Semi-Supervised Learning of Bearing Anomaly Detection via Deep Variational Autoencoders

Zhang, S.; Ye, F.; Wang, B.; Habetler, T.G.

Abstract

Most of the data-driven approaches applied to bearing fault diagnosis up to date are established in the supervised learning paradigm, which usually requires a large set of labeled data collected a priori. In practical applications, however, obtaining accurate labels based on real-time bearing conditions can be far more challenging than simply collecting a huge amount of unlabeled data using various sensors. In this paper, we thus propose a semi-supervised learning approach for bearing anomaly detection using variational autoencoder (VAE) based deep generative models, which allows for effective utilization of dataset when only a small subset of data have labels. Finally, a series of experiments is performed using both the Case Western Reserve University (CWRU) bearing dataset and the University of Cincinnati's Center for Intelligent Maintenance Systems (IMS) dataset. The experimental results demonstrate that the proposed semi-supervised learning scheme greatly outperforms two mainstream semi-supervised learning approaches and a baseline supervised convolutional neural network approach, with the overall accuracy improvement ranging between 3% to 30% using different number of labeled samples.

arXiv

Semi-Supervised Learning of Bearing Anomaly Detection
via Deep Variational Autoencoders

Shen Zhang [1] **Fei Ye** [2] **Bingnan Wang** [3] **Thomas G. Habetler** [1]

Abstract

Most of the data-driven approaches applied to bearing fault diagnosis up to date are established in the supervised learning paradigm, which usually requires a large set of labeled data collected *a priori*. In practical applications, however, obtaining accurate labels based on real-time bearing conditions can be far more challenging than simply collecting a huge amount of unlabeled data using various sensors. In this paper, we thus propose a semi-supervised learning approach for bearing anomaly detection using variational autoencoder (VAE) based deep generative models, which allows for effective utilization of dataset when only a small subset of data have labels. Finally, a series of experiments is performed using both the Case Western Reserve University (CWRU) bearing dataset and the University of Cincinnati's Center for Intelligent Maintenance Systems (IMS) dataset. The experimental results demonstrate that the proposed semi-supervised learning scheme greatly outperforms two mainstream semi-supervised learning approaches and a baseline supervised convolutional neural network approach, with the overall accuracy improvement ranging between 3% to 30% using different number of labeled samples.

1. Introduction

Electric machines are widely employed in a variety of industry applications and electrified transportation systems, and they consume approximately 60% of all electric power produced. However, on certain occasions these machines may operate at unfavorable conditions, such as high ambient temperature, high moisture, and overload, which can eventually result in motor malfunctions that lead to high maintenance costs, severe financial losses, and various safety hazards. Among the many types of electric machine failures that include stator inter-turn faults due to winding insulation breakdown, broken-rotor bar faults, bearing faults, and air gap eccentricity, etc., it is revealed that the bearing fault is the most common fault type that accounts for 30% to 40% of all the machine failures, according to several surveys conducted by the IEEE Industry Application Society (IEEE-IAS) (IEEE, 1985) and the Japan Electrical Manufacturers' Association (JEMA) (JEMA, 2000)

In the last few years, many data-driven approaches have been applied to enhance the accuracy and reliability of bearing fault diagnosis (Zhang et al., 2019; Saufi et al., 2019). Specifically, many time-series signals used for monitoring the bearing condition have high-dimensional feature spaces, which makes it ideal to exploit deep learning algorithms to perform feature extraction and classification. While many models have presented satisfactory results after proper training sessions, most of them are established in the supervised learning paradigm, which requires a large set of labeled data collected *a priori* under all possible fault conditions.

