Development of Bearing Fault Diagnosis Model Using Low Frequency Data Based on Knowledge Distillation

Yongjae Jeon[1], Secheol Yang[2], and Sang Won Lee[3]

[1,2]*Department of Mechanical Engineering, Sungkyunkwan University, Suwon-si, Gyeonggi-do, 16419, Republic of Korea*
rhrhgudwp@g.skku.edu
didtp1028@skku.edu

[3]*School of Mechanical Engineering, Sungkyunkwan University, Suwon-si, Gyeonggi-do, 16419, Republic of Korea*
sangwonl@skku.edu

ABSTRACT

Bearings are critical components for ensuring smooth rotational motion in mechanical systems, and reliable operation requires continuous condition monitoring for fault diagnosis. Recently, there has been growing interest in diagnosing bearing conditions using artificial intelligence, particularly deep learning-based approaches. However, in real industrial environments, limitations such as high sensor cost and restricted data storage often lead to the use of low sampling frequency sensor data, which poses challenges in developing accurate diagnosis models. To address this issue, this paper proposes a bearing fault diagnosis method based on knowledge distillation to enhance the utility of low sampling frequency data. High-frequency acceleration data were collected under both normal and faulty conditions and subsequently downsampled for knowledge distillation. A 1D CNN-based teacher model was trained using high-frequency data, and multiple loss functions were designed to distill both final predictions and intermediate features into a student model trained on low sampling frequency data. The performance comparison between models with and without knowledge distillation verified the effectiveness of the proposed approach. The results demonstrate the feasibility of developing fault diagnosis models using low sampling frequency data in real industrial settings and suggest an effective knowledge distillation strategy.

Keywords: Knowledge Distillation, Bearing Fault Diagnosis, Low Sampling Frequency Data, 1D Convolutional Neural Network

1. INTRODUCTION

In rotating machinery, bearings are critical components that support the rotational motion of the shaft, reduce friction, and ensure precision and stability. In particular, rolling bearings offer advantages such as low friction loss, high load-carrying capacity, and excellent mechanical efficiency, making them widely used across various industrial applications. However, faults in bearings pose serious threats to the reliability and safety of the entire system, often leading to unexpected equipment shutdowns or cascading failures. Such faults can result in reduced equipment availability, decreased productivity, and increased maintenance costs. Therefore, early fault diagnosis and preventive maintenance of bearings are essential tasks in industrial settings (Jia, Lei, Shan & Lin, 2015).

In recent years, Artificial Intelligence (AI), especially Deep Learning (DL)-based diagnostic techniques, has gained significant attention in the field of bearing fault diagnosis. DL models can automatically learn and generalize from complex vibration signals and fault patterns, achieving higher diagnostic accuracy and flexibility compared to conventional methods. Zhang, Zhang, Wange, and Habetler (2019) conducted a comprehensive review comparing Machine Learning (ML) and DL algorithms, concluding that DL methods outperform conventional machine learning techniques in feature extraction and classification performance for bearing fault diagnosis. Siddique, Saleem, Umar, Kim, and Kim (2025) proposed a hybrid DL architecture combining Continuous Wavelet Transform (CWT)-based time–frequency feature extraction with attention-enhanced BiLSTM and 1D convolutional ResNet, demonstrating robust diagnostic performance under noisy environments and nonstationary signal conditions. Cui, Zhang, Zhong, Hou, Chen, Cai, and Kim (2025) developed a lightweight DL model that incorporates a multiscale feature extraction structure with a Ghost module and Efficient

Channel Attention (ECA), achieving over 99.4% diagnostic accuracy while maintaining computational efficiency, thereby significantly improving practical applicability.

Although AI-based bearing fault diagnosis techniques have shown high performance across diverse conditions, most studies rely on high sampling frequency data. While high sampling frequency data can precisely capture variations in rotational speed, load, and environmental factors, they are often expensive to obtain, require substantial resources for data acquisition, storage, and real-time processing, and face installation constraints in industrial environments. In many real-world cases, only low sampling frequency data are available, resulting in the loss of high frequency fault components and a significant degradation in diagnostic performance.

