Machine Fault Diagnosis Based on a Multi-Input Multi-Branch Deep Learning Network with Attention Mechanisms

Lianghui Zou[1], Haoxin Qin[2], Qidong Lu[3], and Zhiliang Qin[1,4,*]

[1]Weihai Beiyang Electrical Group Co., Ltd, Weihai, Shandong, China
[2]Weihai Ivy Foreign Language School, Weihai, Shandong, China
[3]Weihai Research Institute of Industry Technology, Shandong University, China
[4]School of Mechanical, Electrical and Information Engineering, Shandong University, China
Email: zoulianghui@beiyang.com (L.H.Z.); qinhaoxin73@gmail.com (H.X.Q.); lqd19922012@163.com (Q.D.L.); qinzhiliang@beiyang.com (Z.L.Q.)
*Corresponding author
Manuscript received December 12, 2023; revised January 12, 2024; accepted January 30, 2024; published March 22, 2024

***Abstract*—In this paper, we propose a multi-input attention-based deep-learning architecture for machine fault diagnosis, which is a challenging task due to environmental noises and signal interference that are inevitable in practical industry environments. Specifically, we develop an attention mechanism to focus on the most effective signal characteristics and achieve highly accurate fault classifications under various working conditions. First, we construct multi-dimensional features to characterize both time-domain and frequency-domain properties of machine vibration signals. Afterwards, we design a novel multi-input multi-branch (MIMB) architecture incorporating multiple sub-networks to enhance the learning of discriminant capabilities. The derived features from each sub-network are fused to form an input to the final classification layer. To verify the effectiveness of the proposed approach, we conduct comprehensive experiments on the bearing database of Universität Paderborn, Germany, which is generally recognized as the benchmark to compare the performance of various algorithms. Numerical results show that the proposed approach achieves the state-of-the-art classification accuracy and has a noticeable performance gain over the previous schemes in the literature.**

***Keywords*—fault diagnosis, feature extraction, attention mechanism, multi-input network**

I. Introduction

Machines play indispensable roles in modern industrial production. Machine Fault diagnosis is of great significance to avoid the occurrence of accidents and ensure the stability of manufacturing activities. Most papers in the literature on machine fault analysis are based on traditional analytical approaches, e.g., graph theories and expert systems [1, 2]. In recent years, data-driven techniques become increasingly important and attract significant attentions of academic researchers and practitioners in the industry. In particular, deep learning has made breakthroughs in the fields of image and natural language processing. It has been widely applied to various fields for its powerful capability to analyze complex data and offer effective approaches to predict precisely short-term working conditions and Remaining Useful Life (RUL) of machinery components [3, 4].

In general, machine fault diagnosis can be achieved from the following perspectives: (1) Acoustic emission detection to detect sound activities based on the recorded audio signals [5]. (2) Temperature monitoring, which seeks to understand the operating status of machinery based on infrared thermometry and thermal imagery [6]. (3) Lubricant analysis to track the health status of bearing by observing the chemical indicators, e.g., the contamination degree, and the spectrum condition of the lubricant. Other factors such as the film resistance or the impact pulse size also vary with the physical conditions of bearing rolling surface [7, 8]. (4) Feature-based analysis to extract temporal and spectral domain-specific characteristics of acoustic or vibration signals. Nowadays, signal acquisition becomes much more convenient and efficient due to the rapid development of the sensing technology. The vast amount of data collected on-site for rotating machinery analysis provides flexible diagnostic approaches based on various forms of information including electrical current [9], sound [10], vibration [11] and even a fusion of multi-sensory signals [12, 13]. As a most important step to realize fault analysis, feature extraction has a significant impact on the system performance and computational complexities. In [14], Wavelet Packet Transformation (WPD) was utilized to decompose the signal into several components by using a uniform frequency bandwidth, based on which a gear faulty feature vector is generated using an entropy indicator. In [15, 16], statistical characteristics were shown to be capable of highlighting differences between physical quantities and thus providing an intuitive diagnostic criterion. In [17], a compressive sensing technique was proposed as an efficient signal reconstruction method to significantly reduce sampling data size while preserving important features and reliably exploiting the similar sparsity structure of the acquired signal. In [18, 19], a non-linear scattering transform method was used to build invariance to geometric transformations and thus enhance the resistance to affine transformations and achieve a high degree of discriminability. The design of manually extracted signal features, however, requires a substantial amount of domain-specific expert knowledge.

Automatic feature representation has proved to be more scalable and empowers the model to deliver discriminant capabilities in a timely manner. Natalia [20] presented a Multi-Layer Perceptron (MLP) classifier for machine faults diagnosis based on certain pre-computed signal characteristics. Al-Tubi *et al*. [21] proposed a hybrid method to diagnose faulty centrifugal pumps, where wavelet transforms were applied to extract features and a Support Vector Machine (SVM) was used to classify denoised signals. A light-weight gradient boosting machine (LGBM) approach was developed in [22] based on the concepts of the Fourier

doi: 10.18178/ijfcc.2024.13.1.612

transform multi-filtering decomposition and joint mutual information maximization. Recently, deep learning methods have made notable achievements in various fields ranging from computer visions to Natural Language Processing (NLP). It demonstrates an overwhelming number of potentials as well when applied to machine fault diagnoses by exploiting various forms of signal features. To tackle the frequently encountered scenario of unbalanced samples, a Generative and Adversarial Network (GAN)-based approach was used to generate supplementary data with the aid of a global optimization scheme [23]. As machine faulty data is usually acquired by on-site sensors, the received signals generally take the format of one-dimensional (1-D) time series with a high temporal resolution. In [24], Short-Time Frequency-Transformation (STFT) was invoked to convert a 1-D vibration signal into a two-dimensional (2-D) frequency-domain representation, which can be viewed as a 2-D red-green-blue (RGB) image following channel-wise stacking and pixel-wise normalization, thus unleashing the impressive power of image classification as manifested by deep learning. In [25], Zhang addressed the issue of how to improve transfer learning when the data in the training and testing stages take different probabilistic distributions by proposing a domain-adaptive Convolutional Neural Network (CNN) model.

In this paper, we propose a novel Multi-Input Multi-Branch (MIMB) architecture based on the concept of ensembled learning, which incorporates a parallel concatenation of both 1D and 2D CNN models working on a comprehensive combination of temporal and spectral characteristics extracted from raw signals. An attention mechanism operating on channel-wise features significantly improves the convergence rate of the training procedure by adaptively adjusting the ratio of feature maps involved in the computation. The learned deep features from each sub-network are combined before input to the final classification layer to learn effectively both global and local temporal dependencies inherent in machine vibration signals. Numerical results on a challenging dataset on machine fault analysis show that the proposed approach achieves the state-of-the-art accuracy and performs noticeably better than the schemes in the literature. Moreover, the proposed architecture can be well generalized to analyze other categories of signals, e.g., biological signals and seismic signals. The rest of the paper is organized as follows. In Section II, a brief description of feature engineering is introduced. Meanwhile, the proposed MIMB architecture incorporating multi-dimensional sub-networks and self-adaptive attention mechanisms is presented. In Section III, numerical results on the Bearing Database of Universität Paderborn (BDUP) [41] are presented and comprehensive comparisons with other methods are also made. The conclusion is drawn in Section IV.