For bearing anomaly detection, however, it is often times much easier to collect a large amount of data than to accurately obtain their corresponding labels, and this is especially the case for faults that evolved naturally over time. For instance, even though an electromechanical system will be eventually turned off due to the consequences of a fault, it is not easy to determine precisely when the first trace of such a fault actually shows up and how long it lasts at the incipient stage. Additionally, it may also suffer from data ambiguity issues due to challenges associated with properly labeling these data, especially at a transition stage when they exhibit early signs of fault features but are far from obvious when compared to those features of a fully developed fault. Therefore, for accuracy concerns, only the very beginning and the very end of the collected data can be confidently assigned to their corresponding labels, leaving many of the intermediate data unlabeled, which apparently cannot be exploited using supervised learning.

To train a bearing fault classifier with a good level of gen-

---

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta GA, USA [2]California PATH, University of California, Berkeley, Berkeley CA, USA [3]Mitsubishi Electric Research Laboratories, Cambridge MA, USA. Correspondence to: Shen Zhang <shenzhang@gatech.edu>.

eralization using only a small subset of labeled data, one approach is to apply specific algorithms that can make the most out of the labeled data and computing resources available. For example, data augmentation techniques such as generative adversarial networks (GAN) (Goodfellow et al., 2014), which is already applied in (Liu et al., 2018) and (Verstraete et al., 2019), as well as some cross validation methods such as leave-one-out and Monte Carlo. Additionally, another promising route is to apply the semi-supervised learning paradigm to leverage the manually labeled data and the massive unlabeled data.

Semi-supervised learning specifically considers the problem of classification when only a small subset of data have corresponding labels, and as of now only a handful of semi-supervised learning paradigms have been applied to bearing anomaly detection. For example, a support vector data description (SVDD) model proposed in (Liu and Gryllias, 2019) utilizes cyclic spectral coherence domain indicators to build the feature space and fits a hyper-sphere, which later calculate the Euclidean distances in order to isolate data from healthy and faulty conditions. Moreover, graph-based methods are employed in (Chen et al., 2018) and (Zhao et al., 2017) to construct a graph connecting similar samples in the dataset, so the class labels can propagate through the graph from labelled to unlabelled nodes. However, these approaches are sensitive to the graph structure and require eigen-analysis of the graph Laplacian, which limits the scale to which these methods can be applied. Instead of graph-based methods, (Razavi-Far et al., 2019a) captures the structure of the data using α-shape, which is mostly utilized for the surface estimation that reduces parameter tuning.

Additionally, the recent semi-supervised deep ladder network (Rasmus et al., 2015) is implemented in (Razavi-Far et al., 2019b) to identify one-stage parallel shaft helical gear faults in induction machine systems. The ladder network is accomplished through the integration of the supervised and unsupervised learning strategies by modeling hierarchical latent variables. However, the unsupervised component of the ladder network may not be helpful in the semi-supervised setting if the raw data do not exhibit significant clustering on the 2D manifold, which is often not the case for bearing vibration signals. While GANs can be also implemented for semi-supervised learning, it has been reported in (Dai et al., 2017) that a good semi-supervised classifier and a good generator cannot be obtained at the same time. In other words, good semi-supervised learning actually requires a bad generator, and the well-known difficulties to train GANs further adds to the uncertainty of applying them to semi-supervised learning tasks.

The motivation for the proposed research is both broad and specific, as we seek to approach the problem of bearing anomaly detection with solid theoretical explanation, and
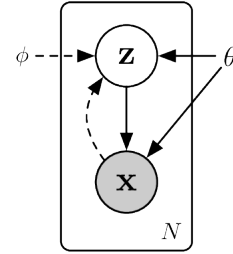

Figure 1. Architecture of VAE. The prior of $\mathbf{z}$ is regarded as part of the generative model (solid lines), thus the whole generative model is denoted as $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})\, p_\theta(\mathbf{z})$. The approximated posterior (dashed lines) is denoted as $q_\phi(\mathbf{z}|\mathbf{x})$.

leverage properties of both the labeled and unlabeled data to make the classifier performance more accurate than that based on the labelled data alone. Therefore, we adopt the semi-supervised learning algorithm as the deep generative model, which is built upon solid Bayesian theory and scalable variational inference. Although some prior work employing VAE can be found in (San Martin et al., 2018) and (Ellefsen et al., 2019), they are only utilizing the discriminative feature in the latent space for dimension reduction and use these features to train other external classifiers, while in this work we also make use of the generative function of VAE and train itself as a classifier through an integrated approach.

