Semi-Supervised Bearing Fault Diagnosis and Classification using Variational Autoencoder-Based Deep Generative Models

Zhang, Shen; Ye, Fei; Wang, Bingnan; Habetler, Thomas G

TR2021-002 January 23, 2021

Abstract

Many industries are evaluating the use of the Internet of Things (IoT) technology with big data to perform remote monitoring and predictive maintenance on their mission-critical assets and equipment, for which mechanical bearings are their indispensable components. Although many data-driven methods have been applied to bearing fault diagnosis, most of them belong to the supervised learning paradigm that requires a large amount of labeled training data to be collected in advance. However, in practical applications, obtaining accurate labels based on real-time bearing conditions can be more challenging than simply collecting large amounts of unlabeled data using multiple sensors. In this paper, we thus propose a semi-supervised learning scheme for bearing fault diagnosis using variational autoencoder (VAE)-based deep generative models, which can effectively utilize a dataset when only a small subset of data have labels. Finally, a series of experiments were conducted using the University of Cincinnati Intelligent Maintenance System (IMS) Center dataset and the Case Western Reserve University (CWRU) bearing dataset. The experimental results show that the proposed semi-supervised learning schemes outperformed some mainstream supervised and semi-supervised benchmarks with the same percentage of labeled data samples. Additionally, the proposed methods can mitigate the label inaccuracy issues when identifying naturally-evolved bearing faults.

IEEE Sensors Journal

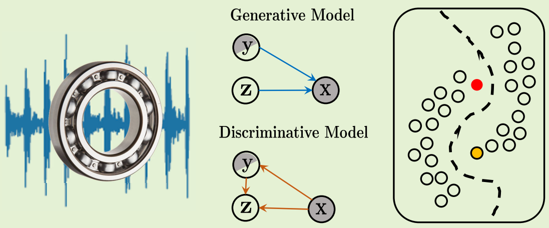
Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

[1]-eps-converted-to.pdf

Semi-Supervised Bearing Fault Diagnosis and Classification using Variational Autoencoder-Based Deep Generative Models

Shen Zhang, *Member, IEEE,* Fei Ye, *Member, IEEE,* Bingnan Wang, *Senior Member, IEEE,* and Thomas G. Habetler, *Fellow, IEEE*

***Abstract*— Many industries are evaluating the use of the Internet of Things (IoT) technology with big data to perform remote monitoring and predictive maintenance on their mission-critical assets and equipment, for which mechanical bearings are their indispensable components. Although many data-driven methods have been applied to bearing fault diagnosis, most of them belong to the supervised learning paradigm that requires a large amount of labeled training data to be collected in advance. However, in practical applications, obtaining accurate labels based on real-time bearing conditions can be more challenging than simply collecting large amounts of unlabeled data using multiple sensors. In this paper, we thus propose a semi-supervised learning scheme for bearing fault diagnosis using variational autoencoder (VAE)-based deep generative models, which can effectively utilize a dataset when only a small subset of data have labels. Finally, a series of experiments were conducted using the University of Cincinnati Intelligent Maintenance System (IMS) Center dataset and the Case Western Reserve University (CWRU) bearing dataset. The experimental results show that the proposed semi-supervised learning schemes outperformed some mainstream supervised and semi-supervised benchmarks with the same percentage of labeled data samples. Additionally, the proposed methods can mitigate the label inaccuracy issues when identifying naturally-evolved bearing faults.**



***Index Terms*— Bearing fault, generative model, semi-supervised learning, variational autoencoders.**

I. Introduction

THE Internet of Things (IoT) is a system that connects many devices together and transfers their data over a network [1]. By connecting these devices, such as simple sensors, smartphones, and wearables to automated systems, it is possible to gather information, analyse it, and take appropriate actions to learn from a process or fulfill a specific task. According to [2], companies in many industries are evaluating the ability to use IoT technology to perform remote monitoring and predictive maintenance on their mission-critical applications. In particular, a key functional component of assets and equipment in many industries is the mechanical bearing, which is responsible for a variety of applications such as planes, vehicles, production machinery, wind turbines, air-conditioning systems, elevator hoists, among others.

These IoT-based bearing diagnosis tasks typically collect a large amount of data from their interconnected sensors, such as the frequently-used vibration [3], [4], acoustic [5], [6], and motor current [7], [8] sensors. These signals are typically rich in high-dimensional features related to bearing defects, which makes it well-suited to leverage deep learning algorithms to extract these fault features and thereafter perform anomaly detection [9]–[11]. Despite their success, most of the existing models are developed in the form of supervised learning, which requires a large set of labeled data collected in advance for each distinct operating condition.

Since data is typically collected by means of sensors without human intervention, it might not be difficult to obtain a sufficient amount of data for supervised bearing fault diagnosis [12]. However, the process of labeling the collected samples can be time-consuming [13], [14] and expensive [13]–[18], and it also requires human knowledge/expertise on the system states [12]. Therefore, the bearing dataset, especially the faulty data, are usually not labeled in real industrial applications [14], [19]. Even attempts are made to label these unlabeled samples, the accuracy of these labels cannot be guaranteed, since they are also subject to confirmational data biases of the engineers interpreting the data [17]. Therefore, both label scarcity and label accuracy issues will pose challenges to the mainstream supervised learning approaches for bearing fault diagnosis.

A promising approach to overcome these challenges is to

S. Zhang and T. G. Habetler are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA (e-mail: shenzhang@gatech.edu, tom.habetler@ece.gatech.edu).

F. Ye is with California PATH, University of California, Berkeley, Berkeley CA, USA (e-mail: fye@berkeley.edu).

B. Wang is with Mitsubishi Electric Research Laboratories, Cambridge MA, USA (e-mail: bwang@merl.com).

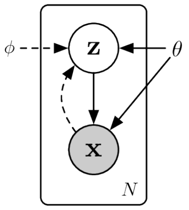


Fig. 1. Architecture of variational autoencoders [20].

apply semi-supervised learning algorithms that leverage the limited labeled data and the massive unlabeled data simultaneously [12]–[19]. Specifically, semi-supervised learning considers the classification problem when only a small part of the data has labels, and so far only a few semi-supervised learning paradigms have been applied to bearing fault diagnosis. For instance, the support vector data description method in [19] uses cyclic spectral coherent domain indicators to construct a feature space and fit a hypersphere, which then calculates the Euclidean distance in order to distinguish the faulty data from the healthy ones. In addition, both [15] and [16] use graph-based methods to construct graphs connecting similar samples in the dataset, so class labels can be propagated from labeled nodes to unlabeled nodes through the graph. However, these methods are very sensitive to graph structure and need to analyze the graph's Laplacian matrix, which limits the scope of these methods. [12] uses α-shape instead of a graph-based method to capture the data structure, and the α-shape is mainly used to perform surface estimation and to reduce the efforts required for parameter tuning.

Moreover, the semi-supervised deep ladder network is also applied in [13] to identify the failure of the primary parallel shaft helical gear in an induction motor system. The ladder network is implemented by modeling hierarchical latent variables to integrate supervised and unsupervised learning strategies. However, the unsupervised components of the ladder network may not help in a semi-supervised environment if those raw data do not show obvious clustering on the 2-D manifold, which is usually the case for vibration signals. Although GAN has been also used for semi-supervised learning in [14], [17], [18], it is reported in [21] that good generators and good semi-supervised classifiers cannot be obtained simultaneously. Additionally, the well-known difficulty in training GANs has further impacted their applications to practical semi-supervised learning tasks [10].