II. Methodology

To enhance the learning capability of the model, domain transformation is frequently used to convert raw time sequences into desirable representations without resorting to the manual selection of signal characteristics. In addition, it enables to the model to fully utilize multiple features learned from various perspectives.

A. Domain Transformation

1) 1D features

The original vibration signal distorted by noises can be divided into segmentations to provide time-domain information. Hence, transformation methods such as the Wavelet Transform (WT) can be introduced to find harmonic components and enhance the overall performance. In particular, WT is capable to process both stationary and transitory signals and is a powerful signal processing technique to extract time-frequency features from 1-D time-domain signals.

Suppose $\psi(t) \in L^2(R)$ with its Fourier Transformation of $\psi(\omega)$ satisfies a weak admissibility condition given by [26],

$$\int_0^{+\infty} \frac{|\psi(\omega)|^2}{|\omega|} \mathrm{d}\omega = \int_{-\infty}^{0} \frac{|\psi(\omega)|^2}{|\omega|} \mathrm{d}\omega = C_\psi < +\infty \tag{1}$$

the wavelet transform of a function $x(t)$ at the scale d and position a is computed by

$$W_f(a,d) = \int_{-\infty}^{+\infty} \frac{x(t)}{\sqrt{d}} \psi^* \left(\frac{t-a}{d}\right) \mathrm{d}t \tag{2}$$

Specifically, the zeroth-order, first-order and second-order scattering coefficient C_0, C_1, C_2 of wavelet scattering transform are given by[18] [19],

$$\begin{cases} C_0 = s \circledast \phi_{J[n]} \\ C_1 = \rho(s \circledast \psi_{\lambda_1}^{(1)}) \times \phi_{J[n]} \\ C_2 = \rho(\rho(s \circledast \psi_{\lambda_1}^{(1)}) \circledast \psi_{\lambda_2}^{(2)}) \circledast \phi_{J[n]} \end{cases} \tag{3}$$

where s denotes the analyzed signal, ⊛ means the periodic convolution, $\phi_J[n]$ denotes a lowpass filter with integer $J > 0$ specifying the averaging scale of the filtering coefficients, $\lambda_1$and λ_2 are frequency indices, and $\rho(t)$ is a non-linearity function. Hence, the coefficients obtained in the WT projecting space can be used as the 1-D feature of fault diagnosis. Furthermore, the concept of multi-scale decomposition makes it feasible for the deep learning network to derive coefficients through multiple channels as shown in Fig. 1, where A presents the approximation signal components and D presents the detailed signal components. For more details on WT, please refer to [26, 27].

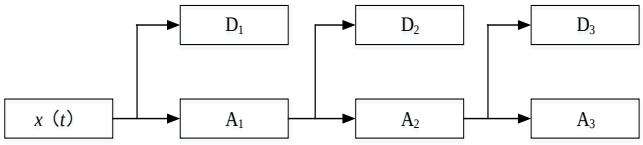

Fig. 1. Wavelet multi-scale decomposition.

2) 2D features

To obtain 2-D frequency-domain features, Short-Time Fourier Transform (STFT) is typically used to partition the signal waveform into segments and perform the FFT over each segment based on the stationary assumption [28]. STFT conducts the framing process of the original signal based on the assumption that each frame is a statistically stationary signal. The STFT of a signal x is given by,

$$STFT_x(t,f) = \int_{-\infty}^{+\infty} x(\tau)h(\tau - t)e^{-j2\pi f\tau}\, d\tau \tag{4}$$

where $h\times(\tau - t)$ is a short time analysis window localized around t and shifting with time. The Mel-frequency spectrogram [29] may be viewed as an improved STFT approach, which uses a Mel filtering bank to obtain a high-resolution image corresponding to the original signal. The transformation can be viewed as a biometric mechanism to simulate the human auditory function, where the Mel frequency is nonlinearly related to the physical frequency as,

$$\begin{cases} M(f) = 2595\times lg(1 + f/700) \\ F(m) = 700\times 10^{(m/2595)} - 700 \end{cases} \tag{5}$$

where f and m represent the actual frequency and Mel frequency respectively. The Mel spectrogram can be obtained through the following four steps: 1) Pre-emphasis, i.e., filter out low-frequency components in the data to make high-frequency characteristics more prominent; 2) Frame, i.e., specify a certain number of sampling points to be analyzed; 3) Obtain frequency components by the Fast Fourier Transform (FFT) operation over each short-term analysis window; 4) Transformation to obtain the Mel spectrogram by passing the results through a bank of Mel filters. For the discrete system, the procedure is usually based on the ordinals of spectrogram lines. The critical step is formulated as follows,

$$B_i(k) = \begin{cases} 0, & k < mc_{i-1} \\ \frac{k - mc_{i-1}}{mc_i - mc_{i-1}}, & mc_{i-1} \le k < mc_i \\ \frac{k - mc_{i+1}}{mc_i - mc_{i+1}}, & mc_i \le k < mc_{i+1} \\ 0, & k \ge mc_{i-1} \end{cases} \tag{6}$$

where $B_i(k)$ denotes the value of the ith filter with frequency k as the independent variable, and mc_i is the Mel central frequency of the ith filter obtained from its corresponding frequency. Therefore, the final Mel spectrogram is given by,

$$L_i = ln(\int_0^{+\infty} B_i(k)\times|X(k)|dt \tag{7}$$

where $X(k)$ denotes the FFT of the signal x. For illustration purposes, we provide an example of the steps to generate the Mel spectrogram as Fig. 2, where the Mel filter banks and the logarithmic operation are introduced between (c) and (d).

In Fig. 2, we show the Mel spectrogram extracted from a signal. Compared with other frequency-domain representations such as the FFT as shown in Fig. 2 (c), the Mel spectrogram is well presented as a 2-D heatmap that manifests rich information on the relationship between frequency and time domains. Following normalization and scaling operations, it can be transformed into a 3-channel image with pixel values ranging from 0 to 255 and hence can be used as the input to an image classification model.

B. Multi-Input Architecture with Attentions

1) Attention mechanism

Inspired by the attention mechanism embedded in human auditory and visual systems, an attention module tries to find out the most conspicuous information on an image or in a signal while ignoring the inconsequential parts [30–33]. In this paper, we propose to use the global attention module aligned with a Long Short-Term Memory (LSTM) model in the signal classification task as shown in Fig. 3.

