These large flaking sizes corresponded to the operating conditions closest to extrapolation. In flaking-size estimation, maintaining accuracy for sizes below 1 pitch is critically important for predicting the remaining useful life of bearings. Therefore, the impact of reduced extrapolation performance on excessively large flaking sizes is limited. Overall, these findings confirm that a deep learning model can accurately estimate flaking sizes from bearing vibration data under a random train-test split that does not consider the time-domain feature distribution.

Figure 4 shows an example of the Grad-CAM result when test data with a flaking size of 0.769 pitch, rotational speed of 1,200 min^{-1}, and an interval of approximately 376 points between the feature vibrations were input to the trained model in scenario A. The estimated flaking size was 0.719 pitch, which indicates a high estimation accuracy. The top and bottom rows show 8,192 points of input normalized acceleration and velocity, and the middle row shows 512 points of Grad-CAM results, with these waveforms corresponding to the time-direction position. A part of each row has a colored area, which indicates the actual timing of the rolling element entry and exit from flaking. The Grad-CAM results show the importance of the feature vibration when the rolling element enters and exits the flaking region. These results indicate that the trained model with a high size estimation performance in scenario A emphasizes physically meaningful features, similar to those used in expert rule-based methods for flaking size estimation.

Figure 5 presents an example of the Grad-CAM result obtained in scenario A when the trained model received input with data containing a flaking size of 1.440 pitch, a rotational speed of 1,500 min^{-1}, and a feature vibration interval of approximately 564 points. The model estimated a flaking size of 1.304 pitch, demonstrating reasonable accuracy even for damage exceeding the 1.0 pitch criterion. In this figure, the top and bottom display the 8,192 point normalized acceleration and velocity signals, respectively, whereas the middle panel shows the Grad-CAM output at a resolution of 512 points along the time axis. The colored areas in the figure highlight the timing of the two rolling elements entering and exiting the flaking area, with the Grad-CAM activation maps emphasizing the importance of the corresponding feature vibrations. These results indicate that the high-performance model in scenario A effectively aggregates feature vibrations from multiple rolling elements to estimate larger flaking sizes, a task that has proven challenging for previous rule-based approaches.

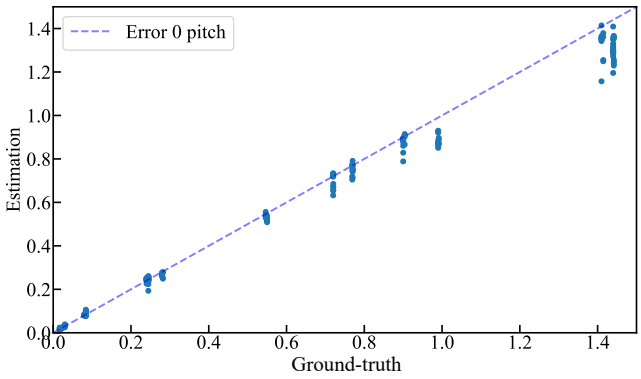


Figure 3. Comparison of estimates and truth of flaking size compared to pitch between rollers for each tests in scenario A. Pitch basis MAE = 0.039.

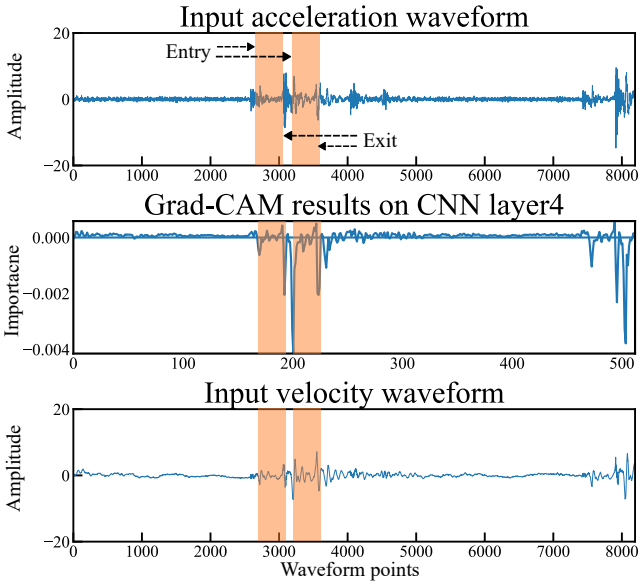


Figure 4. Example of Grad-CAM results in scenario A. Flaking size: Truth: 0.769 pitch, Estimation: 0.719 pitch