The motivation of the proposed research is both broad and specific, as we strive to solve the problem of bearing fault diagnosis through a solid theoretical explanation, and leverage the fault features of both labeled and unlabeled data to make the classifier more accurate and robust. Therefore, we adopt a deep generative model based on solid Bayesian theory and use scalable variational inference in a semi-supervised environment. Although some existing work using variational autoencoders (VAE) for bearing fault diagnosis can be found in [22]–[24], they only use the discriminative features in the latent space for dimension reduction, and then use these features to train other external classifiers. In this work, however, we also take an integrated approach to train the VAE model itself as a classifier by also exploiting its generative capabilities.

This paper tackles both label scarcity and label accuracy issues in bearing fault diagnosis. Detailed technical contributions of this work are summarized as follows:

1) *Semi-supervised deep generative model implementation:* This paper applies two semi-supervised VAE-based deep generative models to leverage properties of both the labeled and unlabeled data for bearing fault diagnosis. To mitigate the "KL vanishing" problem in VAE models and further promote the accuracy and robustness of the semi-supervised classifier, this study also adapt the KL cost annealing techniques [25], [26] on top of the original model presented in [27].
2) *Strong performance mitigating the label scarcity issue:* This work utilizes the CWRU dataset to create test scenarios where only a small subset of data for each fault category has labels, which corresponds to the label scarcity issue discussed in [12]–[19] for real-world applications. The results show that the M2 model can greatly outperform the baseline unsupervised and supervised learning algorithms. Additionally, the VAE-based semi-supervised generative M2 model also compares favorably against four state-of-the-art semi-supervised learning methods.
3) *Solid performance mitigating the label accuracy issue:* This study also uses the IMS dataset with naturally-evolved bearing defects to create test scenarios with the label accuracy issue discussed in [17]. The results demonstrate that incorrect labeling will inevitably reduce the classifier performance of supervised learning algorithms, while adopting semi-supervised deep generative models can be an effective way to mitigate the label accuracy issue. This conclusion can be supported by M2 model's consistent dominance over CNN when a lot of healthy data were mislabeled as faulty ones.

The rest of the paper is organized as follows. In Section II, we introduce some of the background knowledge of VAE. Next, in Section III, we present the architecture of two VAE-based deep generative models in the semi-supervised setting, with detailed discussions on leveraging a dataset including both labeled and unlabeled data. In Section IV, two comparative studies of the proposed models against other popular machine learning and deep learning algorithms are performed using both the University of Cincinnati's Center for Intelligent Maintenance Systems (IMS) dataset [28] and the Case Western Reserve University (CWRU) bearing dataset [29]. Section V concludes the paper by highlighting its technical contributions.

II. Background of Variational Autoencoders

The variational inference technique is often used in the training and prediction process, which is effective for solving the posterior of the distribution obtained from neural networks [20]. The VAE's architecture is demonstrated in Fig. 1, which specifies a joint distribution $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ over observations $\mathbf{x}$ and latent variables $\mathbf{z}$, which are usually sampled from a prior density $p(\mathbf{z})$ subject to a multivariate unit Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. These latent variables are also related to the observed variables $\mathbf{x}$ through the likelihood $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, which can be regarded as a probabilistic decoder, or generator, to decode $\mathbf{z}$ into a distribution over the observation $\mathbf{x}$. A neural network parameterized by $\boldsymbol{\theta}$ will be used to model the decoder.

After specifying the decoding process, it is necessary to perform inference, or to calculate the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ of latent variables $\mathbf{z}$ given the observations $\mathbf{x}$. In addition, we also seek to optimize the model parameters $\boldsymbol{\theta}$ with respect to $p_{\boldsymbol{\theta}}(\mathbf{x})$, which is obtained by marginalizing out the latent variables $\mathbf{z}$ in the likelihood function $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$. Since the prior $p(\mathbf{z})$ is a Gaussian non-conjugate process, the true posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ becomes analytically intractable. Therefore, the technique of variational inference should be used to approximate a posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ with optimized variational parameters $\boldsymbol{\phi}$, which minimizes the Kullback-Leibler (KL) divergence of the approximated posterior to the true posterior. This posterior approximation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ can be also observed as an encoder with distribution $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi}(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_{\phi}(\mathbf{x})))$, of which $\boldsymbol{\mu}_{\phi}(\mathbf{x})$ and $\boldsymbol{\sigma}_{\phi}(\mathbf{x})$ will be also optimized using neural networks.

By definition, the KL divergence measures the similarity between two distributions, which is expressed as an expectation of the log of the first distribution minus the log of the second distribution. Thus the KL divergence of the approximated posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ with respect to the true posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ is shown Eqn. (1), after applying the Bayes' theorem.

After moving $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ to the left hand side of Eqn. (1), it can be written as the sum of a defined term known as the evidence lower bound (ELBO) and the KL divergence, which satisfies $D_{\mathrm{KL}}\left[q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z}|\mathbf{x})\right] \geq 0$. Specifically, based on Jensen's inequality, the optimal $q_{\phi}(\mathbf{z}|\mathbf{x})$ that maximizes the ELBO is $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$, which also simultaneously makes the KL divergence term equal to zero. Therefore, maximizing Eqn. (2) with respect to $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{\phi}$ is analogous to minimizing the KL divergence, and this optimization can be performed using stochastic gradient descent.

III. Semi-Supervised Deep Generative Models based on Variational Autoencoders

This section presents two semi-supervised deep generative models based on VAE [27]. When only a small subset of training data have labels, both models can exploit VAE's generative power to enhance the classifier's performance. By learning a good variational approximation of the posterior, the VAE's encoder can embed the input data $\mathbf{x}$ as a set of low-dimension latent features $\mathbf{z}$. The approximated posterior $q_{\phi}(\mathbf{z}|\mathbf{x})$ is formed by a nonlinear transformations, which can be modeled as a deep neural network $f(\mathbf{z}; \mathbf{x}, \boldsymbol{\phi})$ with variational parameters $\boldsymbol{\phi}$. Similarly, the VAE's generator takes a set of latent variables $\mathbf{z}$ and reproduces the observations $\mathbf{x}$ using $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, which can be also modeled as a deep neural network $g(\mathbf{x}; \mathbf{z}, \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$.

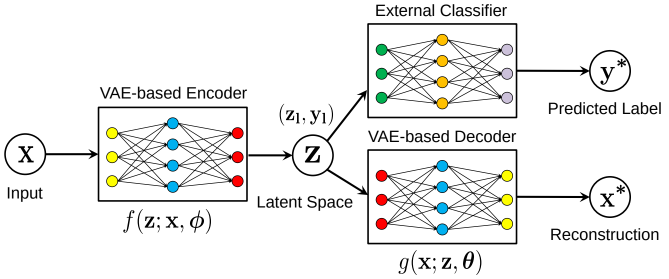


Fig. 2. Illustration of the latent-feature discriminative M1 model.

A. Latent-feature discriminative M1 model

The M1 model [27] trains the VAE-based encoder and decoder in an unsupervised manner. The trained encoder will provide an embedding of input data $\mathbf{x}$ in the latent space, which is defined by the latent variables $\mathbf{z}$. In most cases, the dimension of $\mathbf{z}$ is much smaller than that of $\mathbf{x}$, and these low-dimensional features can often increase the accuracy of supervised learning models.

As shown in Fig. 2, after training the M1 model, the actual classification task will be carried out in an external classifier, such as support vector machine (SVM), polynomial regression, etc. Specifically, the VAE encoder will only process the labeled data $\mathbf{x}_\mathbf{l}$ to determine their corresponding latent variable $\mathbf{z}_\mathbf{l}$, then they are combined with their corresponding labels y_l to train this external classifier. The M1 model is considered a semi-supervised method, since it leverages all available data to train the VAE-based encoder and decoder in an unsupervised manner, and thereafter it also takes the labeled data $(\mathbf{z}_\mathbf{l}, y_l)$ to train an external classifier in a supervised fashion. When compared with purely supervised learning methods that can only be trained using a small subset of data with labels, the M1 model usually promotes more accurate classification, since the VAE structure is also able to learn from the vast majority of unlabeled data, enabling the extraction of more representative latent features to train its subsequent classifier.