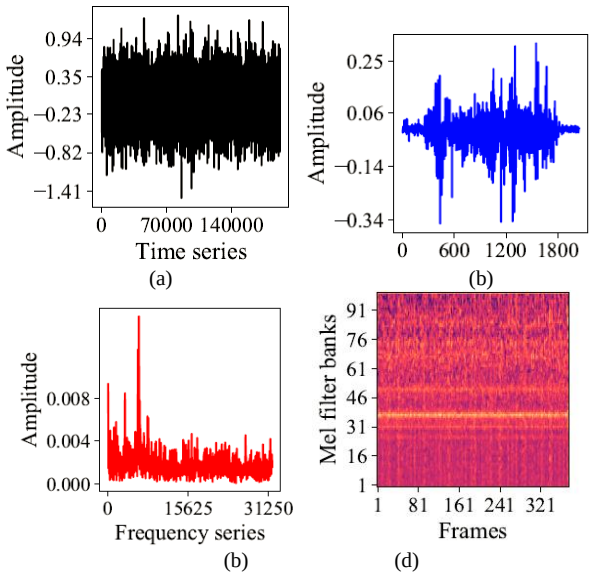

Fig. 2. Mel transformation, (a) Original signal, (b) Frame and add window, (c) The Fourier transform, (d) Mel spectrum.

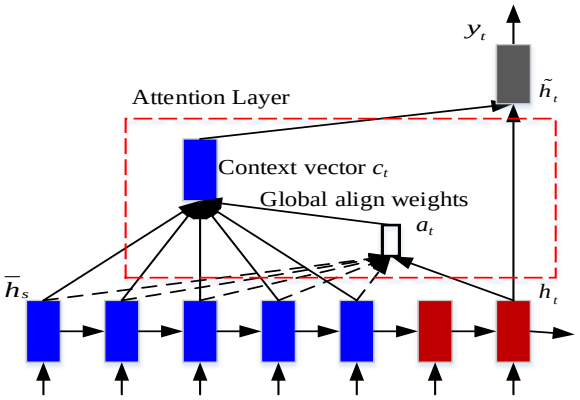

Fig. 3. Global attention model.

where h_s and h_t denote the source-side hidden states and target states of a LSTM layer with the input sequential information. The attention module seeks to derive a context vector c_t to capture relevant hidden state information to predict the current information y_t. First, all the source-side hidden states h_s and the current state h_t are used to generate the attention intensity a_t on the front source-side hidden states, which is formulated by,

$$a_t = \frac{exp\left(score\left(h_t, \bar{h}_s\right)\right)}{\sum_{s'} exp\left(score\left(h_t, \bar{h}_{s'}\right)\right)} \tag{8}$$

where score is referred as a content-based function with three different alternatives,

$$score\left(h_t, \bar{h}_s\right) = \begin{cases} h_t^{\mathrm{T}}\bar{h}_s & dot \\ h_t^{\mathrm{T}}W_a\bar{h}_s & general \\ W_a\left[h_t; \bar{h}_s\right] & concat \end{cases} \tag{9}$$

Second, the context vector c_t is computed as the weighted average over all the source states, while the attention hidden state is produced by combining c_t and h_t as follows,

$$\widetilde{h_t} = \tanh(W_c[c_t; h_t]) \tag{10}$$

Finally, the attention vector is used to produce the predictive distribution as,

$$p(y_t|y_{<t}, x) = \text{softmax}(W_s\widetilde{h_t}) \tag{11}$$

2) Multi-Input Multi-Branch (MIMB) model

In this section, we present the proposed multi-input neural network model that accepts various features obtained from 1-D and 2-D signal transformations. The architecture shown in Fig. 4 contains three network parts: (1) multi-branch sub-network models; (2) attention layer; (3) Long Short-Term Memory (LSTM) layer.

In Fig. 4, the original signal is cut into clips of equal length from which both 1D and 2D features are generated by the feature extraction techniques. Here we propose to use the wavelet coefficients as 1D features; while the STFT and the Mel spectrogram are used as 2D features ranging in the increasing order of computational complexities. For illustration purposes, we present in Fig. 5 the STFT and the Mel spectrograms as 2-D inputs to the proposed model, where the brightness of the pixel represents the power intensity in the frequency domain. Since the attention mechanism has the function of focusing on sensitive features automatically, it is not necessary to put into much effort to search for an optimal network structure. Both 1-D and 2-D sub-network models are composed of sequential convolution layers with reference to the VGG model [34] with max-pooling layers inserted in-between to adjust the dimensions at the input and the output of each module

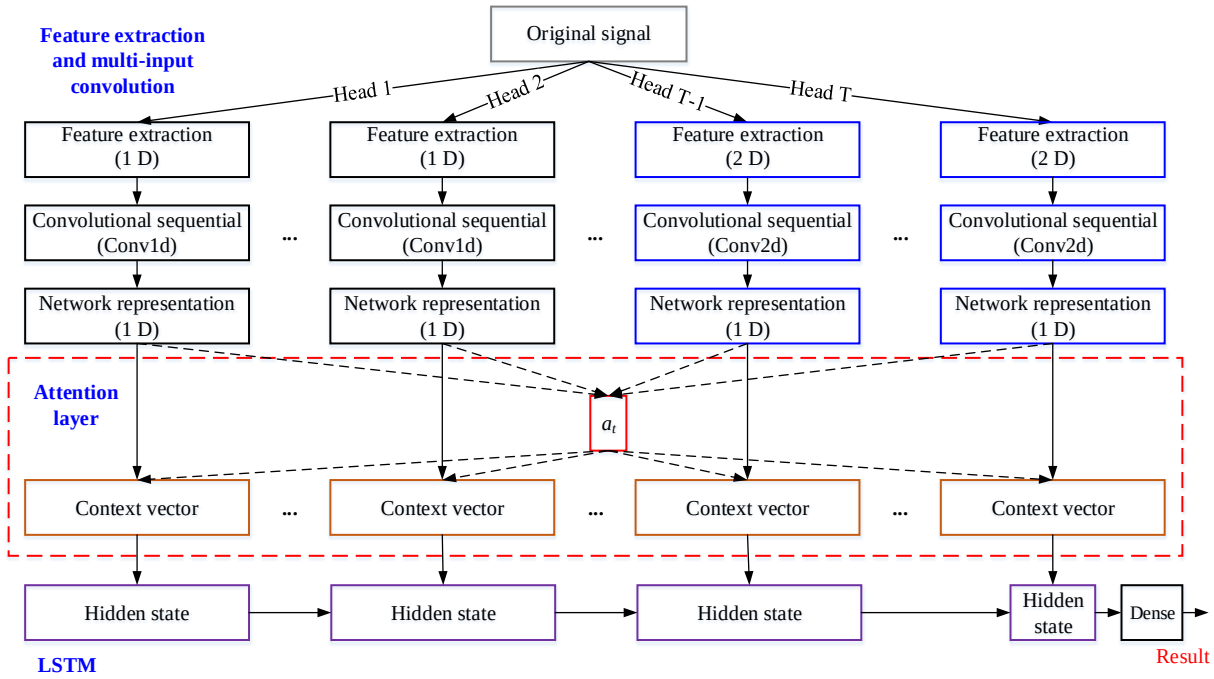

Fig. 4. Diagram of the proposed multi-input multi-branch (MIMB) model with attention mechanism.