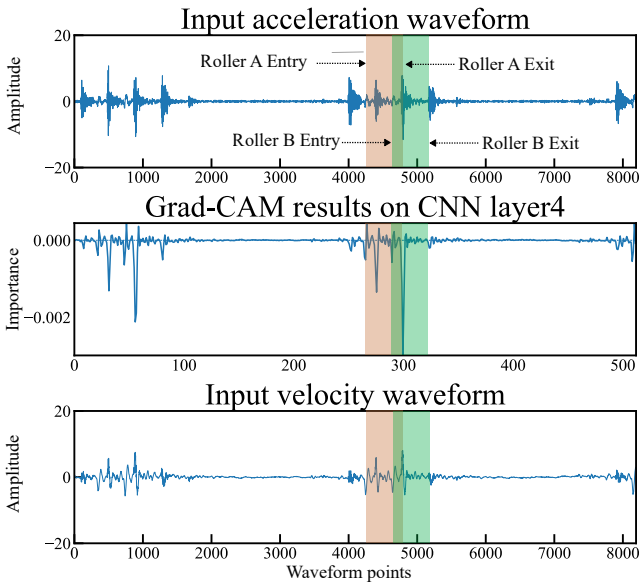


Figure 5. Example of Grad-CAM results in scenario A. Flaking size: Truth: 1.440 pitch, Estimation: 1.304 pitch

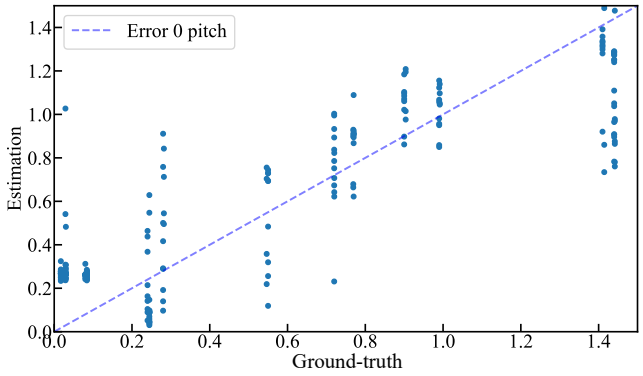


Figure 6. Comparison of estimates and truth of flaking size compared to pitch between rollers for each tests in scenario B. Pitch basis MAE = 0.239.

3.2. Results in Scenario B

Figure 6 shows the estimated and ground-truth flaking lengths for each test data under scenario B. A five-fold cross validation arrangement was designed such that the flaking-size distributions in the training data differed from those in the test data. Compared to scenario A, the estimation performance across a wide range of flaking sizes was lower in scenario B. The effect of the domain shift on the time-domain features between the train and test data is likely. This reduction in estimation performance adversely affects the prediction of the remaining life because errors in smaller flaking sizes are amplified over time. These results confirm that the proposed model has limitations when attempting to estimate flaking sizes that are not well represented within the time-domain feature distribution of training data.

Figure 7 shows an example of the Grad-CAM result for the test data frame with a flaking size of 0.08 pitch and a rotational speed of 1,200 min^{-1} under scenario B. The estimated flaking size was 0.238 pitch, which was an error. The Grad-CAM results showed that the importance was correctly high at the time of the flaking entry point of the colored area, whereas the colored exit timing was higher at a different position than the exit timing. This mismatch suggests that when the training data do not include similar time-domain feature distributions, the model struggles to extract the correct feature vibrations, leading to reduced estimation performance.