B. Semi-Supervised Generative M2 model

As briefly mentioned earlier, the major limitation of the M1 model is the disjoint nature of its training process, as it needs to train the VAE network first and thereafter the

$$\begin{aligned}
D_{\mathrm{KL}}\left[q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z}|\mathbf{x})\right] &= \mathbb{E}_{z\sim q_{\phi}(\mathbf{z}|\mathbf{x})}\left[\log q_{\phi}(\mathbf{z}|\mathbf{x}) - \log p(z|x)\right] \\
&= \mathbb{E}_{z\sim q_{\phi}(\mathbf{z}|\mathbf{x})}[\log q_{\phi}(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}) - \log p_{\theta}(\mathbf{x}|\mathbf{z})] + \log p_{\theta}(\mathbf{x}) \qquad (1)
\end{aligned}$$

$$\begin{aligned}
\log p_{\theta}(\mathbf{x}) &= -\mathbb{E}_{\mathbf{z}\sim q_{\phi}(\mathbf{z}|\mathbf{x})}\left[\log q_{\phi}(\mathbf{z}|\mathbf{x}) - \log p(z) - \log p_{\theta}(\mathbf{x}|\mathbf{z})\right] + D_{\mathrm{KL}}\left[q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z}|\mathbf{x})\right] \\
&= \underbrace{\mathbb{E}_{z\sim q_{\phi}(\mathbf{z}|\mathbf{x})}\left[\log p_{\theta}(\mathbf{z}|\mathbf{x}) + \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})\right]}_{\text{Evidence Lower Bound (ELBO)}} + \underbrace{D_{\mathrm{KL}}\left[q_{\phi}(\mathbf{z}|\mathbf{x}) \| p_{\theta}(\mathbf{z}|\mathbf{x})\right]}_{\geq 0} \qquad (2)
\end{aligned}$$

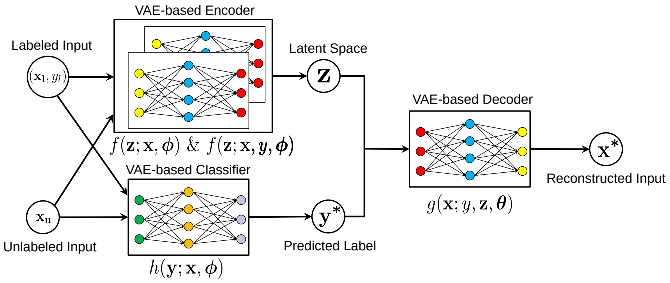


Fig. 3. Illustration of the semi-supervised generative M2 model.

external classifier. Specifically, the initial VAE training phase of the M1 model's VAE-based encoder and decoder is a purely unsupervised process and does not involve any scarce labels y_l, which is completely separated from the subsequent classifier training phase that actually takes y_l. To address this issue, another semi-supervised deep generative model, referred to as the M2 model, is also proposed in [27]. The M2 model can handle two situations at the same time: one where the data have labels, and the other where these labels are not available. Therefore, there are also two ways to construct the approximated posterior q and its variational objective.

1) Variational Objective with Unlabeled Data: When labels are not available, two separate posteriors $q_{\boldsymbol{\phi}}(y|\mathbf{x})$ and $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ will be approximated during the VAE training stage, where $\mathbf{z}$ is still the latent variables similar to the M1 model, while y is the unobserved label y_u. This newly defined posterior approximation $q_{\boldsymbol{\phi}}(y|\mathbf{x})$ will be used to construct the best classifier as our inference model [27]. Given the observations $\mathbf{x}$, the two approximated posteriors of the corresponding class labels y and latent variables $\mathbf{z}$ can be defined as

$$\begin{aligned} q_\phi(y|\mathbf{x}) &= \mathrm{Cat}\left(y|\pi_\phi(\mathbf{x})\right) \\ q_\phi(\mathbf{z}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{z}|\mu_\phi(\mathbf{x}), \mathrm{diag}\left(\sigma^2_\phi(\mathbf{x})\right)\right) \end{aligned} \tag{3}$$

where $\mathrm{Cat}\left(y|\pi_\phi(\mathbf{x})\right)$ is the concatenated multinomial distribution, $\pi_\phi(\mathbf{x})$ can be modeled by a neural network parameterized by ϕ. Combining the above two posteriors, a joint posterior approximation can be defined as

$$q_\phi(y, \mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}|\mathbf{x}) q_\phi(y|\mathbf{x}) \tag{4}$$

Therefore, the revised $\mathrm{ELBO_U}$ that determines the variational objective of the unlabeled data can be written as Eqn. (5), where $\mathcal{L}(\mathbf{x}, y)$ is the original ELBO in Eqn. (2).

2) Variational Objective with Labeled Data: Since the goal of semi-supervised learning is to train a classifier using a limited amount of labeled data and the vast majority of unlabeled data, it would be beneficial to also include the scarce labels in the training process of this deep generative M2 model. Similarly, Eqn. (6) shows the revised $\mathrm{ELBO_L}$ that determines the variational objective for the labeled data.

3) Combined Objective for the M2 Model: In Eqn. (6), the distribution $q_{\boldsymbol{\phi}}(y|\mathbf{x})$, which is used to construct the discriminative classifier, is only included in the variational objective of the unlabeled data. This is still an undesirable feature, since the labeled data will not be involved in learning this distribution or the variational parameter $\boldsymbol{\phi}$. Therefore, an additional loss term should be superimposed on the combined model objective, such that both the labeled and unlabeled data can contribute to the training process. Hence, the final objective of the semi-supervised deep generative M2 model is:

$$\mathcal{J}^\alpha = \sum_{\mathbf{x} \sim \tilde{p}_\mathbf{u}} \mathcal{U}(\mathbf{x}) + \sum_{(\mathbf{x}, y) \sim \tilde{p}_l} \left[\mathcal{L}(\mathbf{x}, y) - \alpha \cdot \log q_\phi(y|\mathbf{x})\right] \tag{7}$$

in which the hyper-parameter α controls the relative weight between the generative and the discriminative learning. A rule of thumb is to set α to be $\alpha = 0.1 \cdot N$ in all experiments, where N is the number of labeled data samples.

With this combined objective function, we can integrate a large number of $\mathbf{x}$ as a mini-batch to enhance the stability of training two neural networks used as an encoder and a decoder. Finally, we'll run stochastic gradient descent to update the model parameters $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{\phi}$. The structure of the M2 model is presented in Fig. 3.

C. Model Implementations

1) M1 Model Implementation: The M1 model constructs its encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and decoder $p_\theta(\mathbf{x}|\mathbf{z})$ by using two deep neural networks $f(\mathbf{z}; \mathbf{x}, \boldsymbol{\phi})$ and $g(\mathbf{x}; \mathbf{z}, \boldsymbol{\theta})$, respectively. The encoder has 2 convolutional layers and 1 fully connected layer using ReLu activation, aided by batch normalization and dropout layers. The decoder consists of 1 fully connected layer followed by 3 transpose convolutional layers, where the first 2 layers use ReLU activation and the last layer uses linear activation.