2. Background of Variational Autoencoders

Variational inference techniques (Kingma and Welling, 2013) are often adopted in the process of training and prediction, which are efficient methods to solve posteriors of the distributions derived by neural networks. The architecture of VAE is shown in Fig 1, which is formulated in a probabilistic perspective that specifies a joint distribution $p_\theta(\mathbf{x}, \mathbf{z})$ over the observed variables $\mathbf{x}$ and latent variables $\mathbf{z}$, such that $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$. Latent variables are drawn from a prior density $p(\mathbf{z})$ that is usually a multivariate unit Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and these local latent variables $\mathbf{z}$ are related to their corresponding observations $\mathbf{x}$ through the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$, which can be viewed as a probabilistic decoder (generator) to decode $\mathbf{z}$ into a distribution over the observation $\mathbf{x}$. The exact form of decoder $p_\theta(\mathbf{x}|\mathbf{z})$ can be modeled as a neural network with parameters θ consisting of the weights $\mathbf{W}$ and biases $\mathbf{b}$ of this network.

Having specified the decoding process, it is also desirable to perform inference, or to compute the posterior $p_\theta(\mathbf{z}|\mathbf{x})$ of latent variables $\mathbf{z}$ given the observed variables $\mathbf{x}$. Additionally, we also wish to optimize the model parameters θ with respect to $p_\theta(\mathbf{x})$ by marginalizing out the latent variables $\mathbf{z}$ in $p_\theta(\mathbf{x}, \mathbf{z})$. However, since the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$ is analytically intractable due to the Gaussian non-conjugate

prior $p(\mathbf{z})$, then the variational inference technique is used to find an approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ with optimized variational parameters ϕ that minimizes its Kullback-Leibler (KL) divergence to the true posterior. This approximated posterior $q_\phi(\mathbf{z}|\mathbf{x})$ can be also viewed as an encoder that is usually assumed to be $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{x})))$, where values of $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\sigma}_\phi(\mathbf{x})$ are optimized by neural networks.

Due to similar intractability issues such that computing the KL divergence explicitly would require the log marginal likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x})$, we'll maximize an alternative objective function known as the evidence lower bound (ELBO) of this log likelihood, which is defined as

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \underbrace{\log p_{\boldsymbol{\theta}}(\mathbf{x}) - \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \,\big\|\, p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right]}_{\mathrm{ELBO}} \quad (1)$$

Specifically, the ELBO is a lower bound on the probability of data given a specific model, which can be further written as

$$\begin{aligned}\mathrm{ELBO} &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) + \log p_{\boldsymbol{\theta}}(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right]\end{aligned}$$
$$(2)$$

Therefore, maximizing it with respect to model parameters $\boldsymbol{\theta}$ approximately maximizes the log marginal likelihood. Additionally, maximizing it with respect to variational parameters ϕ is equivalent to minimizing the KL divergence. The maximization of ELBO requires its gradients with respect to $\boldsymbol{\theta}$ and ϕ, which are also generally intractable. Currently, the dominant approach for circumventing this is by Monte Carlo (MC) integration (Geweke, 1989) to approximate the expectation of the gradients, as shown in Eqn (3), where $\mathbf{z}^{(l)}, l = 1 \ldots L$ are samples from $q_\phi(\mathbf{z}|\mathbf{x})$, and then perform stochastic gradient descent with the repeated MC gradient estimates.