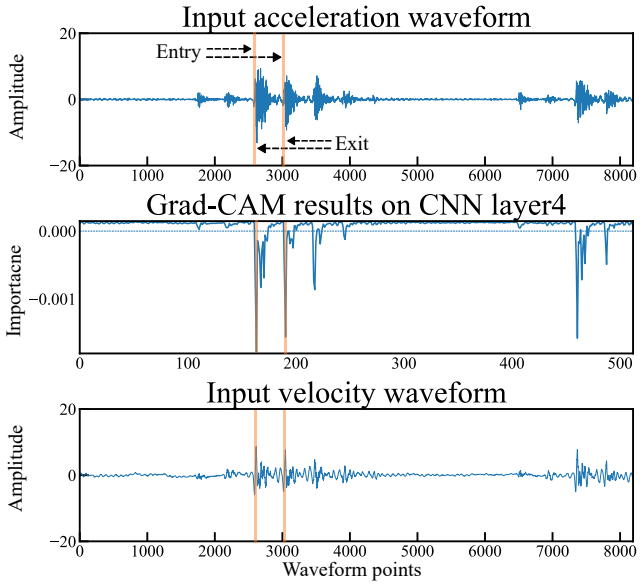
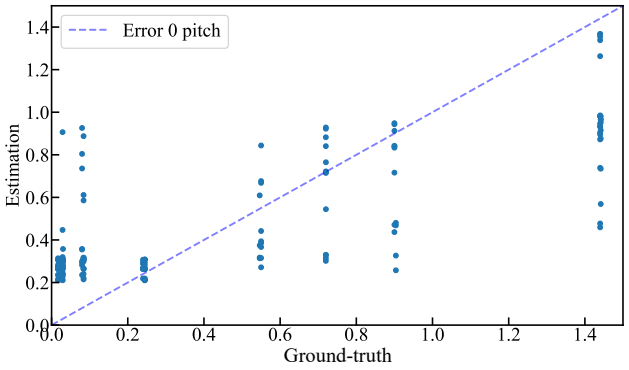
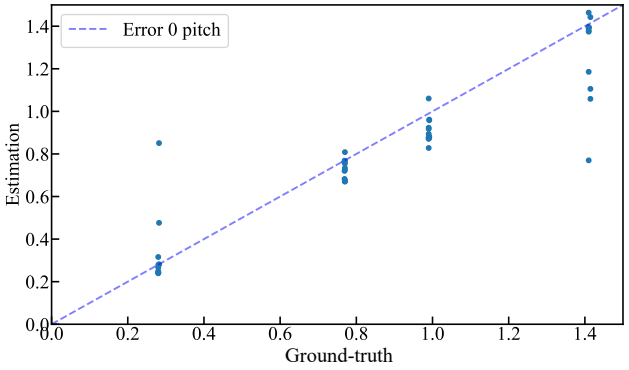


Figure 7. Example of Grad-CAM results in scenario B. Flaking size: Truth: 0.08 pitch, Estimation: 0.238 pitch



(a) Train: Test piece B, Test: Test piece A(168 conditions), Pitch basis MAE = 0.265



(b) Train: Test piece A, Test: Test piece B(52 conditions), Pitch basis MAE = 0.113

Figure 8. Comparison of estimates and truth of flaking size compared to pitch between rollers for each tests in scenario C: Data were split by test piece.

3.3. Results in Scenario C

Figure 8 shows the estimated and truth flaking size in scenario C (i.e., two-fold cross validation), where the dataset was split based on the test pieces. In figure 8a, the distribution of time-domain features is limited to 52 conditions, which are the number of flaking sizes and operating conditions of the training data. Therefore, the estimation performance of the test data was low (MAE = 0.265). By contrast, as shown in Figure 8b, when the number of flaking sizes and operating conditions included in the training data was 168, the performance of the flaking size estimation in the test data was improved (MAE = 0.113). These results indicate that even moderate improvements in the distribution of time-domain features and number of conditions present in the training data can improve the performance of the model.

Figure 9 shows an example of the Grad-CAM results for scenario C, which was trained with test piece data for 168 flaking sizes and operating conditions. The input data had a flaking size of 0.77 pitch and a rotational speed of 1,200 min^{-1}. The estimated flaking size was 0.756 pitch, which is a high estimation performance. The model correctly extracted the time-domain positions corresponding to both events of the rolling element entering and exiting flaking. This feature extraction trend was similar to that in scenario A, where the fold data were randomly split. This confirms that enhancing the time-domain feature distribution in the training data allows the trained model to effectively extract features that are in accordance with physical phenomena.

3.4. Results in Ablation and Comparative Studies

Table 4 presents the results of the ablation study in which nine variants of the proposed model were evaluated under the same data-split conditions as in scenario A. Specifically, the study compared a full model incorporating all seven technical components, a baseline model omitting all of them, and seven additional models, each with one element removed from the full configuration. The performance was quantified using the MAE of the estimated flaking size, normalized by the rolling element pitch. The results indicated that models incorporating individual components, such as the velocity waveform channel, SE block, and batch normalization, consistently yielded improved estimation performance and mitigated overfitting compared with the baseline. In particular, the exclusion of batch normalization led to a marked decline in performance, whereas dropout and data augmentation via window sliding, although offering moderate benefits, did not exert a strong influence.