Due to the "KL vanishing" problem, it is often difficult to achieve a good balance between the likelihood and the KL divergence, as the KL loss can be undesirably reduced to zero, though it is expected to remain a small value. To overcome this problem, the implementation of M1 model uses the "KL

$$\begin{aligned} \mathrm{ELBO_U} &= \mathbb{E}_{q_\phi(y,\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{x}|y, \mathbf{z}) + \log p_\theta(y) + \log p_\theta(\mathbf{z}) - \log q_\phi(y, \mathbf{z}|\mathbf{x})\Big] \\ &= \mathbb{E}_{q_\phi(y|\mathbf{x})}\big[-\mathcal{L}(\mathbf{x}, y)) - \log q_\phi(y|\mathbf{x})\big] \\ &= \sum_y q_\phi(y|\mathbf{x})(-\mathcal{L}(\mathbf{x}, y)) + \mathcal{H}(q_\phi(y|\mathbf{x})) = -\mathcal{U}(x) \end{aligned} \tag{5}$$

$$\begin{aligned} \mathrm{ELBO_L} =& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},y)}\left[\log p_\theta(\mathbf{x}|y, \mathbf{z}) + \log p_\theta(y) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}, y)\right] \\ =& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|y, \mathbf{z}) + \log p_\theta(y) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] = -\mathcal{L}(\mathbf{x}, y) \end{aligned} \tag{6}$$

cost annealing" or "β VAE" [25], which includes a new weight factor β for the KL divergence. The revised ELBO function for "β VAE" is

$$\text{ELBO} = \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]}_{\text{Reconstruction}} - \beta \cdot \underbrace{D_{KL}\left(q_\phi(z|x) \| p(z)\right)}_{\text{KL Regularization}} \quad (8)$$

During training, β will be manipulated to gradually increase from 0 to 1. When $\beta < 1$, the latent variables $\mathbf{z}$ are trained with an emphasis on capturing useful features for reconstructing the observations $\mathbf{x}$. When $\beta = 1$, the $\mathbf{z}$ learned in earlier epochs can be taken as a good initialization, which enables more informative latent features to be used by the decoder [26].

After training the M1 model that is able to balance its reconstruction and generation features, the latent variable $\mathbf{z}$ in latent space will be used as discriminative features for the external classifier. This paper uses an SVM classifier, though any personally preferred classifier can be also used. The advantage of the M1 model is that the discriminative feature extraction function of VAE can be used to reduce the dimensionality of the input data to a lower value. In this study, the input data has a dimension of 1,024, which will be reduced to 128 in the latent space.

2) M2 Model Implementation: The deep generative M2 model uses the same structure for $q_\phi(\mathbf{z}|\mathbf{x})$ as the M1 model, while the decoder $p_\theta(\mathbf{x}|y, \mathbf{z})$ also has the same settings as M1's $p_\theta(\mathbf{x}|\mathbf{z})$. In addition, the classifier $q_\phi(y|\mathbf{x})$ is comprised of 2 convolutional layers and 2 max-pooling layers with dropout and ReLU activation, followed by the final Softmax layer.

Two independent neural networks are used, one for labeled data and one for unlabeled data, with the same network structure, but different input/output specifications and loss functions. For instance, for labeled data, both $\mathbf{x_l}$ and y are considered as input to minimize the labeled $(\mathbf{x}, y) \sim \tilde{p}_l$ part in Eqn. (7), and the output will be the reconstructed as $\mathbf{x_l^*}$ and y^*. For unlabeled data, $\mathbf{x_u}$ is the only input to reconstruct $\mathbf{x'_u}$. Other hyper-parameters of the M2 model are also selected empirically. We use a batch size of 200 for training, the latent variable $\mathbf{z}$ has a dimension of 128. For optimizer settings, we use RMSprop with a 10^{-4} initial learning rate.

3) M1 vs. M2 Model: By comparing the M1 and M2 models, it's obvious to find that the significance of the M1 model lies in its simpler and clear network structure, which is easy to implement and saves training time. As shown in Fig. 2, the M1 model is a simple and straightforward implementation of VAE that only includes an encoder and a decoder trained in an unsupervised manner, then the learned latent features and labels $(\mathbf{z_l}, y_l)$ of the labeled data are subsequently used to train an external classifier.

On the other hand, M2 deals with both labeled and unlabeled data by using two identical encoder networks for both labeled and unlabeled data. Additionally, it also has a built-in classifier to perform inference on the approximated posterior $q_\phi(y|\mathbf{x})$. Therefore, despite the fact that the M2 model tends to have a superior performance than the M1 model, it also suffers from increased model complexity and prolonged training time. Since both of them have their strengths and weaknesses, it is worthwhile to compare how they perform in the context of semi-supervised bearing.

TABLE I
CLASS LABELS SELECTED FROM THE CWRU DATASET

Class label	Fault location			Fault diameter (mils)		
	Ball	IR	OR	0.007	0.014	0.021
1	✓	–	–	✓	–	–
2	✓	–	–	–	✓	–
3	✓	–	–	–	–	✓
4	–	✓	–	✓	–	–
5	–	✓	–	–	✓	–
6	–	✓	–	–	–	✓
7	–	–	✓	✓	–	–
8	–	–	✓	–	✓	–
9	–	–	✓	–	–	✓
10	Normal					

IV. EXPERIMENTAL RESULTS USING THE CWRU DATASET

In this section, we seek to use the CWRU dataset to verify the effectiveness of the two VAE-based semi-supervised deep generative models for bearing fault diagnosis. The developed diagnostic framework will be described in detail, and the performance of the classifier will be first compared with three baseline supervised/unsupervised algorithms, including principal component analysis (PCA), autoencoder (AE), and convolutional neural network (CNN). Then, we'll also compare the proposed methods against some state-of-the-art semi-supervised learning algorithms, such as low density separation (LDS) [30], safe semi-supervised support vector machine (S4VM) [31], SemiBoost [32], and semi-supervised smooth alpha layering (S3AL) [12].

A. CWRU Dataset

The CWRU dataset contains the vibration signals collected from the drive-end bearing and fan-end bearing in a 2 hp induction motor dyno setup [29]. Single-point defects are manually created onto the bearing inner race (IR), outer race (OR), and rolling elements by electro-discharge machining. Different defect diameters of 7 mil, 14 mil, 21 mil, 28 mil, and 40 mil were used to simulate different levels of fault severity. Two accelerometers mounted on the drive-end and fan-end of the motor housing were used to collect vibration data at a motor load of 0 to 3 hp and a motor speed from 1,720 to 1,797 rpm at a sampling frequency of 12 kHz or 48 kHz.

The purpose of the proposed bearing diagnostic model is to reveal the location and severity of bearing defects, vibration data collected for the same failure type but at different speeds and load conditions will be considered to have the same class label. Based on this standard, 10 classes are specified according to the size and location of the bearing defect, and TABLE I identifies a detailed list of all 10 classes featured in this study.