To address this limitation, this study proposes a method for achieving high diagnostic accuracy in low sampling frequency environments by applying the Knowledge Distillation (KD) technique. Specifically, knowledge acquired by a teacher model trained on high sampling frequency data is transferred to a student model trained on low sampling frequency data, thereby compensating for information loss and overcoming the performance limitations of conventional low frequency models. This approach enables cost-effective, lightweight sensor systems to deliver stable and accurate fault diagnosis in industrial applications, which is the core contribution of this work.

2. METHODOLOGY

2.1. Knowledge Distillation

Knowledge Distillation was first proposed by Hinton, Vinyals, and Dean (2015) as a technique to transfer the knowledge learned by a complex and high-performing large-scale model (Teacher) to a smaller and more lightweight model (Student), thereby achieving model compression and faster inference while maintaining performance. Traditionally, knowledge distillation aims to reduce the number of parameters or memory usage without significant loss in accuracy. Romero, Ballas, Kahou, Chassang, Gatta, and Bengio (2015) extended this concept by proposing feature-based distillation, which transfers not only the output predictions but also the intermediate feature maps from the Teacher to the Student, allowing the Student to better mimic the representational capacity of the Teacher.

The loss function of knowledge distillation generally consists of the Cross-Entropy (CE) loss with respect to the original labels and the distillation loss that mimics the Teacher's output probability distribution. In this study, an additional Mean Squared Error (MSE) loss is applied to the intermediate feature maps, and the final loss function is defined as follows:

$$L_{total} = (1-\beta)\left[(1-\alpha)\cdot L_{CE}(\sigma(Z_s),\hat{y}) + \alpha T^2 \cdot L_{CE}\left(\sigma(Z_s/T),\sigma(Z_t/T)\right)\right] + \beta \cdot MSE(F_s,F_t) \quad (1)$$

The total loss function L_{total} consists of two main components:

1. Output-based knowledge distillation term: This combines the standard classification loss $L_{CE}(\sigma(Z_s),\hat{y})$, with the distillation loss$L_{CE}\left(\sigma(Z_s/T),\sigma(Z_t/T)\right)$, which encourages the Student model to replicate the Teacher's softened output probability distribution. The weighting parameter α balances these two terms, while the temperature scaling factor T adjusts the smoothness of the output distribution to convey richer information about inter-class relationships. The multiplication by T^2 compensates for the gradient scaling effect introduced by temperature scaling.
2. Feature-based distillation term: This is formulated as the mean squared error (MSE) between the intermediate feature maps of the Student and Teacher, $MSE(F_s,F_t)$, encouraging the Student to replicate the Teacher's internal representations. The parameter β controls the trade-off between output-based and feature-based distillation losses.

By adjusting α and β, the framework allows flexible control over the influence of hard label supervision, output-level distillation, and feature-level distillation. This design ensures that the Student model learns not only the final decision boundaries of the Teacher but also its intermediate feature representations, which is particularly beneficial in scenarios with low sampling frequency input data.

2.2. Proposed Framework

In this study, knowledge distillation is applied not for model compression, but to overcome limitations in data frequency resolution. In real industrial environments, it is often difficult to reliably collect high sampling frequency data due to constraints such as sensor cost, installation space, and data transmission bandwidth. To address this, the proposed method trains the Teacher model using high sampling frequency data only during the training stage, and in the inference stage, performs fault diagnosis using only the Student model that takes low sampling frequency data as input, as illustrated in Figure 1, thereby achieving high diagnostic performance without requiring high-frequency data during deployment.

The proposed framework consists of the following steps:

1. Teacher model training: Train the Teacher model using high-sampling-frequency data to effectively extract fault-related features in the high frequency domain. In this study, 1D Convolutional Neural Network (1D CNN)-based architecture is used.
2. Student model construction: Design the Student model with the same architecture as the Teacher model but use

low sampling frequency input data obtained through downsampling to reflect sensor constraints in industrial environments.

3. Distillation training process: During Student training, use both the output-based distillation loss, which mimics the Teacher model's final output probability distribution, and the feature-based distillation loss, which mimics the intermediate layer features.
4. Model deployment stage: Once training is complete, the Teacher model and high sampling frequency data are no longer required, and fault diagnosis is performed solely by the Student model using low sampling frequency data.