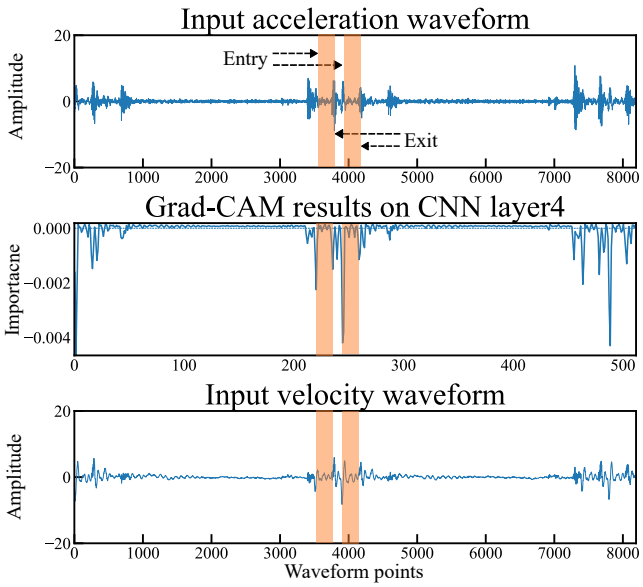


Figure 9. Example of Grad-CAM results in scenario C 168 conditions train. Flaking size: Truth: 0.770 pitch, Estimation: 0.756 pitch

Table 4. Ablation study results with test metric in scenario A for the proposed model variants

Variants name	Pitch basis MAE
With all elements	0.0445
Without all elements	0.0814
Without velocity ch.	0.0434
Without SE block	0.0428
Without drop out	0.0477
Without batch normalization	0.1261
Without average pooling (with max)	0.0509
Without window sliding	0.0490
Without weight decay (L2)	0.0541

Table 5 presents a comparative evaluation of three distinct models—WDCNN, transformer encoder, and the proposed CNN-LSTM—applied under identical data split conditions in scenario A. The evaluation metric was defined as the MAE, which was computed as the ratio of the estimated flaking size to the rolling element pitch in the test data. All methods incorporated effective overfitting prevention techniques, such as batch normalization, as evaluated in our ablation study. Notably, the proposed CNN-LSTM model achieved superior performance in flaking size estimation compared to the transformer encoder, which is recognized for its efficiency in handling temporal information, and the CNN-based WDCNN. These results substantiate that the CNN-LSTM architecture is well suited to the dataset scale and extraction of time-domain features in this study, providing an appropriate framework for examining the influence of time-domain feature distributions.

Table 5. Compalative study results with test metric in scenario A for the proposed model variants

Model name	Pitch basis MAE
WDCNN	0.0565
Transformer	0.0937
CNN-LSTM (Proposed)	0.0445

3.5. Discussion

In summary, these results demonstrate that the distribution of time-domain features in the training data critically affects the ability of the proposed model to estimate flaking sizes in rolling bearings. The estimation performance remained high under random splits without significant domain shifts. However, when substantial portions of the time-domain feature distribution are absent from the training data, the estimation errors increase, and the Grad-CAM results reveal a focus on nonmeaningful vibrations. Furthermore, improving the variety of operational conditions in the training set can overcome these limitations, thereby increasing the performance of the flaking size estimation and aligning the extracted features with those used in rule-based approaches.

Finally, it is important to note that all experiments in this study were performed on a single test rig with a fixed bearing specification, where changes in the frequency-domain features were relatively small. To ensure broader applicability for industrial use, future work should simultaneously consider the frequency- and time-domain feature distributions under diverse machine, installation, and measurement conditions. Additionally, verifying the benefits of domain generalization with broader datasets may not only improve flaking-size estimation but also advance fault diagnosis and remaining useful life prediction across various types of rotating machines.

4. CONCLUSION

We proposed a CNN-LSTM model for estimating the flaking size of rolling bearings from vibration signals. We controlled the distribution of time-domain features in the training data by splitting the dataset under multiple scenarios. We verified that the estimation performance improved when the time-domain feature distribution was diverse and the domain shift between the training and test data was reduced. We also applied Grad-CAM to confirm that the features contributing to the accurate estimation were aligned with the physically meaningful vibrations observed in rule-based diagnostics. In scenarios with a high estimation accuracy, the model highlights the essential shock events corresponding to the entry and exit of rolling elements into the flaked area.