In the proposed architecture, the attention operation is conducted on the network outputs and acts as a link between the convolution networks and the LSTM layer. Suppose the output feature from the sub-networks are x_t, the attention vector a_t is formulated by,

$$\begin{cases} s_{t,i} = \frac{e^{-x_{t,i}}}{\sum_{t=1}^{T} e^{-x_{t,i}}} \\ a_t = \frac{1}{I}\sum_{i=1}^{I} s_{t,i} \end{cases} \tag{12}$$

where $x_{t,i}$ denotes the ith member of x_t, $t = 1,2,\cdots,T$, $T \geq 2$ and I is the vector length. Subsequently, the context vectors that are input to the LSTM layer are computed by,

$$C_t = a_t \cdot x_t \tag{13}$$

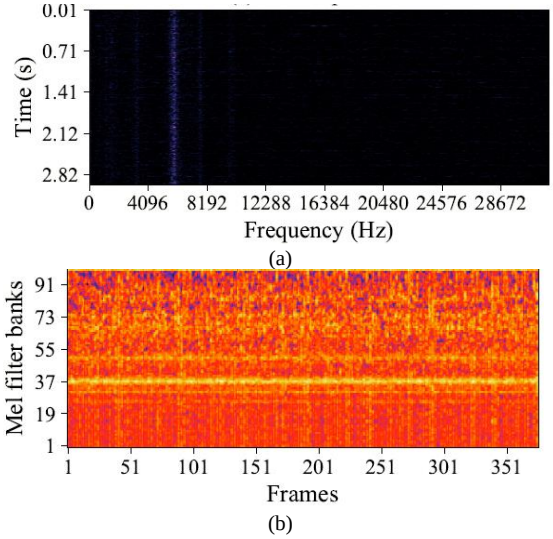

Fig. 5. STFT and Mel spectrograms, (a) STFT spectrum, (b) Mel spectrum.

Fig. 7 shows the flowchart of the attention module in the proposed architecture, where three types of arithmetic operations, i.e., soft-max, vector mean, and scalar multiplication, are represented by x–s, s–a and x– c. Finally, the LSTM layer will operate on the context vectors to obtain the final classification results.

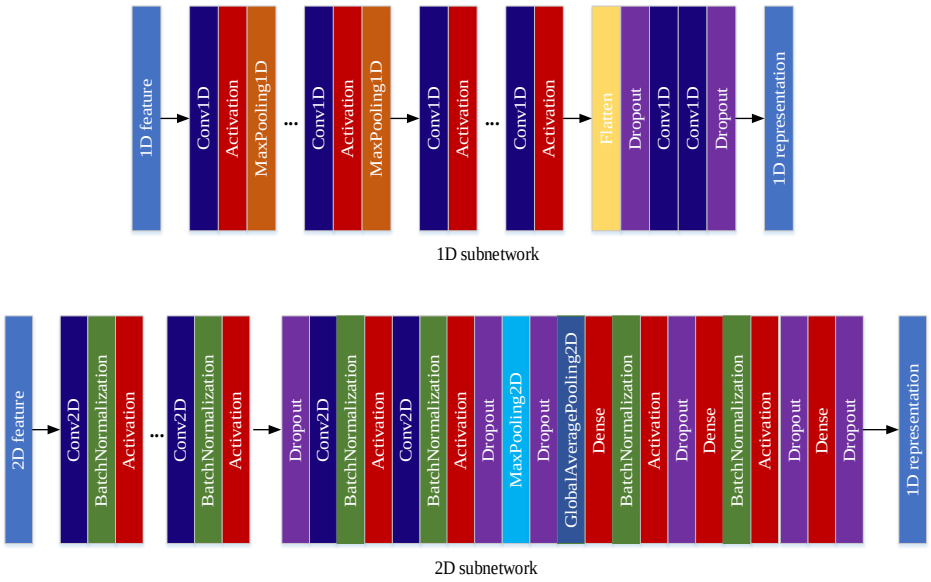

Fig. 6. Sequential structures of 1D and 2D modules in the proposed architecture.

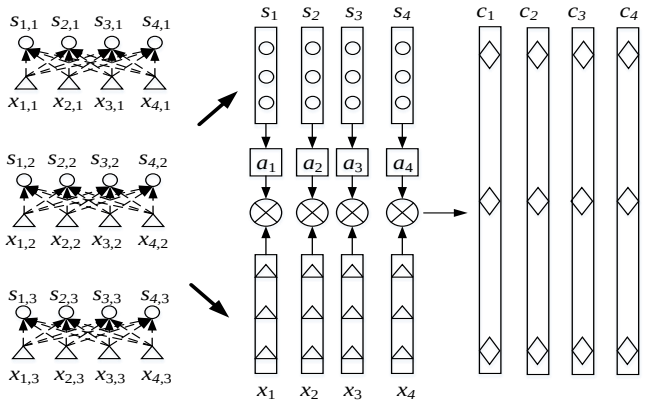

Fig. 7. Flowchart of the attention layer in the context of the proposed architecture.

In summary, the proposed model is essentially a cooperative networking architecture incorporating multiple sub-networks whose inputs are derived from raw signals and represented as multi-dimensional features. As described above, the proposed procedure can be divided into three steps: feature extraction stage to prepare the spectrogram representations of the raw signal; train the proposed model to extract features from deep convolution layers; and pass the features derived from convolution layers into the attention module to facilitate the cooperation of different sub-networks to search for the most discriminant features for the classification.

Note that the Mel spectrogram is provided at the input to the CNN model with inherently embedded time-frequency representations, while the concatenation of domain-specific statistical knowledge and deep representations delivered from the convolution layers enriches signal representations and alleviates overfitting risks, which is of vital importance to improve the generalization capabilities of the proposed architecture. The proposed framework fully leverages the advantages of the meta-learning approach and uses multiple deep-learning models combined with a complementary list of features to gain a noticeable performance improvement over the standalone approach. In Table 1, we summarize the deployment procedure of the proposed model in the scenario of machine fault diagnosis, which is divided into the training and the validation phase as well as an on-site deployment stage.

Table 1. The proposed fault diagnosis algorithm

	Training Stage
1	Obtain original signals from the sensors and cut them into samples of the same length.
2	Compute 1D and 2D features using feature extraction techniques for each sub-network.
3	Design the sequential convolutional neural networks as the feature extraction backbones.
4	Introduce the attention mechanism to fuse the network representations.
5	LSTM works on the fused deep features to further derive temporal patterns.
	Validation Phase
1	Perform the same operations of step 1~ step 2 in the training stage to obtain multiple features.
2	Use the trained attention-based multi-input neural network to classify testing samples.
3	Adjust parameters such as network learning rate and other hyper-parameters and validate the model.
	Model Deployment
1	Conduct the same operation of signal acquisition and pre-processing as the training stage.
2	Compute 1D or 2D features using the existing feature extraction techniques for corresponding sub-networks.
3	Select the best model to classify the online multi-domain features to realize fault diagnosis.

