Smart Structures and Systems, *Vol. 22, No. 2 (2018) 175-183*
DOI: https://doi.org/10.12989/sss.2018.22.2.175

Wavelet-like convolutional neural network structure for time-series data classification

Seungtae Park[1], Haedong Jeong[2], Hyungcheol Min[3], Hojin Lee[1] and Seungchul Lee[*1]

[1]*Department of Mechanical Engineering, Pohang University of Science and Technology, Pohang, Republic of Korea*
[2]*Department of System Design and Control, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea*
[3]*Korea Electric Power Corporation Research Institute, Daejeon, Republic of Korea*

(Received May 8, 2017, Revised March 15, 2018, Accepted March 19, 2018)

Abstract. Time-series data often contain one of the most valuable pieces of information in many fields including manufacturing. Because time-series data are relatively cheap to acquire, they (e.g., vibration signals) have become a crucial part of big data even in manufacturing shop floors. Recently, deep-learning models have shown state-of-art performance for analyzing big data because of their sophisticated structures and considerable computational power. Traditional models for a machinery-monitoring system have highly relied on features selected by human experts. In addition, the representational power of such models fails as the data distribution becomes complicated. On the other hand, deep-learning models automatically select highly abstracted features during the optimization process, and their representational power is better than that of traditional neural network models. However, the applicability of deep-learning models to the field of prognostics and health management (PHM) has not been well investigated yet. This study integrates the "residual fitting" mechanism inherently embedded in the wavelet transform into the convolutional neural network deep-learning structure. As a result, the architecture combines a signal smoother and classification procedures into a single model. Validation results from rotor vibration data demonstrate that our model outperforms all other off-the-shelf feature-based models.

Keywords: convolutional neural networks; machine learning; deep learning; time-series analysis

1. Introduction

Discrete time signals are a typical type of data acquired during the diagnosis of machine health conditions in the field of prognostics and health management (PHM). To analyze a given signal, most data-driven approaches aim to extract important features from a finite length of time windows of the given signal and then carry out advanced data analytics using techniques such as machine learning or statistical data mining (Zhou *et al*. 2013). From a machine-learning perspective, a given signal is categorized into supervised classification problems for a given labeled time-series dataset. The common choices for features in the time domain are based on certain orders of the signal moment and those in the frequency spectrum are based on decomposition by harmonic frequency components. In many cases, key pieces of information are given in the form of time-series data.

However, in many cases, signals are not obtained by the direct measurements of the desired physical quantities of a target system. Signals are typically corrupted by various noise channels because of environmental factors, innate limitations of measurement devices, etc. Furthermore, feature extraction from time-series data is not directly related to the classification algorithm used. For example, wavelet coefficients may be used as features for a given set of time signals. However, the types of mother wavelets are often chosen based on domain knowledge, which is acquired by applying several machine-learning methods to the wavelet coefficients for classification, even in many PHM applications. Here, the mother wavelets and the classification algorithms are selected separately. Therefore, key pieces of information contained in time-series data may not be well extracted, and thus, the classification algorithm may not be fully compatible with the time-series data for the estimation and prediction of machine health conditions in PHM problems.

Smoothing is a typical preprocessing step performed on noise-corrupted time-series data by applying low-pass filters. Smoothing should be performed carefully so as to not remove any key features of signals in high-frequency bands. Inappropriate smoothing procedures might wash away key pieces of information in signals. For example, the application of a low-pass filter might cause the failure of the characterization of signal features such as bumps or spikes (Donoho 1993). Furthermore, even if signals are successfully reconstructed from noise, the feature-searching step performed by human experts can introduce serious biases into the analysis. Therefore, we believe that features from time signals should be extracted in conjunction with a classification model to ultimately provide better classification performance.

Among the many signal-processing methodologies available for handling time-series data, we focus on the discrete wavelet transform (DWT). In this paper, the term

*Corresponding author, Professor
E-mail: seunglee@postech.ac.kr

ISSN: 1738-1584 (Print), 1738-1991 (Online)

"wavelet transform" is used to denote the DWT. Wavelet transform shows better signal smoothing performance than other traditional methods such as the spline method or the low-pass filter in Fourier analysis (Donoho 1994) mainly because of its unique structure called "residual fitting". Also its applicability is proven through many discrete time signal classification problems (Borghetti *et al* 2008, Rafiee *et al* 2010, Zhang *et al* 2004, Jung *et al* 2014). The scheme is illustrated in Fig. 1. First, it extracts high-frequency components using the mother wavelet and then removes the high-frequency components from the original signal. Then, it extracts lower-frequency components from the residual signal and proceeds to remove them. Although such a wavelet transform method is widely used in practice, the aforementioned classification performance highly depends on the shape of the mother wavelet, which is predefined.

Therefore, to overcome such limitations, we propose to apply such a residual fitting scheme to the convolutional neural network (CNN) architecture (Fig. 3). The CNN is one of the most famous deep neural network models and is the greatest contributing factor to the success of computer vision (Matsugu *et al*. 2003). A neural network or deep learning does not require manually engineered well-selected features. It contains a complicated learning mechanism that facilitates autonomous feature extraction and classification. In the wavelet-like CNN, the mother wavelet can be considered as the convolutional filter that is to be optimized to minimize the classification cost in the framework of the deep-learning model. Therefore, more adequate mother wavelets (or kernels) can be expected, resulting in better classification performance. In the PHM field, studies have been conducted to utilize generic CNN models (Janssens *et al.* 2016, Chen *et al*. 2015, Jeong *et al.* 2016). However, these studies are mostly limited to image data and do not carefully account for the unique structures of CNN models representing time-series data gathered for machine health condition monitoring.