B. Data Preprocessing

The diagnosis process starts from data segmentation, which divides the collected vibration signal into multiple segments of equal length. For the CWRU dataset, the number of data samples of the drive-end vibration signal for each kind of bearing failure is approximately 120,000 at three different

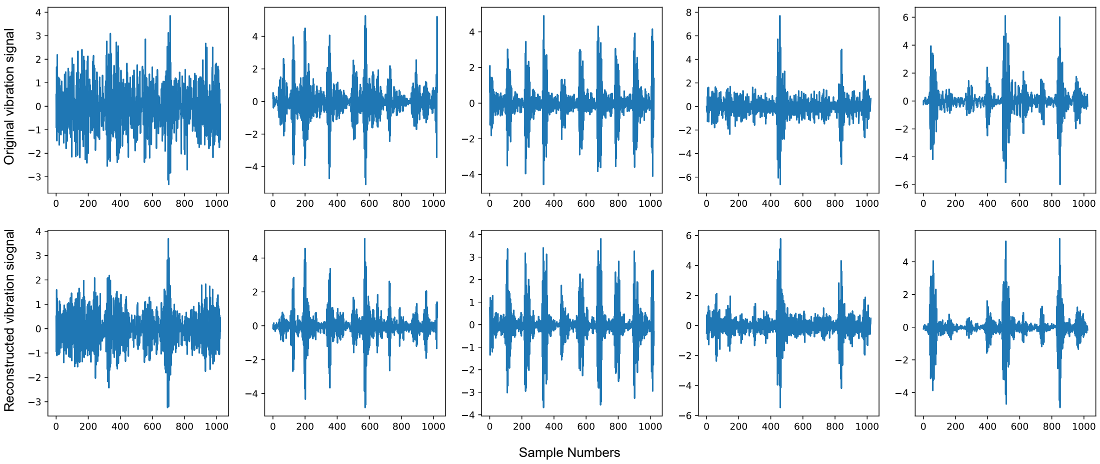


Fig. 4. Comparison of the original (top row) and the reconstructed (bottom row) bearing vibration signals after training the VAE M1 model.

speeds (i.e., 1,730 rpm, 1,750 rpm, and 1,772 rpm). Data collected at these speeds constitute the complete data for each class, which will be later segmented in a fixed window size of 1,024 samples (or 85.3 ms time span for a sampling rate of 12 kHz) and with a sliding rate of 0.2. Finally, the number of training and test data segments is 12,900 and 900, respectively.

All test data will be labeled. Although the percentage of test data appears to be small at first glance – approximately 7%, only a maximum of 2,150 training data segments will have labels in the later experiment, indicating the percentage of test data to labeled training data is still around 30%.

After the initial data import and segmentation stage, these data segments are still arranged in the order of their classes labels (fault types). Therefore, data shuffling needs to be carried out to ensure that both the training set and the test set can represent the overall distribution of the CWRU dataset, which enhances the model generalization and makes it less prone to overfitting. Classical standardization techniques are also implemented to the training and test set to ensure the vibration data have zero mean and unit variance, which is enabled by subtracting the mean of the original data and then dividing the result by the original standard deviation.

C. Experimental Results

After training the VAE-based generative M1 model, the reconstructed bearing vibration signal should be very similar to the actual vibration signal, and their comparisons are demonstrated in Fig. 4. Although a perfect reconstruction may impact the VAE's generative capabilities and reduce its versatility, a reasonably close reconstruction with a small error indicates that VAE has achieved a balance between reconstruction and generation, which is critical to leverage the generative features of the algorithm.

The network structure of the VAE-based deep generative M1 and M2 models have been discussed in detail in Section III. C. In addition to implementing these models to perform bearing fault diagnosis, other popular unsupervised learning schemes such as PCA and autoencoder, as well as the supervised CNN, are also trained to serve as baselines. Their parameters are either selected to be consistent with the M1 and M2 model, or obtained through parameter tuning. For example, we use the same optimizer settings as the VAE model (RMSprop with an initial learning rate of 10^{-4}) to train both CNN and autoencoder benchmarks. More details are provided as follows:

1) PCA+SVM: the PCA+SVM benchmark is trained using low-dimensional features extracted from the labeled data segments (each consists of 1,024 data samples) using PCA. The dimension of the feature space is 128, which is consistent with the M1 and M2 model's latent space dimension. The SVM uses a radial basis function (RBF) kernel, and through modest parameter optimization, its regularization parameter is set to $C = 10$. Additionally, the kernel coefficient is set to "sample" $(1/128/X.var())$, where $X.var()$ is the variance of the input data.
2) Autoencoder (AE): Thanks to the structural similarity with VAE, the AE baseline inherits the same network structure (encoder-decoder) as the M1 model, as well as its SVM-based external classifier.
3) CNN: the CNN benchmark treats each time-series vibration data segment (consisting of 1,024 data samples) as a 2-D 32x32 image, which is a common practice to apply the vanilla CNN on bearing fault diagnosis, as discussed in detail in [33]. Specifically, the CNN baseline has two convolution layers with ReLU activation, each with 2×2 convolutions and 32 filters, followed by a 2×2 max-pooling layer and a 0.25 dropout layer. Next, we have a fully connected hidden layer with a dimensionality of 512, and the output of which is fed into a softmax layer. The cross-entropy loss is adopted, and the batch size is set to 10, which is also obtained via parameter tuning.

A total of 10 rounds of experiments are performed on the same training and test sets shuffled randomly. Of the 129,000 training samples, only a small portion of the labels are actually used in different algorithms to construct bearing fault classifiers. Seven case studies were conducted using 50, 100, 300, 516, 860, 1,075, and 2,150 labels, representing 0.39%, 0.78%, 2.34%, 4%, 6.67%, 8.33%, and 16.67% of the training data have labels.

Tab. II lists the average accuracy and standard deviation of different algorithms after 10 rounds of experiments, in which the latent feature discriminant model (M1) performs

TABLE II
EXPERIMENTAL RESULTS OF VAE-BASED SEMI-SUPERVISED CLASSIFICATION ON CWRU BEARING DATASET WITH LIMITED LABELS

N	Labeled Data %	PCA+SVM	Autoencoder	CNN	VAE M1	**VAE M2**
50	0.39%	26.57 ± 0.38%	33.40 ± 1.24%	30.74 ± 2.88%	32.93 ± 2.01%	**40.57 ± 2.91%**
100	0.78%	33.77 ± 2.81%	37.96 ± 0.65%	34.50 ± 2.16%	36.91 ± 1.37%	**60.04 ± 3.57%**
300	2.34%	44.43 ± 1.59%	44.06 ± 2.61%	60.28 ± 3.42%	47.03 ± 1.22%	**87.63 ± 2.80%**
516	4%	53.27 ± 2.21%	50.88 ± 2.03%	75.41 ± 2.74%	57.06 ± 1.76%	**94.16 ± 1.66%**
860	6.67%	61.70 ± 2.31%	58.89 ± 1.81%	87.39 ± 0.93%	67.19 ± 1.70%	**96.77 ± 0.38%**
1075	8.33%	67.83 ± 1.27%	62.83 ± 1.61%	91.07 ± 1.46%	71.97 ± 1.40%	**97.86 ± 0.51%**
2150	16.67%	82.40 ± 1.47%	77.09 ± 0.98%	97.19 ± 0.99%	86.59 ± 1.43%	**98.06 ± 0.88%**

TABLE III
COMPARISON OF DIFFERENT SEMI-SUPERVISED LEARNING ALGORITHMS WITH DIFFERENT LABELED DATA PERCENTAGE ν

Algorithms	$\nu = 5\%$	$\nu = 10\%$	$\nu = 20\%$	Overall	Rank
LDS [30]	69.60 ± 15.43%	74.49 ± 13.72%	77.90 ± 12.98%	74.21 ± 17.77%	6
S4VM [31]	70.85 ± 12.54%	87.52 ± 12.48%	92.44 ± 8.18%	83.60 ± 11.59%	5
SemiBoost [32]	79.32 ± 18.16%	85.59 ± 14.63%	90.76 ± 10.25%	85.22 ± 14.56%	4
S3AL [12]	85.60 ± 14.48%	90.47 ± 10.73%	94.25 ± 6.96%	90.10 ± 11.68%	2
VAE M1	78.75 ± 7.75%	88.53 ± 8.37%	92.40 ± 7.08%	86.56 ± 9.52%	3
VAE M2	**87.17 ± 7.18%**	**97.82 ± 4.63%**	**99.80 ± 0.47%**	**94.93 ± 7.40%**	1

TABLE IV
DATA CHARACTERISTICS OF EACH SCENARIO BASED ON [12]

Scenarios	Signal length	Defect width		Motor load	
		0.007 *in*	0.014 *in*	0 *hp*	1 *hp*
SCN1	1024 (100)	✓	–	✓	–
SCN2	1024 (100)	✓	–	–	✓
SCN3	1024 (100)	–	✓	✓	–
SCN4	1024 (100)	–	✓	–	✓

as good as the unsupervised model (autoencoder), if not better, showcasing that the M1 model's latent space is able to provide robust features to enable good classification. It is also worth mentioning that, initially, the VAE-based M1 model has advantages over CNN until the number of labeled samples is $N = 100$. Then the performance is degraded, which contradicts the results obtained on the classic MNIST dataset in [27]. One explanation for this deviation may be that the CWRU dataset has many explicit feature representations, which can be easily captured by some powerful supervised learning schemes. Therefore, the CNN only requires 1,000 labeled training data segments to achieve an accuracy over 90%.