III. Numerical Results

A. Experimental Condition

We conducted the experiments on the bearing working database of Universität Paderborn, Germany and the structure of the test rig is shown in Fig. 8. There are primarily five modules: (1) a permanent magnet synchronous motor; (2) torque-measurement device, which delivers the torque value of the power transmission shaft; (3) rolling bearing test module to sample experimental data from ball bearings with different types of damage. (4) flywheel to simulate the inertia

driven equipment; (5) load motor to provide a constant radial load, which can be continuously adjusted up to 10 kN. In a typical experimental setting, the acceleration signal, which is viewed as the object of investigation, is measured at the top end of the rolling bearing module using a piezoelectric accelerometer (Model No. 336C04 @ PCB Piezotronics.) and a charge amplifier (Type 5015A @ Kistler Group) with a low-pass filter at 30 kHz. Meanwhile, the signal is digitalized and saved with the sampling rate of 64 kHz.

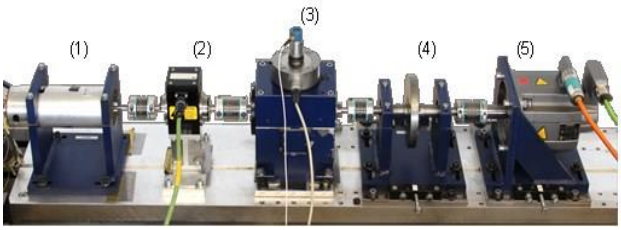

Fig. 8. The test rig [35].

In Fig. 8, bearings mounted in the test module can be replaced by the damaged counterparts to artificially introduce machine faults. The machine works under complex and volatile conditions leading to the various operating parameters measured on the components in terms of rotational speeds, load torques, and radial forces. To ensure a fair comparison with the literature, it is necessary to execute the setup as shown in Table 2 while taking into considerations as well multiple operating conditions needed for the experiments.

Table 2. Setting of multiple conditions [35]

No.	Rotational Speed	Load Torque	Radial Force	Setting Name
0	1500	0.7	1000	N15 M07 F10
1	900	0.7	1000	N09 M07 F10
2	1500	0.1	1000	N09 M01 F10
3	1500	0.7	400	N15 M07 F04

In this paper, a total number of 14 bearings with real damages sampled from accelerated lifetime tests are selected as the faulty experimental objects. These bearings are categorized by the faulty processing methods as shown in Table 3.

TABLE 3. Setting of multiple conditions

Bearing	Dam	BE	Comb	Arra	DE	CD
KA0	F:P	OR	S	No	1	SP
KA15	PD:I	OR	S	No	1	SP
KA16	F:P	OR	R	rand	2	SP
KA22	F:P	OR	S	No	1	SP
KA30	PD:I	OR	R	rand	1	Ds
KB23	F:P	OR + IR	M	rand	2	SP
KB24	F:P	OR + IR	M	No	3	Ds
KB27	PD:I	OR + IR	M	rand	1	Ds
KI04	F:P	IR	M	No	1	SP
KI14	F:P	IR	M	No	1	SP
KI16	F:P	IR	S	No	3	SP
KI17	F:P	IR	R	rand	1	SP
KI18	F:P	IR	S	No	2	SP
KI21	F:P	IR	S	No	1	SP

The descriptions of the acronyms used in the experimental settings is presented in Table 4, where KA04 and KA22 have the different damage geometries with width 3mm and width 1mm respectively, while KI04 and KI14 are similarly distinguished by length 2mm and length 1mm. Since abnormal detection is the first step for the fault diagnosis, the normal samples are needed for the diagnostic model. The normal data is labeled by bearing codes K001–K006, which is sampled during six run-in periods.

Table 4. Setting of multiple conditions

Acronyms	Descriptions
F:P	fatigue: pitting, which stands for the mode and symptom of the Dam (damage)
PD: I	plastic deform: indentation, which is another mode of the damage
OR	outer ring of the BE (bearing element)
IR	inner ring of the BE
S	single damage, which means one single component is affected by single damage
R	repetitive damage, i.e., damages that are repeated at several places on the same bearing component
M	different damages occur or identical damage symptoms occur on different bearing components
Arra	arrangement of the repetitive and multiple damages, which characterizes the arrangement of the damage symptoms on each component.
No	no repetition. Namely, the damage occurs only once.
rand	random distribution of local damage symptoms
DE	damage extent to represent the extent of bearing damage
SP	single point. The characteristic of damage (CD) is characterized by a small extend at a localized position.
Ds	Distributed damages characterized by generalized roughness

In Fig. 9, we present the samples of the time-domain bearing waveforms belonging to a total number of 15 categories as covered in this paper. Note that signals across several categories are hardly visually discernible, which poses a significant challenge to the task of achieving a high classification accuracy over this dataset with a diverse and complex background. Hence it is necessary to employ the effective signal processing techniques to enhance the representational capacity of the extracted signal features.

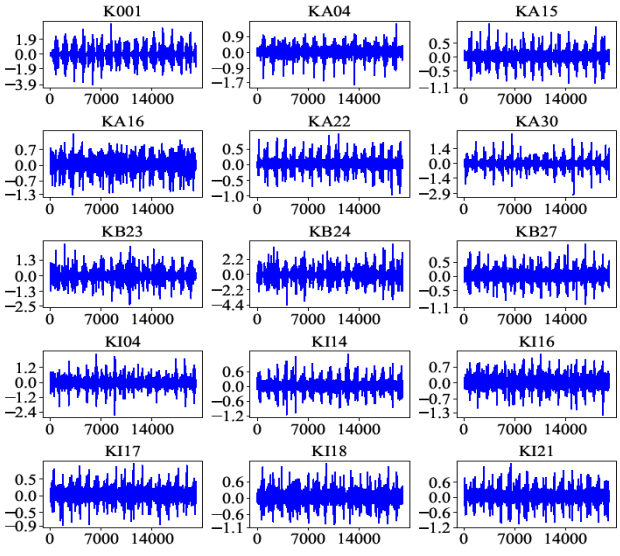

Fig. 9. Bearing time-domain waveforms.

B. Experimental Settings

In training a deep learning model, it is of importance to maintain a balanced dataset by eliminating an excessively high proportion of samples belonging to a specific class, thus enabling the model to respond equally to data points distributed across the domain. The balanced settings of the training samples, the validation samples, and the test samples are shown in Table 5.