The proposed wavelet-like CNN model has two advantages. First, it directly approximates mother wavelets or kernels for signals, whereas the wavelet transform uses a rather predefined form of the mother wavelet regardless of the characteristics of signals and classification problems. Second, it automatically deciphers the abstracted features via a deep-learning model.

The rest of the paper is organized as follows. In Section 2, we briefly discuss the residual fitting of the wavelet transform and CNN models. This section provides a unified view of the CNN and the wavelet transform. Then, in Section 3, we expand the concept described in the previous section to the wavelet-like CNN in detail for time-series data classification. In Section 4, we present a validation of our proposed approach, conducted using vibration signals for rotating machinery.

2. Residual fitting

Residuals are defined as the differences between the observation data and the fit to the observation data. Residual fitting is the scheme for fitting the residuals step by step. In general, the wavelet transform carries out residual fitting by gradually fitting residuals with two main components: filters and convolution operators. At first, corrupted signals are convoluted with a large sized filter to extract low frequency component. Since the large size filter corresponds to a long -time resolution, we denote it by long time module. Then the extracted low frequency component is removed from the corrupted signal and the same procedure is conducted with a smaller sized filter.

The general structure of the wavelet transform is depicted in Fig. 1, where the filters are predefined mother wavelets and the operator is a linear convolutional operator. The module structure of the wavelet transform is depicted in Fig. 2. Our study is to improve the wavelet transform by specifying each time module in more adaptive way.

The configuration can be improved by assuming that the nonlinear convolutional operator H and the filter K_i are free to be optimized according to the problem setting. Under this assumption, the problem is categorized into CNN models using the residual fitting scheme. Because the proposed CNN model follows the same scheme as the wavelet transform, we call it the "wavelet-like CNN" model.

3. Structure of wavelet-like CNN model

In this section, we address the technical problems encountered when realizing our concepts and then propose a new deep neural network model based on a basic CNN structure.

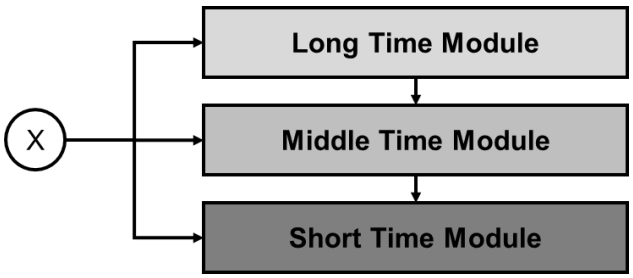


Fig. 1 Residual fitting in wavelet transform

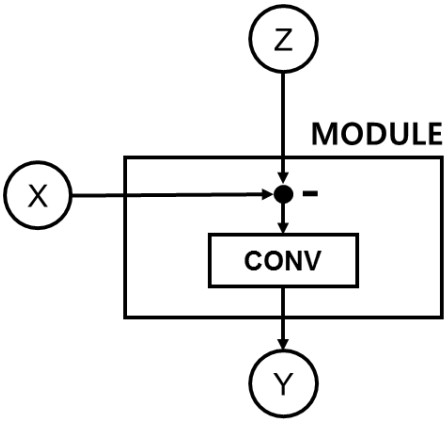


Fig. 2 Module structure of wavelet transform

3.1 CNN model

CNN models are known as one of the biologically inspired models and have been widely used for image pattern recognition problems such as handwritten digit recognition and face recognition (Matsugu *et al*. 2003). In image recognition, a CNN model consists of multiple layers of small parameters, and the model collects information about the parameters to obtain better representations of the original image (Korekado *et al.* 2003). A basic CNN architecture includes pairs of convolutional subsampling layers (LeCun *et al.* 1989). The last subsampling layers are fully connected, and the output vector classifies the input using max pooling between the overall values of the activation function. This hierarchical organization helps to autonomously extract appropriate features in image classification tasks without prior domain knowledge. When dealing with the variability of two-dimensional (2D) shapes, CNN-oriented models typically outperform other models (LeCun *et al.* 1989) because the convolutional operator effectively carries spatial information.

The application of CNN-oriented models to one-dimensional (1D) time-series data is not justified. However, we believe that such an application can be justified if a sufficient sampling frequency rate is used. A sufficient sampling frequency rate retains the temporal correlation that exists among discrete sampled points in original continuous time signals. In this scenario, a convolutional layer can learn temporal information in the same manner as it learns 2D spatial information in an image classification problem.

It has been empirically proven that a CNN model learns about convolutional filters, which effectively characterize image segments such as object edges (Zeiler and Fergus 2014). For 1D time signals, Cui *et al.* (2016) suggests the use of a multiscale CNN for time-series classification. A previous study (Chase *et al.* 2014) convolves each preselected feature and realizes a bagging model by processing concatenated convolved preselected features in fully connected layers.

In our experiment, we classify two types of vibration signals. Cross entropy, which is a typical choice for a loss function when dealing with a classification problem, is defined by

$$L(p,q) = -\sum_{x} p(x)\log q(x) \tag{1}$$

where p and q are the true label and the predicted probability, respectively. It is interpreted as the sum of the information entropy of our predicted probability $-\log q(x)$ for the given labels $q(x)$.