However, by integrating the features learned in the M1 model and the classification mechanism directly into the same model, the conditional generated M2 model can obtain better results than the CNN or M1 model using an external SVM classifier. Specifically, to achieve a fault classification accuracy of about 95%, the M2 model only needs 4% (516) of the training data segments to be labeled, while the best accuracy that can be achieved using other benchmark algorithms is only 75.41% using the same amount of labeled data. In addition, this significant accuracy improvement of around 20% is also very consistent, since its standard deviation is as low as 1.66%.

Additionally, we also compare the proposed VAE-based M1 and M2 model against four state-of-the-art semi-supervised learning methods, including LDS, S4VM, SemiBoost, and S3AL. To make a fair comparison, we used the same CWRU dataset, the same data preprocessing methods in terms of data segmentation (signal length and number) and labeling (based on defect width and motor speed/load), and the same labeled data percentage as [12]. The fault classification accuracy and standard deviation obtained using the CWRU dataset is shown in TABLE III, where ν stands for the percentage of labeled samples. The results for the four semi-supervised benchmark studies were summarized in [12], and the data segmentation and labeling details of which are presented in TABLE IV.

It can be observed from Table III that the best-performing algorithm is the proposed VAE M2 method. Specifically, the average accuracy of VAE M2 is 1.5%, 7.4%, and 5.6% better than the second-best S3AL algorithm, when the percentage of labeled data ν is 5%, 10%, and 20%, respectively. In addition, the standard deviation of VAE M2 is also significantly lower than the benchmark semi-supervised learning studies, demonstrating the robustness and consistency of the VAE M2 model. Moreover, the VAE M1 model also secures third place in this comparison, and its performance is just 1% to 2% shy of S3AL when ν is 10% or 20%. Therefore, it can be concluded that it is not only the adoption of VAE-based networks, but also the integrated training approach of the M2 model that contributed to the largest performance enhancement of deep generative models in semi-supervised bearing fault diagnosis.

V. EXPERIMENTAL RESULTS USING THE IMS DATASET

A. IMS Dataset

In the previous CWRU dataset, bearing damage is artificially initiated in order to accelerate the degradation process. Therefore, the IMS bearing dataset, which contains data collected from naturally evolved bearing defects, is also used in this study to evaluate the performance of the VAE-based generative models. The IMS dataset was collected on the test stand shown in Fig. 5. Specifically, four double-row bearings are installed

TABLE V

BEARING FAULT SCENARIOS AND THEIR DEGRADATION STARTING POINTS IN THE IMS DATASET [34]

Bearing	Subset 1 Bearing 3	Subset 1 Bearing 4	Subset 2 Bearing 1	Subset 3 Bearing 3
Fault type	Inner race	Rolling element	Outer race	Outer race
Endurance duration	34 days 12 h	34 days 12 h	6 days 20 h	45 days 9 h
Number of files	2,156	2,156	984	4,448
Degradation starting point AEC*	2,027	1,641	547	2367
Degradation starting point MAS-Kurtosis†	1,910	1,650	710	N/A
Degradation starting point HMM-DPCA‡	2,120	1,760	539	N/A

*AEC: auto-encoder-correlation-based (AEC) prognostic algorithm.

†MAS-Kurtosis: moving average spectral kurtosis.

‡HMM-DPCA: hidden Markov model with dynamic PCA.

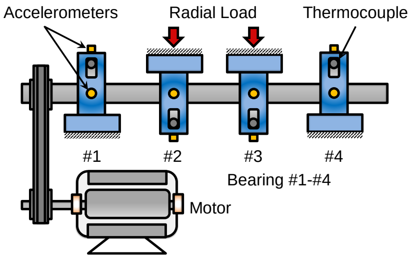


Fig. 5. Experimental setup collecting the IMS dataset [28].

on the same shaft, which is coupled to the motor shaft through a rubber band. Each of the four bearings tested carries 6,000 lbs of radial load. The test can continue even if these bearings show early implications of failure, since they are not used to support the motor or any active rotational motion. In fact, this test will not be terminated until the accumulation of debris on the electromagnetic plug exceeds a certain level of threshold [28]. This is a major distinction when compared to the actual situation where the bearing supports the motor shaft and transmission, since the test needs to be quickly stopped after sensing abnormal conditions.

The IMS dataset consists of 3 subsets that were collected when the motor was running at a constant speed of 2,000 rpm. For subset 1, two high-sensitivity PCB 253B33 Quart ICP accelerometers are installed to measure bearing vibrations in both x and y directions, while subsets 2 and 3 only have one accelerometer installed on each bearing. Data are acquired in 1-second windows and are stored in a separate file every ten minutes. Each file contains 20,480 sampling points, except for the first subset, which collects the first 92 files every five minutes. As mentioned earlier, the IMS test may continue after the bearing is degraded, and it is challenging to label when such a bearing degradation actually happened. In [34], three different algorithms are applied to estimate the degradation starting point, the results demonstrate a high level of uncertainty as the estimated starting points can deviate by more than 100 data segments using different methods. A detailed summary of this finding in [34] is listed in TABLE V.

B. Data Preprocessing

The IMS dataset uses a fixed window size of 1,024 for segmentation and collects data at a frequency of 20.48 kHz. Due to the large amount of noise in the “Subset 3 Bearing 3” condition reported in [34], we only select the first 3 fault conditions in TABLE V to evaluate the performance of semi-supervised VAE models with label uncertainty. In addition, 210 consecutive files are selected for each fault condition. and the last 15 of them are chosen after their degradation starting points determined by the auto-encoder-correlation (AEC) algorithm. For instance, data files 1,832 to 2,042 will be selected for the “Subset 1 Bearing 3” scenario, since its estimated degradation starting point is 2,027. On the other hand, healthy data are picked from the first 110 files of “Subset 1 Bearing 3”.

Each fault scenario has 210 consecutive vibration data files, of which the last 10 files will serve as the test set, and the first 200 files will constitute the entire training set. Each file contains 20,480 data points, which can be divided into 20 data segments. Therefore, each fault scenario has 4,000 data segments, or all 4 categories (healthy, rolling elements, outer race, inner race) have 16,000 data segments.

In order to simulate the challenges related to accurate data labeling in practical applications, labels will be assigned starting from the last of the 40,000 training data segments for each fault scenario and proceed backward. For these data segments, we should have the highest confidence in the accuracy of their labels. Then, by labeling more preceding files, but with lower confidence, more case studies can be performed. The purpose is to investigate whether incorrect labeling will negatively impact the accuracy of the supervised learning benchmark – CNN, and to assess if semi-supervised deep generative models can still improve the accuracy of the bearing fault classifier by leaving these data segments unlabeled.