Table 5. Category labels of samples

Faulty Type	Train Num	Validation Num	Test Num	Label
Normal	234	54	96	0
KA04	240	60	100	1
KA15	240	60	100	2
KA16	240	60	100	3
KA22	240	60	100	4
KA30	240	60	100	5
KB23	240	60	100	6
KB24	240	60	100	7
KB27	240	60	100	8
KI04	240	60	100	9
KI14	240	60	100	10
KI16	240	60	100	11
KI17	240	60	100	12
KI18	240	60	100	13
KI21	240	60	100	14

For the purpose of fair comparisons with other methods, we also ensure strict independence between the training process and the validation process. The validation set is designated to be independent of the model training process and used only as a performance indicator. In order to objectively evaluate the performance of the model on the data that it has not seen, we ensure that the validation set and the training set are strictly non-overlapping with each other. As the training subset contains both normal signals and faulty signals with various degrees of damages on the bearings, we expect that a well-trained model will possess the capabilities of fault detection and the classification of damages as well.

To visualize the scattering effects on a machine vibration signal, we include in Fig. 10 the representations of different orders of wavelet scattering. The wavelet filtering effectively converts the raw waveform distorted by high-frequency noises into a 3-channel RGB image, which can be used at the input of 2D CNN models.

In the proposed model, each convolution layer is followed by a ReLu function to speed up the convergence of the training process, a batch normalization module to perform the scaling of each layer's output based on the specified batch size, and a pooling operation to obtain position invariance over local regions, as well as a dropout operation to reduce dependencies between adjacent layers. The parameters of each layer in the training process are adjusted based on the backward-propagation (BP) algorithm targeted at minimizing a cross-entropy (CE)-based loss function. A total of 150 epochs is adopted to train the model. Moreover, dropout is introduced into fully connected layers to speed up training and prevent over-fitting. The probability of dropped neurons changes randomly at each epoch, which changes the architecture and reduces over-fitting compared with the training process without dropout.

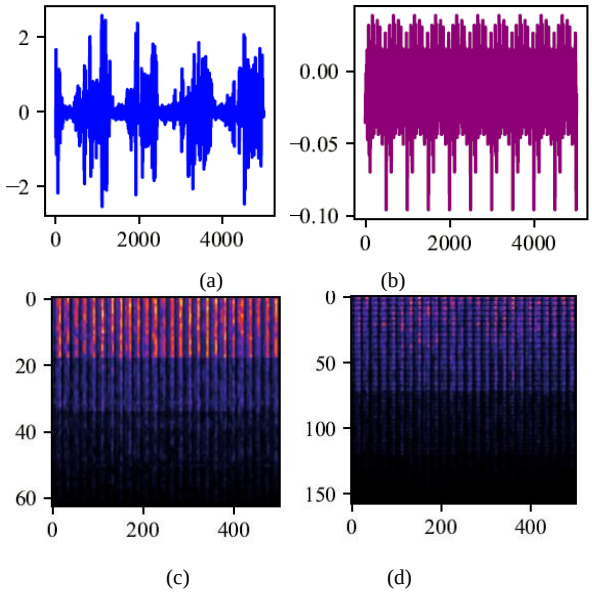

Fig. 10. Graphic representations of wavelet scattering; (a) Raw data, (b) Zeroth-order scattering, (c) First-order scattering, (d) Second-order scattering

C. Numerical Results

To further reduce overfitting risks and improve the generalization capability of the model, we perform extensive data augmentations on raw signals including time-shifting, Gaussian noise injection, pitch changing, speed tuning and dynamic amplitude before forwarding the augmented waveforms to generate the frequency-domain Mel spectrograms. Effective data augmentation increases not only the number but also the diversity of training samples. In the experiments, the number of the Mel frequencies is set to 40 to provide a high resolution over low-frequency region. In Table 6, we present a comprehensive comparison between the proposed model and the representative approaches in the literature in terms of the metrics commonly used to evaluate the performance of classification models. For instance, accuracy is defined as the percentage of the total number of samples that correct results are predicted, while the recall rate, e.g., sensitivity, denotes the ratio of the number of positive predictions that are correctly predicted to the total number of positive examples. Other metrics include F1-score and the argument memory (AM) in terms of kilo-bytes (KB) to denote the memory occupation of various models.

The first approach in Table 6 uses the conventional 1D CNN with raw signal waveforms as the model input. Other approaches featured with 'Mel' use various 2D CNN models as the feature extractors based on the Mel Spectrograms of raw signals, where we have employed classical sequential models, ResNet models with residual connections between convolution layers, and the EfficientNet models proposed recently by Google Deepmind [38]. Note that the model denoted by 'Scattering' [18] resorts to wavelet scattering transform to build invariances to geometric transformations while keeping a high discriminability. Hence, the scattering coefficients are used as the input to the VGG model and the performance is also fine-tuned based on the experimental settings. Table 6 shows that the proposed model, which effectively fuses complementary deep features from both 1D and 2D sub-networks and uses the attention mechanism to capture spatial-temporal information embedded in audio

signals achieves noticeable performance gains over other approaches on the validation set. Moreover, the proposed model yields the state-of-the-art accuracy of 99.87% and the F1−Score of 98.96% at the cost of an impressively moderate memory size to host the network, which are the best results to our knowledge.

Table 6. The results of the contrast experiments

Methods	Accuracy	Recall rate	F1−score	Argument memory
1D CNN + Raw [36]	98.20%	94.79%	93.81%	23645
2D CNN + Mel [37]	98.53%	100.0%	91.43%	2092
EfficientNet + Mel [38]	99.40%	97.26%	97.94%	149451
ResNet + Mel [39]	99.13%	100.0%	97.96%	258225
VGG + Scattering [40]	99.59%	100.0%	100.0%	174366
Proposed	99.87%	98.96%	98.96%	53076

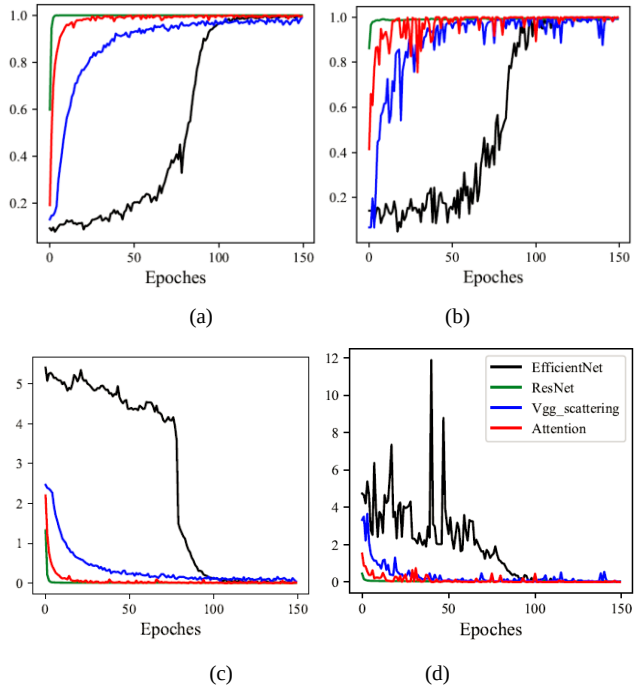

Fig. 11. Accuracy and loss curves of various models in the training process; (a) Accuracy curves of training set, (b) Accuracy curves of validation set, (c) loss curves of training set, (d) loss curves of validation set