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[f(\mathbf{z})\right] \approx \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z}^{(l)}) \quad (3)$$

3. Deep Generative Model Architecture for Semi-supervised Learning

This section presents two models for semi-supervised learning (Kingma et al., 2014) that exploit the generative functionality of VAE to improve the classification performance when only a small subset of data have labels. The labeled data has the form of $(\mathbf{x_l}, y_l)$, and the unlabeled data is represented by $\mathbf{x_u}$.

By learning a close variational approximation of the posterior, the encoder part of the VAE is able to provide an embedding or feature representation of the input data $\mathbf{x}$ as a set of latent features $\mathbf{z}$. The probabilities of the approximated
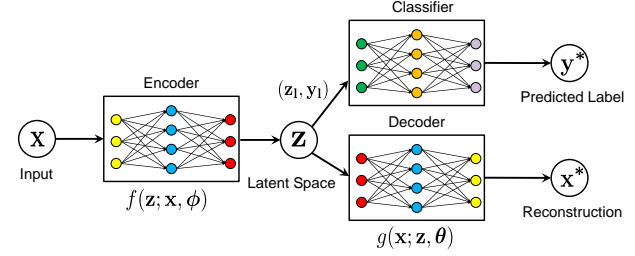

Figure 2. Illustration of the latent-feature discriminative model(M1).

posterior $q_\phi(\mathbf{z}|\mathbf{x})$ are formed by nonlinear transformations that can be modeled as a deep neural network $f(\mathbf{z}; \mathbf{x}, \phi)$ with variational parameters ϕ. On the other hand, the generator part of VAE takes a set of latent variables $\mathbf{z}$ and produces new observations $\mathbf{x}'$ using a suitable likelihood function $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, which can be modeled as another deep neural network $g(\mathbf{x}; \mathbf{z}, \boldsymbol{\theta})$ with model parameters $\boldsymbol{\theta}$.

3.1. Latent-feature discriminative model (M1)

A very straightforward implementation of VAE in semi-supervised learning, also referred to as the M1 model in (Kingma et al., 2014), is to train the model using both the labeled and unlabeled training data. When both the encoder and generator parameters are optimized, only the encoder part is utilized to provide a clustering of input data in the latent space. In most cases, the dimension of latent space variables $\mathbf{z}$ is much smaller than that of the input data $\mathbf{x}$, and these low-dimensional embeddings can be handled more easily in a supervised learning problem.

After training VAE, the classification task will be handed over to an external classifier. Only the labeled data $\mathbf{x_l}$ will be processed by the VAE encoder to determine their corresponding latent variables $\mathbf{z_l}$, which will be used as input features to train this classifier, such as a (transductive) SVM or multinomial regression, aided by their corresponding labels y_l, as shown in Fig. 2. Since the M1 model is trained using all data available in an unsupervised manner, it can generally promote a more accurate classification while only using a limited number of labeled data.

3.2. Generative semi-supervised model (M2)

A major limitation of the M1 model lies its disjoint nature in firstly training the VAE and later the external classifier. Specifically, the initial VAE training stage of the M1 model is a complete unsupervised process that does not directly involve any of the scarce labels y_l, which is entirely separated from the later classifier training stage that actually takes y_l. Therefore, another semi-supervised M2 model was also proposed in (Kingma et al., 2014) to address this problem.

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

$$
\begin{aligned}
\mathrm{ELBO_L} &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},y)}\left[\log p_\theta(\mathbf{x}|y,\mathbf{z}) + \log p_\theta(y) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x},y)\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|y,\mathbf{z}) + \log p_\theta(y) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] \\
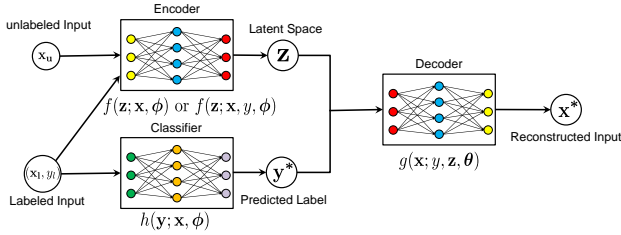&= -\mathcal{L}(\mathbf{x},y)
\end{aligned}
\tag{7}
$$

Figure 3. Illustration of the generative semi-supervised model (M2).