Furthermore, practical condition monitoring of rotating machines requires comprehensive coverage of both time- and frequency-domain features. Consequently, it is necessary to assess the domain generalization capability of the model across diverse conditions including various bearing types, sizes, damage states, operating parameters, and measurement setups. In addition, enhancing the dataset with a precise time-domain damage position is necessary for large-scale quantitative assessments of the extracted features using explainable AI. These efforts aim to broaden the range of feature distributions and facilitate the industrial implementation of the proposed approach. In future research, the robustness of this method can be further enhanced by applying techniques such as representation learning, domain adaptation, and domain generalization to datasets acquired under more varied conditions, thereby extending its applicability to a wide array of rotating machines.

REFERENCES

Chen, B., Liu, T., He, C., Liu, Z., & Zhang, L. (2022, June). Fault diagnosis for limited annotation signals and strong noise based on interpretable attention mechanism. *IEEE Sens. J.*, *22*(12), 11865–11880.

Chen, J., Huang, R., Zhao, K., Wang, W., Liu, L., & Li, W. (2021). Multiscale convolutional neural network with feature alignment for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.*, *70*, 1–10.

Elman, J. L. (1990, March). Finding structure in time. *Cogn. Sci.*, *14*(2), 179–211.

Hochreiter, S., & Schmidhuber, J. (1997, November). Long short-term memory. *Neural Comput.*, *9*, 1735–1780.

Hu, J., Shen, L., & Sun, G. (2018, June). Squeeze-and-excitation networks. In *2018 ieee/cvf conference on computer vision and pattern recognition* (pp. 7132–7141). IEEE.

Ioffe, S., & Szegedy, C. (2015, February). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, *abs/1502.03167*, 448–456.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998, November). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, X., Zhang, W., & Ding, Q. (2019, August). Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Signal Processing*, *161*, 136–154.

Liu, R., Lehman, J., Molino, P., Such, F. P., Frank, E., Sergeev, A., & Yosinski, J. (2018, July). An intriguing failing of convolutional neural networks and the CoordConv solution. *arXiv [cs.CV]*.

Lu, H., Pavan Nemani, V., Barzegar, V., Allen, C., Hu, C., Laflamme, S., ... Zimmerman, A. T. (2023, May). A physics-informed feature weighting method for bearing fault diagnostics. *Mech. Syst. Signal Process.*, *191*, 110171.

Maekawa, T., Mizokuchi, H., Taguchi, K., Miyasaka, T., & Shibasaki, K. (2018). Spall length estimation of rolling element bearings by using vibration signal. *Proc. Symp. Eval. Diagn.(In Japanese)*, *2018.17*(0), 110.

Neil, D., Pfeiffer, M., & Liu, S.-C. (2016, October). Phased LSTM: Accelerating recurrent network training for long or event-based sequences. *arXiv [cs.LG]*.

Randall, R. B., & Antoni, J. (2011, February). Rolling element bearing diagnostics—a tutorial. *Mech. Syst. Signal Process.*, *25*(2), 485–520.

Sawalhi, N., & Randall, R. B. (2011, April). Vibration response of spalled rolling element bearings: Observations, simulations and signal processing techniques to track the spall size. *Mech. Syst. Signal Process.*, *25*(3), 846–870.

Selvaraju, R. R., Cogswell, M., Das, A., & others. (n.d.). Grad-cam: Visual explanations from deep networks via gradient-based localization.

Shi, Y., Liu, Y., & Gao, X. (2021). Study of wind turbine fault diagnosis and early warning based on SCADA data. *IEEE Access*, *9*, 124600–124615.

Smith, W. A., Hu, C., Randall, R. B., & Peng, Z. (2015). Vibration-based spall size tracking in rolling element bearings. In *Proceedings of the 9th iftomm international conference on rotor dynamics* (pp. 587–597). Springer International Publishing.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., & Liu, Y. (2024, February). RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, *568*(127063), 127063.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017, June). Attention is all you need. *Neural Inf Process Syst*, 5998–6008.

Yoshimatsu, O., Taguchi, K., Yoshihiro, S., & Yairi, T. (2023, September). Size estimation of flaking in rolling bearings using deep learning with explainability. *PHMAP_CONF*, *4*(1).

Zhang, F., Huang, J., Chu, F., & Cui, L. (2020, December). Mechanism and method for the full-scale quantitative diagnosis of ball bearings with an inner race fault. *J. Sound Vib.*, *488*, 115641.

Zhang, W., Peng, G., Li, C., Chen, Y., & Zhang, Z. (2017, February). A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors*, *17*(2).

Zhou, A. Y., & Farimani, A. B. (2023, December). FaultFormer: Pretraining transformers for adaptable bearing fault classification. *IEEE Access*, *12*, 70719–70728.