C. Experimental Results

A total of 10 rounds of independent semi-supervised experiments were performed using the IMS dataset, and 10 case studies are conducted by labeling the last 40, 100, 200, 400, 800, 1,000, 2,000, 4,000, and 8,000 data segments of the training set, which accounts for 0.25%, 0.63%, 1.25%, 2.5%, 5%, 6.25%, 12.5%, 25%, and 50% of the training data, respectively.

TABLE VI presents the classification results after 10 rounds of experiments. The performance of the M1 is better than that of PCA, but it has almost the same performance as the vanilla autoencoder. This shows that the VAE’s discriminant feature space has no obvious advantage over the vanilla autoencoder’s encoded space. Nevertheless, the performance of the M1 model

TABLE VI
EXPERIMENTAL RESULTS OF SEMI-SUPERVISED CLASSIFICATION ON IMS BEARING DATASET WITH LIMITED LABELS

N	Labeled Data %	PCA+SVM	Autoencoder	CNN	VAE M1	VAE M2
40	0.25%	$61.60 \pm 0.63\%$	$64.54 \pm 2.07\%$	$59.08 \pm 2.68\%$	$\mathbf{72.01 \pm 1.91\%}$	$66.27 \pm 8.31\%$
100	0.63%	$67.22 \pm 1.15\%$	$67.07 \pm 1.49\%$	$62.93 \pm 4.10\%$	$\mathbf{76.61 \pm 1.48\%}$	$71.15 \pm 6.24\%$
200	1.25%	$70.31 \pm 0.49\%$	$73.42 \pm 0.94\%$	$68.64 \pm 5.40\%$	$\mathbf{78.74 \pm 1.25\%}$	$76.54 \pm 3.58\%$
400	2.50%	$75.38 \pm 0.90\%$	$78.42 \pm 1.17\%$	$74.20 \pm 3.17\%$	$81.66 \pm 1.02\%$	$\mathbf{82.78 \pm 2.21\%}$
800	5.00%	$77.85 \pm 0.62\%$	$84.81 \pm 0.78\%$	$78.73 \pm 2.98\%$	$85.03 \pm 1.15\%$	$\mathbf{88.45 \pm 1.71\%}$
1000	6.25%	$78.19 \pm 0.59\%$	$85.83 \pm 0.85\%$	$81.29 \pm 4.18\%$	$86.61 \pm 1.27\%$	$\mathbf{89.66 \pm 1.54\%}$
2000	12.50%	$78.50 \pm 0.30\%$	$86.61 \pm 0.77\%$	$86.62 \pm 4.11\%$	$87.20 \pm 1.18\%$	$\mathbf{90.87 \pm 1.97\%}$
4000	25.00%	$78.96 \pm 0.72\%$	$83.72 \pm 0.89\%$	$87.74 \pm 0.54\%$	$85.14 \pm 0.96\%$	$\mathbf{92.01 \pm 0.92\%}$
8000	50.00%	$79.06 \pm 0.65\%$	$84.00 \pm 1.21\%$	$81.56 \pm 2.79\%$	$85.36 \pm 1.17\%$	$\mathbf{88.11 \pm 3.47\%}$

is also superior to the supervised learning algorithm CNN. By incorporating the vast majority of unlabeled data in the training process, the improvement is approximately 5% to 15% when the number of labeled data segments varies from $N = 40$ to $N = 1,000$, and the standard deviation is also much smaller.

Similar to the comparison results shown in TABLE II, the classifier's performance of the VAE-based M2 model is superior to the other four algorithms, showing the advantage of integrating the training process of the VAE model and its built-in classifier. Critical observations can be drawn when the number of labeled training data increases from $N = 4,000$ to $N = 8,000$, the accuracy of the supervised algorithm CNN is reduced by more than 6%, and the loss of the semi-supervised VAE M2 model is 4%. The performance of unsupervised learning algorithms, which do not use labels in their training process, remains intact. This can be largely attributed to many healthy data are incorrectly labeled as faulty data, which also creates a dilemma that impairs the classifier's accuracy using either insufficient data or more data with inaccurate labels. Specifically, the best attainable accuracy for three baselines algorithms are 87.74% when $N = 4,000$ and 84% when $N = 8,000$, while the VAE-based M2 model can achieve an average of 92.01% and 88.11%, respectively.

In summary, the experimental results obtained using the IMS dataset consistently supports the previous findings on the CWRU dataset, that is, taking advantage of the large amount of unlabeled data can effectively enhance the classifier's performance using semi-supervised VAE-based deep generative models, especially the M2 model. In addition, the results also imply that inaccurate labeling can reduce the accuracy of supervised learning algorithms. Therefore, in diagnosing naturally evolved bearing faults in real-world applications, it is desirable to leverage semi-supervised learning methods, which only requires a small set of data that we can label with confidence while retaining the majority of data unlabeled.

VI. Conclusion

This paper implemented two semi-supervised deep generative models based on VAE for bearing fault diagnosis with limited labels. The results show that the M2 model can greatly outperform the baseline unsupervised and supervised learning algorithms, and this advantage can be up to 27% when only 2.3% of training data have labels. Additionally, the VAE-based M2 model also compares favorably against four state-of-the-art semi-supervised learning methods in terms of identifying bearing faults using data with limited labels.

The CWRU dataset only contains vibration data from manually initiated bearing defects, which is inconsistent with the real-world scenario where these defects are evolved naturally over time. Therefore, we also used the IMS dataset to verify the performance of the two VAE-based semi-supervised deep generative models. The results demonstrate that incorrect labeling will reduce the classifier performance of mainstream supervised learning algorithms, while adopting semi-supervised deep generative models and keeping data with label uncertainties unlabeled can be an effective way to mitigate this issue.

References

[1] M. Burgess, "What is the internet of things? wired explains," [Online]. Available: https://www.wired.co.uk/article/internet-of-things-what-is-explained-iot, 2018, Accessed: May. 2020.

[2] J. Krakauer, "Data in action: Iot and the smart bearing," [Online]. Available: https://blogs.oracle.com/bigdata/data-in-action-iot-and-the-smart-bearing, 2018, Accessed: Apr. 2020.

[3] J. Harmouche, C. Delpha, and D. Diallo, "Improved fault diagnosis of ball bearings based on the global spectrum of vibration signals," *IEEE Trans. Energy Convers.*, vol. 30, no. 1, pp. 376–383, March 2015.

[4] F. Immovilli, A. Bellini, R. Rubini, and C. Tassoni, "Diagnosis of bearing faults in induction machines by vibration or current signals: A critical comparison," *IEEE Trans. Ind. Appl.*, vol. 46, no. 4, pp. 1350–1359, July 2010.

[5] M. Kang, J. Kim, and J. Kim, "An FPGA-based multicore system for real-time bearing fault diagnosis using ultrasampling rate AE signals," *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2319–2329, April 2015.

[6] A. Ming, W. Zhang, Z. Qin, and F. Chu, "Dual-impulse response model for the acoustic emission produced by a spall and the size evaluation in rolling element bearings," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6606–6615, Oct 2015.

[7] M. Blodt, P. Granjon, B. Raison, and G. Rostaing, "Models for bearing damage detection in induction motors using stator current monitoring," *IEEE Trans. Ind. Electron.*, vol. 55, no. 4, pp. 1813–1822, April 2008.

[8] S. Zhang, B. Wang, M. Kanemaru, C. Lin, D. Liu, M. Miyoshi, K. H. Teo, and T. G. Habetler, "Model-based analysis and quantification of bearing faults in induction machines," *IEEE Trans. Ind. Appl.*, vol. PP, no. PP, pp. PP–PP, March 2020.