The M2 model can simultaneously deal with two cases: one where the labels are available, and one where they are not provided. Thus there are also two different approaches when constructing the approximate posterior q as well as the variational objective.

3.2.1. VARIATIONAL OBJECTIVE WITH UNLABELLED DATA

For the case where labels are missing, two separate posteriors $q_\phi(\mathbf{z}|\mathbf{x})$ and $q_\phi(y|\mathbf{x})$ will be involved in the VAE training phase, where $\mathbf{z}$ is still a vector latent variable similar to the M1 model, and y is the unobserved label y_u. This newly defined posterior approximation $q_\phi(y|\mathbf{x})$ will be used to construct the best classifier possible as our inference model (Kingma et al., 2014). Specifically, the two approximated posteriors of class labels y and latent variables $\mathbf{z}$ are defined as

$$
\begin{aligned}
q_\phi(y|\mathbf{x}) &= \mathrm{Cat}\left(y|\pi_\phi(\mathbf{x})\right) \\
q_\phi(\mathbf{z}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{z}|\mu_\phi(\mathbf{x}), \mathrm{diag}\left(\sigma_\phi^2(\mathbf{x})\right)\right)
\end{aligned}
\tag{4}
$$

where $\mathrm{Cat}\left(y|\pi_\phi(\mathbf{x})\right)$ is the concatenated multinomial distribution, $\pi_\phi(\mathbf{x})$ is similar to $\mu_\phi(\mathbf{x})$ and $\sigma_\phi(\mathbf{x})$ in such a way that it is also defined by a neural network parameterized by ϕ, which we need to learn from the training phase. By

combining the above two posteriors, a joint approximated posterior can be defined as

$$
q_\phi(y,\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}|\mathbf{x})q_\phi(y|\mathbf{x})
\tag{5}
$$

Similar to Eqn. 2, the revised $\mathrm{ELBO_U}$ that determines the variational objective for the unlabeled data can be written as Eqn. 6, where $\mathcal{L}(\mathbf{x},y)$ is the original ELBO defined in Eqn. 2,

3.2.2. VARIATIONAL OBJECTIVE WITH LABELLED DATA

Since the goal of semi-supervised learning is to train a classifier using a limited amount of labelled data combined with a vast majority of unlabeled data, it would be better to also incorporate the scarce labels to train the VAE classifier, which is unfortunately a featuer that cannot be accomplished using the M1 model. Similar to Eqn. 2, the revised $\mathrm{ELBO_L}$ that determines the variational objective for the labeled data can be written in the form of Eqn. 7.

3.2.3. COMBINED GENERATIVE SEMI-SUPERVISED MODEL OBJECTIVE

As can be observed in Eqn. 6, the distribution $q_\phi(y|\mathbf{x})$, which is used to construct the discriminative classifier, is only contained in the variational objective of the unlabeled data. This is still an undesirable property since the labeled data will not be involved to learn this distribution and thus the variational parameter ϕ. Therefore, an extra loss term should be superimposed onto the combined model objective, such that both the labeled and labeled data would contribute to the training process. Thus the final objective of the generative semi-supervised model is:

$$
\mathcal{J}^\alpha = \sum_{\mathbf{x}\sim\tilde{p}_\mathbf{u}}\mathcal{U}(\mathbf{x}) + \sum_{(\mathbf{x},y)\sim\tilde{p}_l}\left[\mathcal{L}(\mathbf{x},y) - \alpha\cdot\log q_\phi(y|\mathbf{x})\right]
\tag{8}
$$

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

With this objective function, we can implement a sufficiently large number of $\mathbf{x}$ as a mini-batch to enhance the stability of training two neural networks that work as an encoder or a decoder. The objective function can be properly evaluated using either labelled or unlabelled data, and finally we run gradient descents to update the network parameters $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{\phi}$. The network of the M2 model is very close to a vanilla VAE with an additional posterior on y and two additional terms on the objective function. An illustration of the M2 generative semi-supervised model is presented in Fig. 3.