In Fig. 11, we present the accuracy curves and loss curves of various models over the training and the validation sets as well. Both curves serve a valuable indicator of the generalization capability of deep-learning models. It is expected that the accuracy curves tend to 100.0% as the number of training iterations increases, while the loss curves are expected to converge to zero and manifest a significant amount of stability. Fig. 11(a) shows that the accuracy curve of the proposed architecture increases closely to 100% with approximately 20 training epochs and does not show volatility for the following epochs, thus achieving better stability capability over other models. The curves in Fig. 11(b) over the validation sets show similar trends, which indicates that the proposed model can be well generalized to unseen samples. In Fig. 11(c) and Fig. 11(d), the proposed model behaves consistently with respect to loss curves and converge very quickly to a minimal loss in the training process. It is noted that the EfficientNet model exhibits a large degree of volatility when evaluated on the validation data.

The performance of the proposed model can be visually represented through a confusion matrix evaluated on the test data, which can be divided into 15-class data as shown in Table 4. The confusion matrix as shown in Fig. 12 is a specific table layout that allows a direct visualization of the performance in each class. It reports the errors and confusions among different classes by calculating the correct and incorrect classification of the test samples for each class.

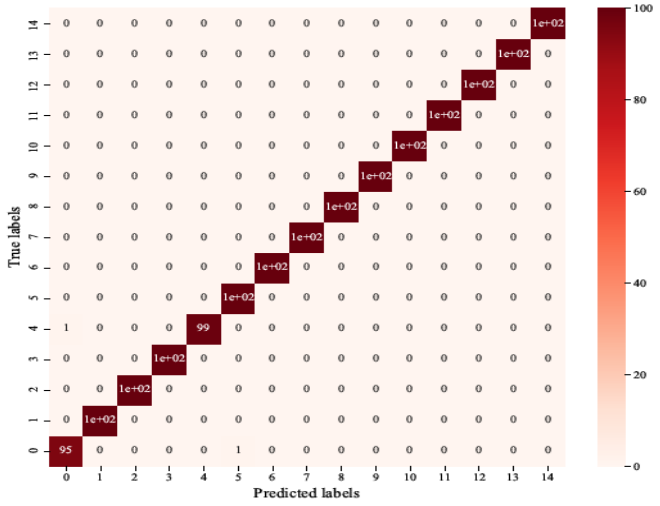

Fig. 12. Confusion matrix of the proposed algorithm on the test dataset.

The horizontal axis in Fig. 12 represents the predicted samples and the vertical axis represents the true samples. The probabilities of correctly classified results are recorded on the diagonal, while those of incorrect predictions are scattered through the matrix. Among a total number of 96 normal signals, there is one signal that is mis-classified from the zeroth class to the 5th class. Besides, a sample belonging to the 4th class is incorrectly predicted as a normal sample of the zeroth class. A visual inspection of machine fault signals shows that the differences between certain classes are sufficiently subtle even for humans to recognize. For all the other classes of faulty signals, Fig. 12 shows that the proposed model achieves an impressive performance of 99.87% accuracy and nearly reaches an upper bound on this typical large-scale machine vibration dataset even with complex and diverse audio background. The performance gain is accredited to the design of multi-branch sub-networks to

extract complementary deep features from various domains of signal waveforms. The sub-networks have different depths and widths, which enable them to learn discriminant semantic information as compared with a single feature extractor. The deep fusion module, which consists of the attention mechanism embedded in the LSTM layer, can be viewed as a collection of non-linear high-dimensional projections onto the fully connected layers prior to the final classification. The attention module integrates deep features adaptively by capturing the global pattern embedded in the fused features, while the subsequent LSTM layers refines local temporal relations to further enhance the performance. Hence, the proposed architecture manifests an excellent capability to extract multi-scaled temporal dependencies embedded in the input time sequences.

IV. Conclusion

In this paper, we propose a multi-input multi-branch (MIMB) deep-learning architecture incorporating a complimentary feature fusion mechanism and an attention module for the task of machine fault diagnosis. The multi-branch sub-networks serve as deep feature extractors and enable the proposed model to learn effectively both global and local discriminant characteristics of the wavelet-denoised machine vibration signals. Extensive data augmentations are applied to alleviate overfitting and ensure the diversities of training samples. The performance is evaluated on a typical bearing dataset of Universität Paderborn. Compared with other standalone schemes in the literature, the proposed approach achieves a noticeable performance gain and obtains the state-of-the-art accuracy of 99.87% on the independent test data. Moreover, the proposed model shows rapid convergence in the training process and is also feasible for practical implementations due to its moderate memory requirement.

For the future work, we believe that the seamless integration of deep-learning algorithms into pre-emptive maintenance will bring unprecedented industrial opportunities. The proposed method can be generalized to analyze various categories of signals, such as heartbeat sound signals, environmental sounds, and gyrometer sensor data. We will further investigate the efficient fusion of complementary features and the joint optimization strategy of various sub-network models.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Lianghui Zou completed the deep-learning codes for this paper. Haoxin Qin assisted in data processing tasks and numerical experiments. Qidong Lu assisted in the literature survey. Zhiliang Qin designed the experimental framework. All authors approved the final version.

Acknowledgement

This paper was supported by the Shandong Provincial Key Science & Technology Innovative Engineering Foundation (No. 2020CXGC010112).

References

[1] T. Escobet, A. Bregon, B. Pulido, and V. Puig, *Fault Diagnosis of Dynamic Systems*, Springer, 2019.

[2] D. Zhou *et al*., "Fault diagnosis techniques for dynamic systems," *Acta Automatica Sinica*, vol. 35, no. 6, pp. 748–758, 2009.

[3] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 2, pp. 1539–1548, 2017.

[4] J. Deutsch and D. He, "Using deep learning-based approach to predict remaining useful life of rotating components," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 1, pp. 11–20, 2017.

[5] S. R. Saufi, Z. A. B. Ahmad, M. S. Leong, and M. H. Lim, "Low-speed bearing fault diagnosis based on ArSSAE model using acoustic emission and vibration signals," *IEEE Access*, vol. 7, pp. 46885–46897, 2019.

[6] M. Xia, G. Han, Y. Zhang, J. Wan et al., "Intelligent fault diagnosis of rotor-bearing system under varying working conditions with modified transfer CNN and thermal images," *IEEE Transactions on Industrial Informatics*, pp. 3488-3496, Jun. 2020.

[7] S. Ramezani and A. Memariani, "A fuzzy rule based system for fault diagnosis using oil analysis results," *International Journal of Industrial Engineering*, vol. 22, no. 2, 2011.