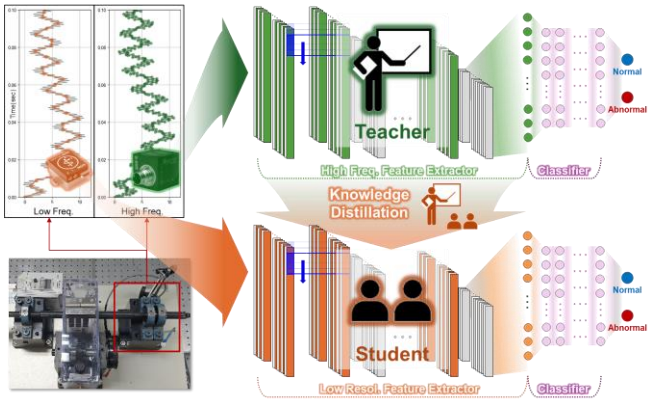


Figure 1. Framework of Knowledge Distillation-based fault diagnosis model

3. EXPERIMENT SETUP AND DATA PREPARATION

3.1. High Frequency Data Acquisition

In this study, a bearing fault diagnosis experiment was conducted using a bearing simulator available in our laboratory, shown as Figure 2. The simulator consists of a driving motor, a shaft, and bearings. The fault type used in the experiment was a ball fault, and data were collected for both healthy and faulty bearings. The shaft rotational speed was set to 2,000 rpm(33.33 Hz), and a high-performance accelerometer was used to acquire continuous vibration signals at a sampling frequency of 1,000 Hz for 100 seconds. The acquired signals were segmented into samples of 0.2 seconds each, resulting in 500 high-frequency samples for both the healthy and faulty conditions. Figure 3 presents examples of high-frequency signals from healthy and faulty bearings.

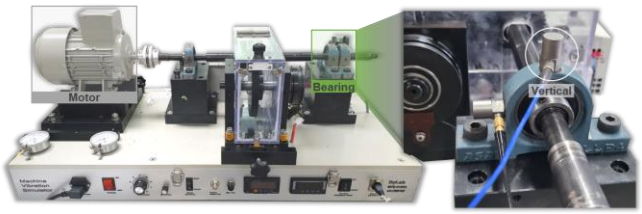
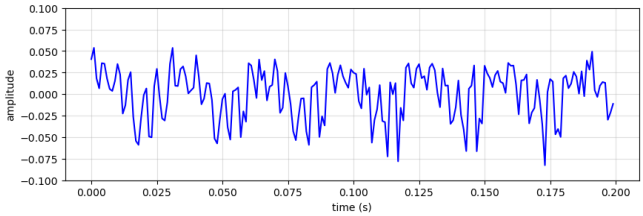
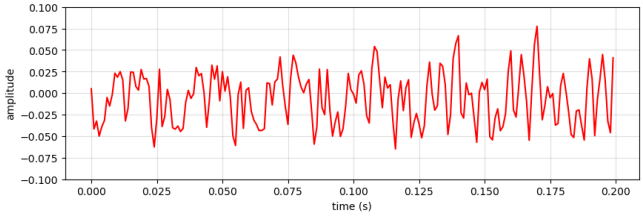


Figure 2. Bearing simulator and sensor setup position



(a) Normal bearing vibration data

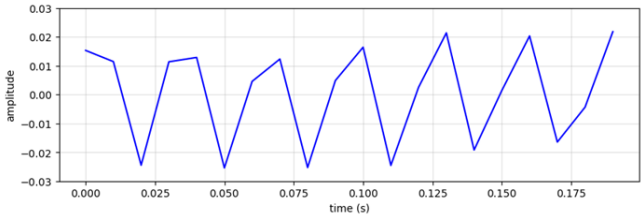


(b) Fault bearing vibration data

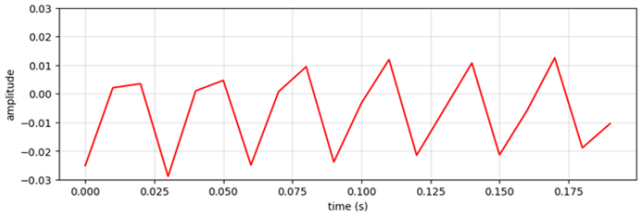
Figure 3. High frequency sensor data graph

3.2. Low Frequency Data Preparation Using Decimation

To emulate the constraints of sensors typically used in industrial environments, the collected high frequency data (1,000 Hz) were converted into low frequency data using a decimation technique. First, a low-pass filter was applied to remove high frequency components and prevent aliasing that may occur during downsampling. Then, the signals were downsampled to a sampling frequency of 100 Hz. Both healthy and faulty datasets underwent the same procedure, and examples of the resulting low-frequency signals are shown in Figure 4.