3.3. Model Implementations

3.3.1. VAE BASED M1 MODEL IMPLEMENTATIONS

The VAE based M1 model consists of two deep neural networks $f(\mathbf{z}; \mathbf{x}, \boldsymbol{\phi})$ and $g(\mathbf{x}; \mathbf{z}, \boldsymbol{\theta})$ to represent the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and decoder $p_\theta(\mathbf{x}|\mathbf{z})$. The encoder structure has 2 convolutional layers and 1 fully connected layer with batch normalization, dropout and ReLU activation. The decoder consists of 1 fully connected layer followed by 3 transpose convolutional layers, among which the first 2 layers adopts ReLU activation and the final layer uses linear activation.

While training this VAE based M1 model, it is often times difficult to train a straight implementation of VAE that equally weighted the likelihood and the KL divergence, as shown in Eqn. 2, due to the "KL-vanishing" problem as the KL loss can undesirably reduce to zero (although it is expected to maintain a small value). To overcome this, the implementation of M1 model uses the "KL cost annealing" or "β VAE" (Bowman et al., 2015), which has a modified ELBO function as

$$\text{ELBO} = \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]}_{\text{Reconstruction}} - \beta \cdot \underbrace{\text{KL}\left(q_\phi(z|x)\|p(z)\right)}_{\text{KL Regularization}} \tag{6}$$

in which the introduced weight factor of the KL divergence term β will gradually increase from 0 to 1 over the course of the training epoch. When $\beta < 1$, the latent variable $\mathbf{z}$ is trained to focus more on capturing useful information for reconstruction of $\mathbf{x}$. When the full VAE objective is considered ($\beta = 1$), $\mathbf{z}$ learned in the earlier epoch earlier can be viewed as VAE initialization; such latent features are much more informative than the random start in constant schedule and thus are ready for the decoder to use (Liu et al., 2019)

After training the M1 model that is able to balance its recon-

struction and generation features, latent variables $\mathbf{z}$ in the latent space will be used as discriminative features that can be passed along to an external classifier. In this paper, the SVM classifier is adopted. While it is flexible to use any classifier of personal preference, the merit of the M1 model is to take advantage of the feature extraction functionality of VAE to reduce the dimension of the input data (1,024), to a much lower feature dimension, which is selected to be 128 in this study.

3.3.2. VAE BASED M2 MODEL IMPLEMENTATIONS

The VAE based M2 model adopts the same structure for $q_\phi(\mathbf{z}|\mathbf{x})$ as in the M1 model, and the decoder $p_\theta(\mathbf{x}|y, \mathbf{z})$ also has the same settings as M1 $p_\theta(\mathbf{x}|\mathbf{z})$. The classifier $q_\phi(y|\mathbf{x})$ has 2 convolutional layers and 2 max pooling layers with dropout and ReLU activation, followed by a final Softmax layer.

Two separate neural networks, one for the labeled data and one for the unlabeled data, are employed with identical network structure but different input/output specifications and loss functions (Keng, 2017). For example, for the labeled data, both $\mathbf{x_l}$ and y are treated as input to minimize the labeled $(\mathbf{x}, y) \sim \tilde{p}_l$ part in Eqn. 8, and the output would be the reconstructed $\mathbf{x_l'}$ and y', while for the unlabeled data, $\mathbf{x_u}$ serves as input to generate the reconstructed $\mathbf{x_u'}$.