3.2 Nonlinear convolution

In general, a convolution operator H is defined as

$$H(x) = f\left(k \otimes x + b\right) \tag{2}$$

If the activation function f is assumed to be an identity map (linear convolution), it averages the local values of a discrete time signal weighted by a parameter k, which is a commonly used smoother. Successive application of the linear convolutional operator with varying parameters k results in wavelet transform. However, assuming the activation function f to be nonlinear is more advantageous than assuming it to be linear. It has been proven that outlier values attributed to a heavy-tailed distribution are better treated by the nonlinear processing of signals than by linear processing (Donoho 1993). Successive application of the nonlinear convolutional operator with varying parameters k results in the wavelet-like CNN. In this study, we set f to an exponential linear unit (ELU) so that the continuity in the input vibration signal is not affected.

3.3 Multiresolution network and residual network

Wavelet transformation is a multiresolution analysis. It shrinks the width of a mother wavelet by half for each residual fitting step. Consequently, it analyzes the signals in a diverse spectrum. The wavelet-like CNN model adopts the same strategy by halving the length of the convolutional filter. After the input signal is processed through the convolutional layer, the residual is passed to the next convolutional layer with a half-length convolutional filter, as illustrated in Fig. 3.

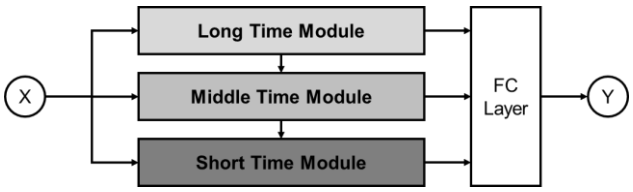


Fig. 3 An architecture of the proposed wavelet-like CNN

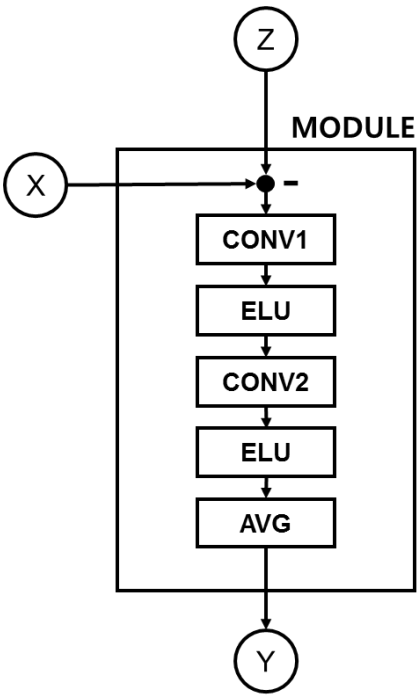
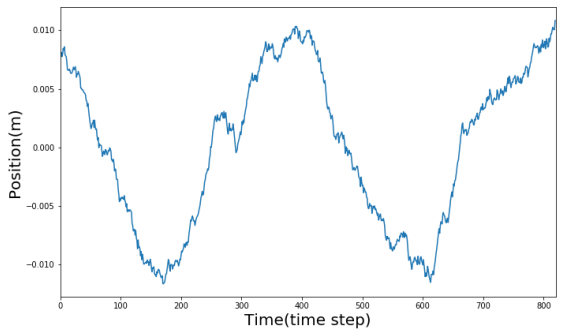
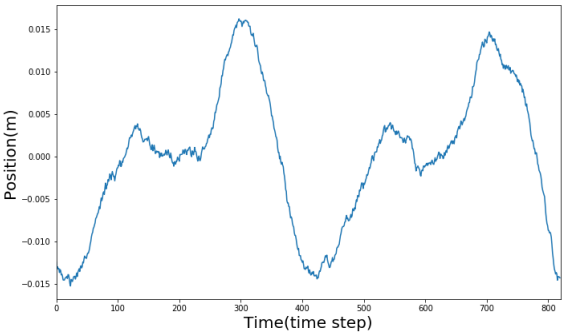


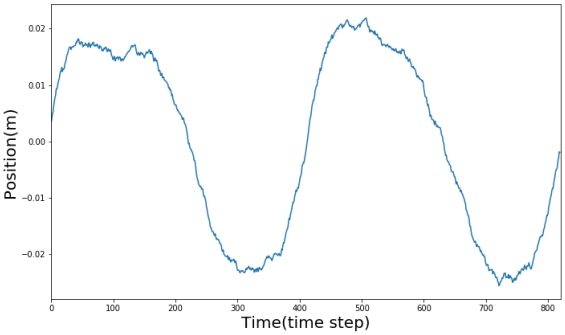
Fig. 4 The module structure



(a)



(b)



(c)

Fig. 5 Vibration data for each mode

In practice, not all convolutional filter sizes ($N/2^{(j)}$ where $i=1,2,\cdots$) need to be defined to prevent the model from overfitting the data. The structure is combined with the wavelet-like CNN. The model for the triple-resolution analysis is illustrated in Fig. 3 where FC layer means fully-connected layer. The module in the structure is illustrated in Fig. 4.

3.4 Multichannel convolution layer

In image data, the variability of 2D shapes is complicatedly entangled in data. Therefore, it is very difficult to learn important features in the images using a single convolutional filter. Therefore, typical CNN-oriented models for image classification problems use multiple convolutional filters to catch a variety of features residing in images. We term this process as "multichannel" analysis to emphasize the fact that images are processed through various "channels" of filters. Usually multiple convolutional filters are represented by a single tensor as follows

$$k \in R^{N\times M} \rightarrow k \in R^{N\times M\times K} \tag{3}$$

where K denotes the number of convolutional filters and N and M denotes height and width of convolutional filters respectively. Similarly, the wavelet-like CNN also uses multiple convolutional filters to learn various features residing in time signals; in the present case, N = 1. Because we have multiple convolutional layers for each time resolution, we use multiple convolutional filters for each time resolution.