(a) Normal bearing vibration data



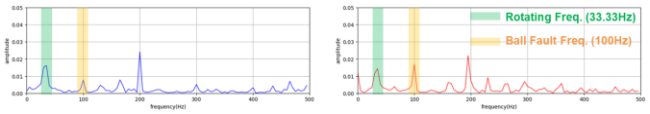
(b) Fault bearing vibration data

Figure 4. Low frequency sensor data graph using decimation

Compared to the original high frequency data, the transformed low frequency data exhibited the following characteristic changes. First, the removal of high-frequency components simplified and smoothed the signal waveform.

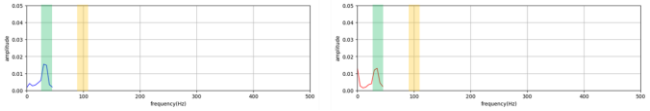
Second, the overall amplitude decreased. Third, fine vibration details and irregular peaks present in the high frequency range disappeared, making it more difficult to directly identify fault indications. Fourth, in the frequency domain, the energy became concentrated in the low frequency range, while the energy in the high frequency range dropped significantly.

In addition, Fast Fourier Transform (FFT) analysis was conducted to compare the frequency-domain characteristics of high frequency and low frequency data. In Figure 5, The results showed that the rotational frequency component at approximately 33.33 Hz was clearly observable in both high frequency and low frequency data. However, the fault frequency component at approximately 100 Hz was distinct in the high frequency data but could not be observed in the low frequency data due to the sampling frequency limitation. This indicates that fault-related features can be lost in low frequency data, which serves as one of the key motivations for applying the knowledge distillation approach in this study.



Normal Fault

(a) High Frequency data



Normal Fault

(b) Low Frequency data

Figure 5. FFT result of high and low frequency data

3.3. Dataset Composition

The dataset obtained from both high-frequency and low-frequency measurements was divided into training and validation sets to develop and evaluate the diagnostic models. The splitting ratio was set to 4:1. As a result, the training set consisted of 400 samples for each condition (normal and fault), while the validation set contained 100 samples for each condition. This ensured that both datasets maintained class balance, preventing bias toward a specific condition during the training process and allowing for a reliable evaluation of model performance.

4. RESULTS

4.1. Model Architecture

In this study, both the Teacher and Student models for bearing fault diagnosis were implemented using an identical 1D Convolutional Neural Network (1D CNN) architecture. The input layer accepts raw time-series sensor data, followed by a feature extraction module composed of four 1D convolutional layers and one max pooling layer.

The first convolutional layer employs four filters with a kernel size of three and uses the ReLU (Rectified Linear Unit) activation function to extract low-level features. The second convolutional layer uses eight filters with the same kernel size and activation function. This is followed by a max pooling layer with a pool size of two and a stride of two, which reduces the temporal dimension by half while retaining critical patterns and suppressing noise. The third and fourth convolutional layers utilize 16 and 32 filters respectively, both with a kernel size of three and ReLU activation, to capture high-level features effectively.

Instead of feeding the extracted features directly into a fully connected layer, a Global Average Pooling (GAP) layer is applied to compute global statistics for each channel, thereby reducing the number of parameters and preventing overfitting. The subsequent hidden layer consists of eight neurons, also using the ReLU activation function for non-linear transformation. Finally, the output layer consists of two neurons with a Softmax activation function to produce probability distributions over the normal and fault classes.

For training, the Adam optimizer and categorical crossentropy loss function were employed, with a batch size of 32, learning rate of 0.001, and 1000 epochs.

4.2. Baseline Model Training Results Using Low Frequency Data

When training the baseline model with only low frequency data (100 Hz), as shown in Figure 6, the validation loss failed to converge and exhibited divergence in the later stages. This instability is attributed to the loss of critical fault frequency components (around 100 Hz) during the downsampling process, which made it difficult to distinguish between normal and faulty conditions. In the training curves of Figure 6, the loss initially decreases slightly but validation loss increases sharply in the later stages, indicating overfitting and classification instability. The final validation accuracy was 60.5%, which is insufficient for reliable fault classification.