Other hyper-parameters for the M2 model are also chosen empirically. We use 200 as the batch size for training, and run for 10 epochs. The dimension of the latent variable $\mathbf{z}$ is 128. We use the RMSprop optimizer with an initial learning rate of 10^{-4}.

4. Experimental Evaluation

In this section, we seek to validate the effectiveness of the variational autoencoder based deep generative model for semi-supervised learning in the context of bearing anomaly detection. The formulation of the diagnosis framework will be presented in detail, and the classifier performance will be compared against other popular semi-supervised learning methods, such as PCA and the vanilla autoencoder (AE), as well as CNN that represents a benchmark supervised learning paradigm.

4.1. Datasets

In this experiment, both the Case Western Reserve University (CWRU) bearing dataset and the University of Cincinnati's Center for Intelligent Maintenance Systems (IMS) datasets are adopted, which contain data from manually initiated and naturally evolved bearing defects, respectively.

Table 1. Selected class labels from the CWRU bearing 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					

Figure 4. Experimental test rig for collecting the IMS dataset.

4.1.1. CWRU DATASET

The CWRU dataset contains mechanical vibration signals collected from the drive-end bearing in a 2-hp induction motor dyno setup. Single point defects are manually created to the bearing inner raceway, the outer raceway, and the rolling element using electro-discharge machining. Different levels of fault severity are modeled using different defect diameters such as 7 mils, 14 mils, 21 mils, 28 mils, and 40 mils. Vibration data are collected with motor loads from 0 to 3 hp, and motor speeds from 1,720 to 1,797 rpm using two accelerometers installed at both the drive end and fan end of the motor housing at a sampling frequency of either 12 kHz or 48 kHz. The generated dataset is recorded and also made publicly available on the CWRU bearing data center website ((CWRU), 2005).

Since the purpose of the bearing diagnostic framework is to accurately reveal both the location and severity of a bearing defect, vibration data collected for the same fault type but at different speeds and loading conditions would be considered as having the same label. Based on this criterion, 10 classes are identified based on the location and size of the bearing defect, and a detailed list of all 10 scenarios with their defect size and locations are demonstrated in Table 1.

4.1.2. IMS DATASET

In most cases, bearing damage is artificially initiated in order to accelerate the degradation process. Therefore, the IMS dataset is also applied in this study to verify the effectiveness of VAE based generative models for detecting naturally evolved bearing defects. The IMS bearing dataset has been collected on an endurance test rig illustrated in Fig. 4. Specifically, there are 4 double row bearings mounted on the same shaft, which is coupled to the motor shaft via a rubber belt. Each of the 4 bearings under test is applied with a radial load of 6,000 lbs, and since they are not used
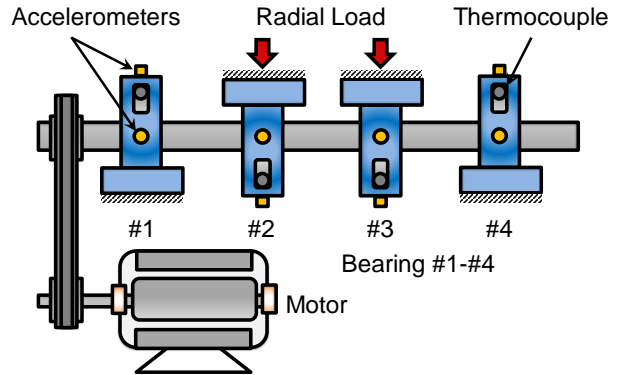
to support the motor or any active rotating movement, the endurance test can continue even these bearings have shown implications of failure. As a matter of fact, the stopping criterion for the test is when the accumulation of debris on a magnetic plug exceeded a certain level (Lee et al., 2007), which is a major difference when compared to a realistic scenario where the bearing is supporting the motor or transmission, and the test needs to be stopped quickly after spotting any anomaly conditions.