4. Experiment

4.1 Data

We acquire and store signals from three malfunction modes of rotor vibrations: (1) normal as a baseline, (2) misalignment, and (3) unbalance. The typical time-series signals of these modes are illustrated in Figs. 5(a)-5(c), respectively. Note that we do not align the phases of vibration data although phase information can be picked up by the Keyphasor probe embedded in the RK 4 Rotor Kit. Because the main goal of this research is to construct a model for signal classification rather than reconstruct or smoothen the original signals, this goal is made clear in the validation process by generalizing the input signals to the maximum extent possible. This generalization can be partially achieved by not synchronizing the phase of each signal.

Every signal has a length of 820 points. Because a single revolution of the rotor is represented by approximately 410 discrete points, by doubling the length, we assume that the semantics of a single revolution can be contained in discrete samples with a length of 820 points, independent of the noise generated by discretization.

4.2 Evaluation measure

We choose five evaluation measures: micro AUC (where AUC is the area under the curve), macro AUC, micro F1 score, macro F1 score, and accuracy. Because the problem pertains to multilabel classification, some variations of the AUC and the F1 score are measured. The detailed theories have been described in previous papers (Chase *et al.* 2014, Lipton *et al.* 2015). All the measures are evaluated based on 10 stratified folds. The measures are briefly described in Table 1.

4.3 Comparison

We compare the evaluation measures obtained using several models. The models are typically categorized into two groups: feature-based machine-learning models and deep-learning models.

To benchmark different classifiers for feature-based models, we use logistic regression and the support vector machine (SVM) with linear and nonlinear radial basis function (RBF) kernels.

Table 1 Evaluation measures

Evaluation measure	Description
Accuracy	Calculate the accuracy classification score.
Micro AUC	Compute the AUC from prediction scores. Calculate metrics globally by considering each element of the label indicator matrix as a label.
Macro AUC	Calculate the AUC for each label, and find their unweighted mean. This does not take label imbalance into account.
Micro F1 score	The F1 score can be interpreted as a weighted average of the precision and the recall; an F1 score reaches its best value at 1 and worst value at 0. Calculate metrics globally by counting the total true positives, false negatives, and false positives.
Macro F1 score	Calculate the F1 score for each label, and find their unweighted mean. This does not take label imbalance into account.

Table 2 Model comparison

Category	Model	Micro AUC	Macro AUC	Micro F1 score	Macro F1 score	Accuracy
Logistic regression (PCA)	Mean	2.63	2.62	2.53	3.34	2.67
SVM with linear kernel	C = 0.1	0.6100	0.5556	0.4800	0.3405	0.4800
	C = 1	0.6150	0.5602	0.4867	0.3447	0.4867
	C = 10	0.6100	0.5556	0.4800	0.3405	0.4800
SVM with RBF kernel (PCA)	C = 0.1	0.8250	0.7546	0.7667	0.5729	0.7667
	C = 1	0.8650	0.8000	0.8200	0.6660	0.8200
	C = 10	**0.8900**	**0.8435**	**0.8533**	**0.7723**	**0.8533**
Logistic regression (wavelet)		0.5500	0.5000	0.4000	0.2950	0.4000
SVM with linear kernel (wavelet)	C = 0.1	0.6200	0.5648	0.4933	0.3512	0.4933
	C = 1	0.6100	0.5556	0.4800	0.3425	0.4800
	C = 10	0.6300	0.5741	0.5067	0.3723	0.5067
SVM with RBF kernel (wavelet)	C = 0.1	0.6200	0.5648	0.4933	0.3512	0.4933
	C = 1	0.6200	0.5648	0.4933	0.3512	0.4933
	C = 10	0.6200	0.5648	0.4933	0.3512	0.4933
MLP	Layer size = [100, 50]	0.9674	0.9225	0.9333	0.9180	0.9333
	Layer size = [200, 100]	0.9275	0.8410	0.9467	0.9333	0.9467
Wavelet coefficient ANN		**0.9564**	**0.9531**	**0.8667**	**0.8442**	**0.8667**
Our model	Kernel size = [100, 200, 400]	**0.9986**	**0.9985**	**0.9933**	**0.9917**	**0.9933**

Variation is achieved by changing the regularization parameter *C*. For deep-learning models, we benchmark the multilayer perceptron (MLP) and a recently suggested deep-learning model using the discrete wavelet coefficient (Jaber and Bicker 2016). We use nine levels of wavelet coefficients, as suggested by Jaber and Bicker (2016) because it is the finest decomposition level that we can achieve using signals with a length of 820 points ($2^9 < 820 < 2^{10}$). The finest decomposition is recommended when a very low frequency band is investigated, which is 25 Hz in our case (Jaber and Bicker 2016). We choose Daubechies' second-order wavelet for analyzing vibration data, as recommended by Jaber and Bicker (2016). We name this model as the "wavelet coefficient ANN."

The comparison results are illustrated in Table 2. In general, linear models show low evaluation measure values, whereas nonlinear models show relatively high evaluation measure values.

Furthermore, deep-learning models show better performance than feature-based ones. It is worth mentioning that the SVM with the RBF kernel based on principal component analysis (PCA) shows better evaluation measures than those based on wavelet coefficients. We argue that this difference occurs because of the high variance generated by the higher dimensions of the wavelet coefficients. Among all the models, our model shows the best performance.