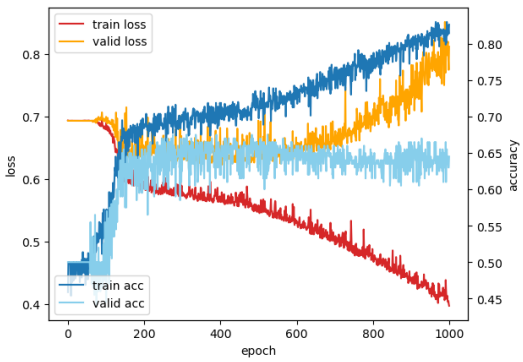


Figure 6. Loss and accuracy of baseline model during training

4.3. Teacher Model Training Results Using High Frequency Data

When the same architecture was trained using high frequency data (1000 Hz), as shown in Figure 7, both training and validation losses decreased steadily, achieving a final validation accuracy of 100%. This demonstrates that high frequency data retains sufficient fault frequency components, enabling clear discrimination between normal and faulty conditions. Therefore, this high frequency trained model was designated as the Teacher model, which was subsequently used to enhance the performance of the low frequency Student model through knowledge distillation.

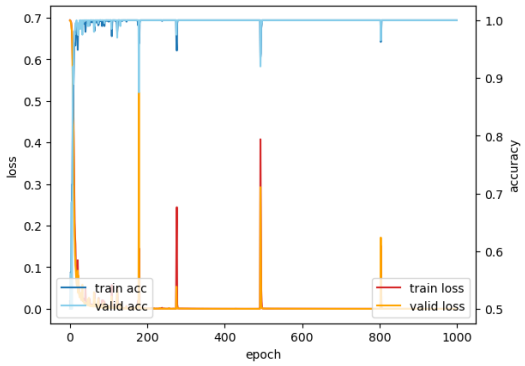


Figure 7. Loss and accuracy of Teacher model during training

4.4. Knowledge Distillation and Optimization

Knowledge distillation was applied to transfer the knowledge from the Teacher model to the Student model, aiming to achieve high classification accuracy using only low frequency data. In this study, both output-based distillation loss and feature-based distillation loss were employed. The weighting coefficient for standard knowledge distillation loss (α), the weighting coefficient for balancing standard and feature distillation losses (β), and the temperature parameter (T) for generating soft targets were all optimized.

The search ranges were set as follows: α from 0.0 to 1.0 with an interval of 0.1, β from 0.0 to 1.0 with an interval of 0.1, T from 1 to 8, and learning rate in {0.01, 0.001, 0.0001, 0.00001}. The optimal parameters were found to be $\alpha = 0.1$, $\beta = 0.1$, temperature = 2, and learning rate = 0.0001. With these settings, as shown in Figure 8, both training and validation accuracies reached 100 percent, indicating that the Student model successfully overcame the information loss caused by downsampling by effectively leveraging high-frequency knowledge from the Teacher model.

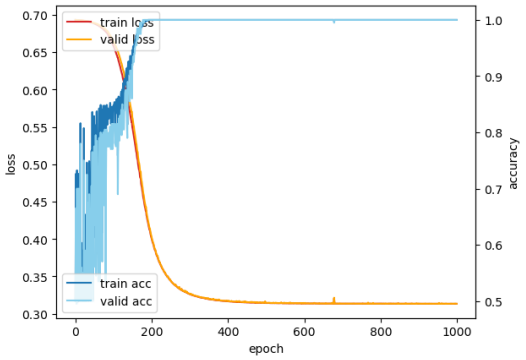
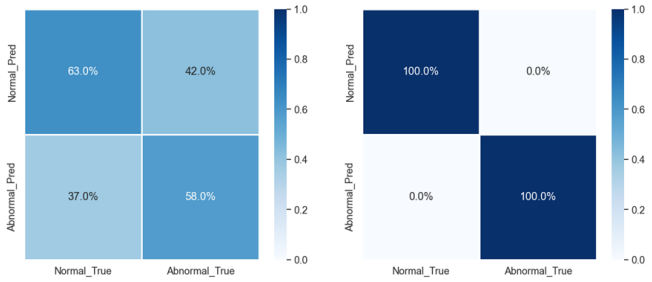


Figure 8. Loss and accuracy of knowledge distillation-based Student model during training

4.5. Performance Comparison

The comparison of confusion matrices in Figure 9 highlights the performance difference before and after applying knowledge distillation. Without distillation as shown in Figure 9 (a), a considerable portion of faulty samples were misclassified as normal, significantly lowering fault detection sensitivity. In fact, over half of the faulty samples were incorrectly labeled, undermining the reliability of the classification.