[8] M. Lopez-Ramirez, R. J. Romero-Troncoso, D. Moriningo-Sotelo, O. Duque-Perez, D. Camarena-Martinez, and A. Garcia-Perez, "Discriminating the lubrication condition from the rotor bearing fault in induction motors using Margenau-Hill frequency distribution and artificial neural networks," *Industrial Lubrication and Tribology*, vol. 69, pp. 970-979, 2017.

[9] A. S. Barcelos and A. J. M. Cardoso, "Current-based bearing fault diagnosis using deep learning algorithms," *Energies*, vol. 14, no. 9, p. 2509, 2021.

[10] K. Shibata, A. Takahashi, and T. Shirai, "Fault diagnosis of rotating machinery through visualization of sound signals," *Mechanical Systems and Signal Processing*, vol. 14, no. 2, pp. 229–241, 2000.

[11] C. Mongia, D. Goyal, and S. Sehgal, "Vibration response-based condition monitoring and fault diagnosis ofrotary machinery," *Materials Today: Proceedings*, vol. 50, pp. 679–683, 2022.

[12] M. Azamfar, J. Singh, I. Bravo-Imaz, and J. Lee, "Multisensor data fusion for gearbox fault diagnosis using 2d convolutional neural network and motor current signature analysis," *Mechanical Systems and Signal Processing*, vol. 144, 2020.

[13] Y. Pang, D. Yang, R. Teng, B. Zhou, and C. Xu, "A deep learning based multiple signals fusion architecture for power system fault diagnosis," *Sustainable Energy, Grids and Networks*, vol. 30, p. 100660, 2022.

[14] A. Afia, C. Rahmoune, D. Benazzouz, B. Merainani, and S. Fedala, "New gear fault diagnosis method based on MODWPT and neural network for feature extraction and classification," *Journal of Testing and Evaluation*, vol. 49, no. 2, 2019.

[15] R. S. Figliola and D. E. Beasley, *Theory and design for mechanical measurements*. John Wiley & Sons, 2020.

[16] J. A. Grajales, H. F. Quintero, C. A. Romero, and E. Henao, "Engine diagnosis based on vibration analysis using different fuel blends," *Advances in Condition Monitoring of Machinery in Non-Stationary Operations*, Springer, 2018, pp. 267–274.

[17] K. F. Brethee, G. R. Ibrahim, and R. A. Mohammed, "Using envelope analysis and compressive sensing method for intelligent fault diagnosis of ball bearing," *Advances in Science Technology and Engineering Systems Journal*, vol. 5, no. 5, pp. 370–375, 2020.

[18] M. Andreux *et al*., "Kymatio: Scattering transforms in python." *Journal of Machine Learning Research*, vol. 21, no. 60, pp. 1–6, 2020.

[19] S. Ma, B. Cheng, Z. Shang, and G. Liu, "Scattering transform and LSPTSVM based fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 104, pp. 155–170, 2018.

[20] N. F. E. Sepúlveda and J. K. Sinha, "Blind application of developed smart vibration-based machine learning (SVML) model for machine faults diagnosis to different machine conditions," *Journal of Vibration Engineering & Technologies*, pp. 1–10, 2020.

[21] M. A. S. ALTobi, G. Bevan, P. Wallace, D. Harrison, and K. Ramachandran, "Fault diagnosis of a centrifugal pump using MLP-GABP and SVM with CWT," *Engineering Science and Technology*, vol. 22, no. 3, pp. 854–861, 2019.

[22] C. Zhang, L. Kong, Q. Xu, K. Zhou, and H. Pan, "Fault diagnosis of key components in the rotating machinery based on Fourier Transform multi-filter decomposition and optimized LightGBM," *Measurement Science and Technology*, vol. 32, no. 1, 2020.

[23] F. Zhou, S. Yang, H. Fujita, D. Chen, and C. Wen, "Deep learning fault diagnosis method based on global optimization GAN for unbalanced data," *Knowledge-Based Systems*, vol. 187, p. 104837, 2020.
[24] H. Tao, P. Wang, Y. Chen, V. Stojanovic, and H. Yang, "An unsupervised fault diagnosis method for rolling bearing using STFT and generative neural networks," *Journal of the Franklin Institute*, 2020.
[25] B. Zhang, W. Li, X.-L. Li, and S.-K. Ng, "Intelligent fault diagnosis under varying working conditions based on domain adaptive convolutional neural networks," *IEEE Access*, vol. 6, pp. 66367–66384, 2018.
[26] A. Bouzida, O. Touhami, R. Ibtiouen, A. Belouchrani, M. Fadel, and A. Rezzoug, "Fault diagnosis in industrial induction machines through discrete wavelet transform," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 9, pp. 4385–4395, 2010.
[27] P. Liang, C. Deng, J. Wu, and Z. Yang, "Intelligent fault diagnosis of rotating machinery via wavelet transform, generative adversarial nets and convolutional neural network," *Measurement*, 2020.
[28] D. Mustafa, Z. Yicheng, G. Minjie, H. Jonas, and F. Jürgen, "Motor current based misalignment diagnosis on linear axes with Short-Time Fourier Transform (STFT)," *Procedia CIRP*, vol. 106, pp. 239–243, 2022.
[29] Y. Atmani, S. Rechak, A. Mesloub, and L. Hemmouche, "Enhancement in bearing fault classification parameters using Gaussian mixture models and Mel frequency cepstral coefficients features," *Archives of Acoustics*, vol. 45, no. 2, pp. 283–295, 2020.
[30] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*,Volume 2, pp. 2204–2212, December 2014.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.
[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.
[33] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.
[34] A. Nasiri, A. Taheri-Garavand, M. Omid, and G. M. Carlomagno, "Intelligent fault diagnosis of cooling radiator based on deep learning analysis of infrared thermal images," *Applied Thermal Engineering*, vol. 163, p. 114410, 2019.
[35] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," *Proceedings of the European conference of the Prognostics and Health Management Society*. Citeseer, 2016, pp. 05–08.
[36] I. H. Ozcan, O. C. Devecioglu, T. Ince, L. Eren and M. Askar. "Enhanced bearing fault detection using multichannel, multilevel 1D CNN classifier," *Electrical Engineering*, vol. 104, pp. 435-447, 2022.
[37] M. T. Pham, J. M. Kim, C. H. Kim, and C. H. Kim. "2D CNN-Based Multi-Output Diagnosis for Compound Bearing Faults under Variable Rotational Speeds," *Machines*, vol. 9, pp. 199, 2021.
[38] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *International Conference on Machine Learning*, pp. 6105–6114, 2019.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[40] T .Bäßler, "Augmented Mel-spectrogram VGG-16 model for axial and radial load classification at wire-race bearings," *SSRN Electronic Journal*, 2022.
[41] https://mb.uni-paderborn.de/en/kat/main-research/datacenter/bearing-datacenter/data-sets-and-download