4.4 Evaluation

Qualitative evaluation should be carried out to evaluate the relevance of the features found by our model. Namely, we should most importantly check whether the convolutional filter conveys key information of the rotor dynamics. The convolutional filters of each convolution layer are illustrated in Figure 6. To appreciate the meaning of the filters, we observe the frequency response of the

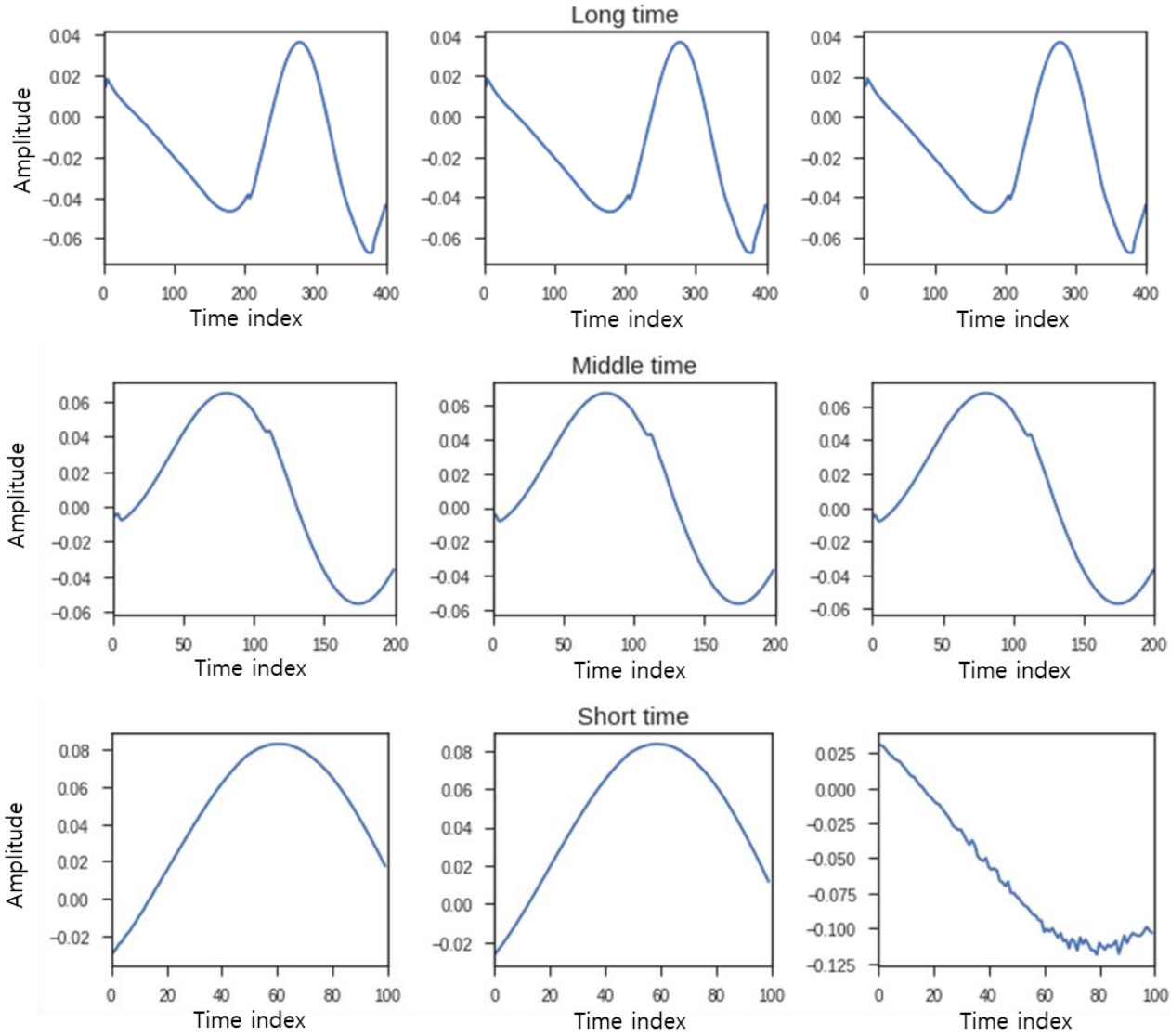


Fig. 6 Convolutional filters

filters (Fig. 7). The red dashed lines represent the 1X and 2X frequencies of the rotor. Long-time filters function like a typical low-pass filter because they multiply each frequency component with a certain decaying factor. They especially emphasize the 1X component. Middle-time filters catch the 1X component of the signal, whereas short-time filters catch the 2X component of the signal. The observations are consistent with the mechanical property that the abnormality of the rotor is typically articulated by variations in the 1X and 2X frequency components. Furthermore, the results are sound in the sense that the 1X component is characterized by relatively longer-time filters (i.e., long- and middle-time filters), whereas the 2X component is characterized by short-time filters.

Fig. 7 indicates that most of the denoising is carried out by the long-time filters. One explanation is the direct connection of the long-time convolution layer with raw signals. Nevertheless, the results suggest that the first convolution step does not possess the intended property (i.e., long-time resolution analysis). Rather, it smoothens signals in advance and proceeds with its analysis using the subsequent middle- and short-time convolutional layers. The process resembles the manner in which the wavelet transform (DWT) works in that the high-frequency components are removed before the signals are convolved with the wavelet of the next level. However, the CNN does not require the subsampling of signals to follow Nyquist's rule. Fig. 6 shows the redundancy among the filters. The figure indicates that the data do not require more than a single convolutional filter. Rather than distorting information, the channels tend to unify different types of behavior within the filters, resulting in a self-regularization effect. Such a regularization effect of convolutional filters is often reported in practice.