In contrast, after applying knowledge distillation as shown in Figure 9 (b), all normal and faulty samples were correctly classified, achieving 100% classification accuracy. This indicates that the high frequency features transferred from the Teacher model allowed the Student model to establish a more precise decision boundary. Consequently, the proposed knowledge distillation-based training strategy effectively compensates for the limitations of low-resolution data and demonstrates strong potential for real-world industrial applications.



(a) Before KD (b) After KD

Figure 9. Confusion matrix of before and after application of knowledge distillation

5. CONCLUSION

This study proposed a knowledge distillation-based fault diagnosis framework capable of achieving high diagnostic accuracy using low sampling frequency sensor data. Through bearing simulator experiments, the Teacher model trained on

high-frequency data successfully transferred its knowledge to the Student model trained on low sampling frequency data, allowing comparable diagnostic performance. While the proposed approach demonstrated the feasibility of accurate diagnosis with low-cost sensing systems, it still requires high sampling frequency data during training and may be sensitive to sensor installation differences. Future research will focus on applying the proposed method in real industrial environments and extending it to various fault types, such as inner race and outer race faults, to further verify its generalizability.

ACKNOWLEDGEMENT

This work was supported by the Technology Innovation Program (20012807, Development of Customized Smart HMI Systems) and the Industry Technology Alchemist Project (20025702, Development of smart manufacturing multiverse platform based on multisensory fusion avatar and interactive AI) by the Ministry of Trade, Industry & Energy (MOTIE, Korea), and the Technology Development Program (S3430541) funded by the Ministry of SMEs and Startups (MSS, Korea).

NOMENCLATURE

L_{total}	total loss
L_{CE}	cross-entropy loss
σ	softmax function
Z_s	logits from student model
Z_t	logits from teacher model
$\hat{y}$	ground-truth label
T	temperature parameter
F_s	intermediate feature map from student model
F_t	intermediate feature map from teacher model
α	ratio between CE loss and distillation loss
β	ratio between output-based distillation loss and feature-based distillation loss

REFERENCES

Jia, F., Lei, Y., Shan, H., & Lin, J. (2015). Early fault diagnosis of bearings using an improved spectral kurtosis by maximum correlated kurtosis deconvolution. *Sensors*, vol. 15(11), pp. 29363-29377.

Zhang, S., Zhang, S., Wang, B., & Habetler, T. G. (2019). Machine learning and deep learning algorithms for bearing fault diagnostics—A comprehensive review. arXiv 2019. *arXiv preprint arXiv:1901.08247*.

Siddique, M. F., Saleem, F., Umar, M., Kim, C. H., & Kim, J. M. (2025). A hybrid deep learning approach for bearing fault diagnosis using continuous wavelet transform and attention-enhanced spatiotemporal feature extraction. *Sensors*, vol. 25(9), pp. 2712.

Cui, Y., Zhang, Z., Zhong, Z., Hou, J., Chen, Z., Cai, Z., & Kim, J. H. (2025). Bearing Fault Diagnosis Based on Multiscale Lightweight Convolutional Neural Network. *Processes*, vol. 13(4), pp. 1239.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). FitNets: Hints for Thin Deep Nets. *arXiv preprint arXiv: 1412.6550*.

BIOGRAPHIES

Yongjae Jeon. He is a Ph.D. candidate of the Department of Mechanical Engineering, Sungkyunkwan University, Suwon, Korea. He received his B.S. in System Management Engineering from Sungkyunkwan University. His research interests are prognostics and health management for smart manufacturing, and process optimization.

Secheol Yang. He is a master student in the Department of Mechanical Engineering, Sungkyunkwan University, Suwon, Korea. He received his B.S. in the School of Civil, Architectural Engineering and Landscape Architecture from Sungkyunkwan University. His research interest is prognostics and health management for smart manufacturing.

Sang Won Lee. He is a Professor of the School of Mechanical Engineering, Sungkyunkwan University, Suwon, Korea. He received his Ph.D. in Mechanical Engineering from the University of Michigan, Ann Arbor, MI, USA. His research interests include smart factory, PHM, cyber-physical system, environmentally-friendly mechanical machining, additive manufacturing, and data-driven design.