[9] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, 2019.

[10] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler, "Deep learning algorithms for bearing fault diagnostics–A comprehensive review," *IEEE Access*, vol. 8, pp. 29 857–29 881, 2020.

[11] S. R. Saufi, Z. Ahmad, M. S. Leong, and M. H. Lim, "Challenges and opportunities of deep learning models for machinery fault detection and diagnosis: A review," *IEEE Access*, vol. 7, pp. 122 644–122 662, 2019.

[12] R. R-Far, E. Hallaji, M. F-Zanjani, and M. Saif, "A semi-supervised diagnostic framework based on the surface estimation of faulty distributions," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1277–1286, March 2019.

[13] R. R-Far, E. Hallaji, M. F-Zanjani, M. Saif, S. H. Kia, H. Henao, and G. Capolino, "Information fusion and semi-supervised deep learning scheme for diagnosing gear faults in induction machine systems," *IEEE Trans. Ind. Electron.*, vol. 66, no. 8, pp. 6331–6342, Aug 2019.

[14] P. Liang, C. Deng, J. Wu, Z. Yang, J. Zhu, and Z. Zhang, "Single and simultaneous fault diagnosis of gearbox via a semi-supervised and high-accuracy adversarial learning framework," *Knowledge-Based Syst.*, vol. 198, p. 105895, 2020.

[15] X. Chen, Z. Wang, Z. Zhang, L. Jia, and Y. Qin, "A semi-supervised approach to bearing fault diagnosis under variable conditions towards imbalanced unlabeled data," *Sensors*, vol. 18, no. 7, p. 2097, 2018.

[16] M. Zhao, B. Li, J. Qi, and Y. Ding, "Semi-supervised classification for rolling fault diagnosis via robust sparse and low-rank model," in *Proc. IEEE Int. Conf. Ind. Inf. (INDIN)*, July 2017, pp. 1062–1067.

[17] D. B. Verstraete, E. L. Droguett, V. Meruane, M. Modarres, and A. Ferrada, "Deep semi-supervised generative adversarial fault diagnostics of rolling element bearings," *Structural Health Monitoring*, pp. 1–22, 2019.

[18] T. Pan, J. Chen, J. Xie, Y. Chang, and Z. Zhou, "Intelligent fault identification for industrial automation system via multi-scale convolutional generative adversarial network with partially labeled samples," *ISA Trans.*, vol. 101, pp. 379 – 389, 2020.

[19] C. Liu and K. Gryllias, "A semi-supervised support vector data description-based fault detection method for rolling element bearings based on cyclic spectral analysis," *Mech. Syst. Signal Process.*, vol. 140, p. 106682, 2020.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov, "Good semi-supervised learning that requires a bad GAN," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6510–6520.

[22] G. San Martin, E. López Droguett, V. Meruane, and M. das Chagas Moura, "Deep variational auto-encoders: A promising tool for dimensionality reduction and ball bearing elements fault diagnosis," *Structural Health Monitoring*, pp. 1092–1128, 2018.

[23] A. L. Ellefsen, E. Bjørlykhaug, V. Æsøy, and H. Zhang, "An unsupervised reconstruction-based fault detection algorithm for maritime components," *IEEE Access*, vol. 7, pp. 16 101–16 109, 2019.

[24] M. Hemmer, A. Klausen, H. V. Khang, K. G. Robbersmyr, and T. I. Waag, "Health indicator for low-speed axial bearings using variational autoencoders," *IEEE Access*, vol. 8, pp. 35 842–35 852, 2020.

[25] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," *arXiv preprint arXiv:1804.03599*, 2018.

[26] X. Liu, J. Gao, A. Celikyilmaz, L. Carin *et al.*, "Cyclical annealing schedule: A simple approach to mitigating KL vanishing," *arXiv preprint arXiv:1903.10145*, 2019.

[27] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 3581–3589.

[28] J. Lee, H. Qiu, G. Yu, and J. Lin, "Rexnord technical services," *Bearing Data Set, IMS, University of Cincinnati, NASA Ames Prognostics Data Repository*, 2007.

[29] "Case Western Reserve University (CWRU) bearing data center," [Online]. Available: https://csegroups.case.edu/bearingdatacenter/pages/project-history, 2005, accessed: Nov. 2020.

[30] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation." in *AISTATS*, vol. 2005. Citeseer, 2005, pp. 57–64.

[31] Y.-F. Li and Z.-H. Zhou, "Towards making unlabeled data never hurt," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 175–188, 2014.

[32] P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu, "Semiboost: Boosting for semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2000–2014, 2008.

[33] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, 2018.

[34] R. M. Hasani, G. Wang, and R. Grosu, "An automated auto-encoder correlation-based health-monitoring and prognostic method for machine bearings," *arXiv preprint arXiv:1703.06272*, 2017.

Shen Zhang (S'13-M'19) received the B.S. degree with the highest distinction in electrical engineering from Harbin Institute of Technology, Harbin, China, in 2014, and the M.S. and Ph.D. degrees in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2017 and 2019.

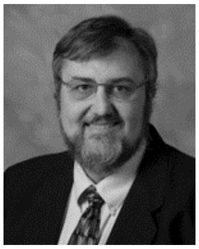
His research interests include design, control, condition monitoring, and fault diagnostics of electric machines, control of power electronics, powertrain engineering in electric vehicles, deep learning and reinforcement learning applied to energy systems.

Fei Ye (S'16-M'19) received the M.S. degree in electrical and computer engineering from Northeastern University, Boston, MA, USA, in 2014, and the Ph.D. degree in intelligent vehicles and transportation systems from University of California at Riverside.

Her research interests include connected and autonomous vehicles, intelligent vehicle trajectory planning, spatial and temporal data mining machine learning, and its application in vehicles and transportation.

Bingnan Wang (M'12-SM'15) received his B.S. degree from Fudan University, Shanghai, China, in 2003, and Ph.D. degree from Iowa State University, Ames, IA, USA, in 2009, both in Physics. He has been with Mitsubishi Electric Research Laboratories (MERL), located in Cambridge, Massachusetts since then, and is now a Senior Principal Research Scientist.

His research interests include electromagnetics and photonics, and their applications to wireless communications, wireless power transfer, sensing, electric machines, and energy systems.



Thomas G. Habetler (S'82-M'83-SM'92-F'02) received the B.S.E.E. degree in 1981 and the M.S. degree in 1984, both in electrical engineering, from Marquette University, Milwaukee, WI, and the Ph.D. degree from the University of Wisconsin-Madison, in 1989.

From 1983 to 1985 he was employed by the Electro-Motive Division of General Motors as a Project Engineer. Since 1989, he was with the Georgia Institute of Technology, Atlanta, where he is currently a Professor of Electrical and Computer Engineering. His research interests are in electric machine protection and condition monitoring, switching converter technology, and drives. He has published over 300 technical papers in the field. He is a regular consultant to industry in the field of condition-based diagnostics for electrical systems.

Dr. Habetler was the inaugural recipient of the IEEE-PELS "Diagnostics Achievement Award," and a recipient of the EPE-PEMC "Outstanding Achievement Award." the 2012 IEEE Power Electronics Society Harry A. Owen Distinguished Service Award, the 2012 IEEE Industry Application Society Gerald Kliman Innovator Award. He has also received one Transactions and four conference prize paper awards from the Industry Applications Society. He has served on the IEEE Board of Directors as the Division II Director, and on the Technical Activities Board, the Member and Geographic Activities Board, as a Director of IEEE-USA, and is a past president of the Power Electronics Society. He has also served as an Associate Editor for the IEEE TRANSACTIONS ON POWER ELECTRONICS.