During an optimization process, the best evaluation measures are achieved relatively earlier, whereas cross entropy does not reach optimality. This phenomenon suggests that our model has a "feature-learning phase. Fig. 8 shows the resultant convolutional filters when all the evaluation measures are more than 0.98 but cross

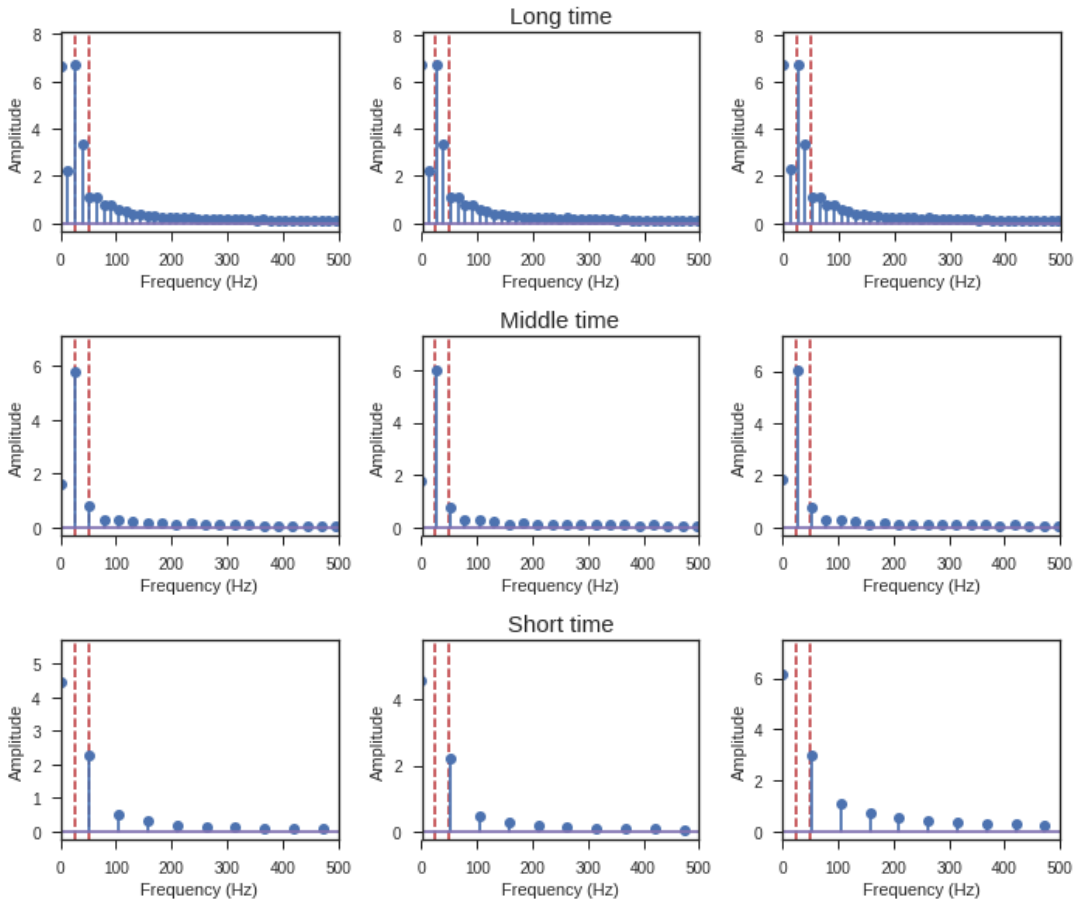


Fig. 7 Frequency spectrum of convolutional filters

entropy is not fully optimized. The spiky shapes found in the filters visually represent the process of the features being searched.

Fig. 9 shows how the wavelet-like CNN activates its neurons while classifying vibration data. CONV1, CONV2, and CONV3 are the convolutional operations in the long-, middle-, and short-time modules, respectively. Because each module consists of two convolutional layers (Figure 4), each module gives two outputs. The actual outputs are CONV1_2, CONV2_2, and CONV3_2. CONV1_2 shows the moderate smoothing step for the original data. The subsequent outputs CONV2_2 and CONV3_2 show the 1X and 2X frequency activation because CONV2_2 is activated at the 1X component more strongly (red), whereas CONV3_2 is activated more frequently. These results coincide with the results of our analysis shown in Figs. 6-8. A summary of the entire activation process is presented in Fig. 10.

5. Conclusions

Our goals are to (1) combine smoothing and classification procedures into a single CNN structure, (2) achieve high accuracy, and (3) evaluate how well the proposed model learns features from time-series data. Validation results based on rotor vibration data suggest that these goals are moderately achieved; the proposed evaluation measures show better performance than both off-the-shelf feature-based models and the deep-learning models recently proposed in the PHM field.

The novelty of our study lies in integrating the "residual fitting" mechanism of wavelet shrinkage into the convolutional neural network structure. Consequently, the wavelet-like CNN architecture combines signal reconstruction and classification procedures into a single model; these procedures are currently carried out exhaustively based on expert knowledge.

Acknowledgments

This work was partially supported by the SW fusion technology upgrading (R&D) for new industry creation (S0503-17-1011), by the Leading Human Resource Training Program of Regional Neo-industry through the National Research Foundation of Korea funded by the Ministry of Science, ICT, and Future Planning (NRF-2016H1D5A1910285), by a technology innovation project (S2439592) of the Small and Medium Business Administration, and the Korea Evaluation Institute of Industrial Technology (project ID: 10080729).

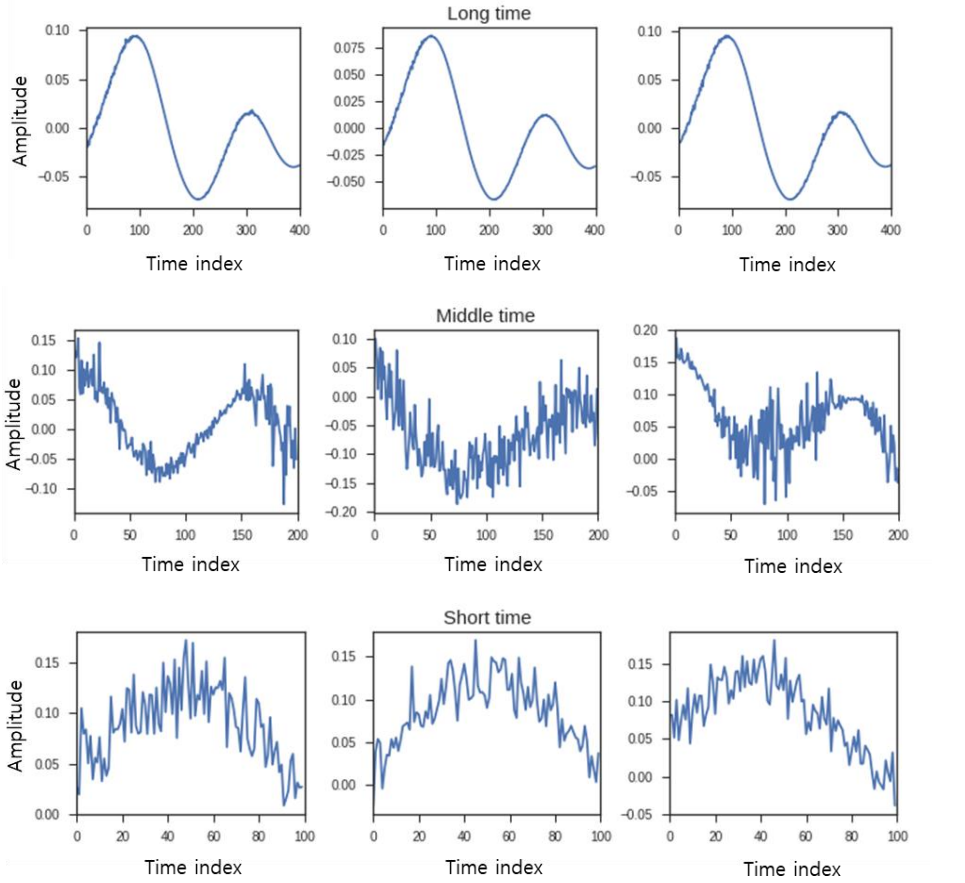


Fig. 8 Convolutional filters when cross entropy is not fully optimized

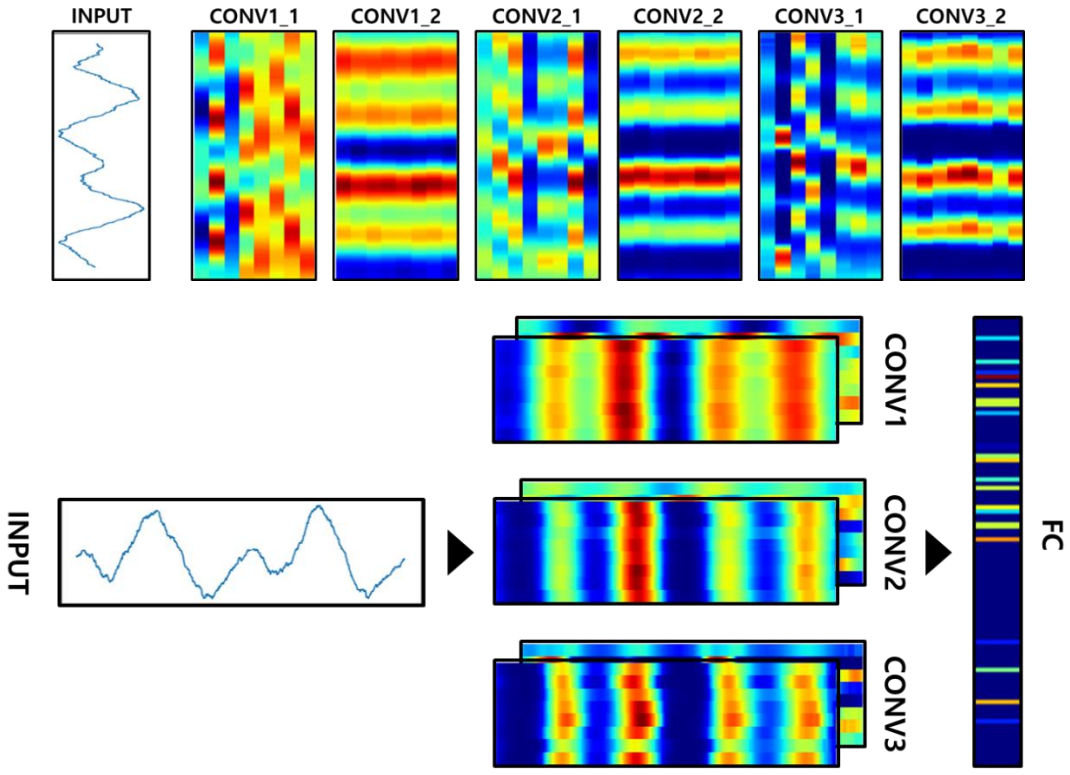


Fig. 9 Entire activation process

References

Borghetti, A., Bosetti, M., Di Silvestro, M., Nucci, C.A. and Paolone, M. (2008), "Continuous-wavelet transform for fault location in distribution power networks: Definition of mother wavelets inferred from fault originated transients", *IEEE T. Power Syst.*, **23**(2), 380-388.

Chase Lipton, Z., Elkan, C. and Narayanaswamy, B. (2014), "Thresholding Classifiers to Maximize F1 Score", *arXiv preprint arXiv:1402.1892.*

Chen, Z., Li, C. and Sanchez, R.V. (2015), "Gearbox fault identification and classification with convolutional neural networks", *J. Shock Vib.*, **2015.**

Cui, Z., Chen, W. and Chen, Y. (2016), "Multi-scale convolutional neural networks for time series classification", *arXiv preprint arXiv:1603.06995.*

Donoho, D.L. (1993), "Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data", *In Proceedings of Symposia in Applied Mathematics.*

Donoho, D L. (1993), "Unconditional bases are optimal bases for data compression and for statistical estimation", *Appl. Comput. Harmon. A.*, **1**(1), 100-115.

Donoho, D.L. and Johnstone, I.M. (1994), "Ideal spatial adaptation by wavelet shrinkage", *Biometrika*, 425-455.

Donoho, D.L. and Johnstone, I.M. (1995), "Adapting to unknown smoothness via wavelet shrinkage", *J. Am. Statist. Sssociation*, **90**(432), 1200-1224.

Donoho, D.L. and Johnstone, J.M. (1994), "Ideal spatial adaptation by wavelet shrinkage", *Biometrika*, **81**(3), 425-455.

Gelman, L., Patel, T.H., Persin, G., Murray, B. and Thomson, A. (2013), "Novel technology based on the spectral kurtosis and wavelet transform for rolling bearing diagnosis", *Int. J. Prognost. Health Management*, 2153-2648.

Gelman, L., Petrunin, I., Jennions, I.K. and Walters, M. (2012), "Diagnostics of local tooth damage in gears by the wavelet technology", *Int. J. Prognost. Health Management*, **3**(52).

He, K., Zhang, X., Ren, S. and Sun, J. (2016), "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jaber, A.A. and Bicker, R. (2016), "Industrial robot backlash fault diagnosis based on discrete wavelet transform and artificial neural network", *Am. J. Mech. Eng.*, **4**(1), 21-31.

Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S. and Van Hoecke, S. (2016), "Convolutional neural network based fault detection for rotating machinery", *J. Sound Vib.*, **377**, 331-345.

Jeong, H., Park, S., Woo, S. and Lee, S. (2016), "Rotating machinery diagnostics u sing deep learning on orbit plot images", *Procedia Manufact.*, **5**, 1107-1118.

Jung, U., and Koh, B. (2014), "Bearing fault detection through multiscale wavelet scalogram-based SPC", *Smart Struct. Syst.*, **14**(3), 377-395.

Korekado, K., Morie, T., Nomura, O., Ando, H., Nakano, T., Matsugu, M. and Iwata, A. (2003), "A convolutional neural network VLSI for image recognition using merged/mixed analog-digital architecture", *In Knowledge-Based Intelligent Information and Engineering Systems*, 169-176. Springer Berlin/Heidelberg.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R E., Hubbard, W. and Jackel, L.D. (1989), "Backpropagation applied to handwritten zip code recognition", *Neural Comput.*, **1**(4), 541-551.

Lipton, Z.C., Kale, D.C., Elkan, C. and Wetzell, R. (2015), "Learning to diagnose with LSTM recurrent neural networks", *arXiv preprint arXiv:1511.03677.*

Matsugu, M., Mori, K., Mitari, Y. and Kaneda, Y. (2003), "Subject independent facial expression recognition with robust face detection using a convolutional neural network", *Neural Networks*, **16**(5), 555-559.

Rafiee, J., Rafiee, M.A. and Tse, P.W. (2010), "Application of mother wavelet functions for automatic gear and bearing fault diagnosis", *Exp. Syst. Appl.*, **37**(6), 4568-4579.

Rafiee, J., Tse, P.W., Harifi, A. and Sadeghi, M.H. (2009), "A novel technique for selecting mother wavelet function using an intelligent fault diagnosis system", *Exp. Syst. Appl.*, **36**(3), 4862-4875.

Sawicki, J.T., Sen, A.K. and Litak, G. (2009), "Multiresolution wavelet analysis of the dynamics of a cracked rotor", *Int. J. Rotat. Machinery*, 2009.

Williams, D.R.G.H.R. and Hinton, G.E. (1986), "Learning representations by back-propagating errors", *Nature*, **323**(6088), 533-538.

Zeiler, M.D. and Fergus, R. (2014), "Visualizing and understanding convolutional networks", *In European conference on computer vision*, 818-833.

Zhang, L., Zhou, W. and Jiao, L. (2004), "Wavelet support vector machine", *IEEE T. Syst. Man. Cy. B*, **34**(1), 34-39.

Zhou, C., Li, H., Li, D., Lin, Y. and Yi, T. (2013), "Online damage detection using pair cointegration method of time-varying displacement", *Smart Struct. Syst.*, **12**(3), 309-325.