The IMS data contains 3 subsets of data when the motor runs at a constant speed of 2,000 rpm, and there are two PCB 253B33 High sensitivity Quart ICP accelerometers installed on each bearing (x and y positions) for the first subset, while the other two subsets only have one accelerometer on each bearing. A 1-second acquisition has been made every ten minutes and stored in a separate data file that includes 20,480 sample points, with an an exception for the first subset for which the first 92 files have been acquired every five minutes. As briefly discussed earlier, since the endurance test may continue after bearing degradation, a summary of the bearing anomaly condition and the approximated degradation starting point is listed in Table 2, where the starting points are estimated using 3 different algorithms adapted from (Hasani et al., 2017) .

4.2. Data Preprocessing

4.2.1. DATA SEGMENTATION

The diagnostic procedure begins with segmentation, during which the collected vibration signals are first divided into a number of segments with an equal length.

For the CWRU dataset, the length of the vibration signal for each type of drive end motor bearing fault shown in Table 1 is roughly 120,000 at three different speeds, namely 1,730 rpm, 1,750 rpm, and 1,772 rpm. Data collected at these speeds would constitute the complete fault data, which is

Table 2. Bearing anomalies and degradation starting point of the IMS dataset

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.

later segmented according to a fixed window size of 1,024 (spans 85.3 ms in the dataset that was collected at 12 kHz) and a sliding ratio of 0.2. Ultimately the number of segments for the training data and the test data are 12,900 and 900, respectively. Although the percentage of test data might seem low (around 7%) at the first glance, only a maximum of around 2,000 labeled training samples will be used in this semi-supervised learning setting, making the ratio of test data to the labeled training data above 40%.

Similarly, the IMS dataset is also segmented using a fixed window size of 1,024 with data collected at 20.48 kHz. Due to a high level of noise present in the "Subset 3 Bearing 3" condition as reported in (Hasani et al., 2017), only the first 3 fault conditions shown in Table 2 are selected for algorithm validation. In addition, in order to simulate a realistic scenario where the equipment would be stopped right after spotting a bearing anomaly condition, 210 files are selected for each fault type, with 15 of them chosen after their degradation starting points using the auto-encoder-correlation-based (AEC) prognostic algorithm. For example, for the "Subset 1 Bearing 3" condition, data files 1,832 to 2,042 would be selected, while the degradation starting point for this fault scenario is 2,027. On the other hand, the healthy data is selected as the first 110 files also from "Subset 1 Bearing 3".

The last 10 files will be treated as the test set while the remaining 200 files form the entire training set for semi-supervised learning algorithms. As each files contains 20,480 data points that can be divided into 20 data segments, there are altogether 4,000 data segments for each class, or 16,000 for all 4 classes (healthy, outer race, inner race, rolling element). The size of the test set is 800 data segments for all 4 classes. Similar to the way of constructing the training set and the test set, although the percentage of test data might seem low (around 5%) at the first glance, this ratio is more significant when compared against the amount of labeled data, since only a limited number of the 160,000

training data segments will be labeled.

In order to simulate the challenges associated with data labeling for real world applications, the training set will be labeled starting from the last one of the 200 files, for which we should have the highest level of confidence regarding the accuracy of its label. Then more test cases will be created by labeling more preceding files but with a decreased level of confidence. The purpose is to validate if the inaccurate labeling would affect the accuracy of supervised learning algorithms, and if these data are left unlabeled, can they still promote the classifier accuracy using semi-supervised learning algorithms.

4.2.2. DATA SHUFFLING AND STANDARDIZATION

After the initial loading and segmentation of the data, the data segments are still sorted according to their class (fault types). Therefore, data shuffling needs to be performed to make sure that the training and test sets are representative of the overall distribution of the dataset, thus reducing the variance and making sure that the trained VAE models would remain general and come with less overfitting.

The conventional standardization technique is applied to both the training set and the test set to make sure the bearing vibration data in each segment have zero-mean and unit-variance, which is accomplished by first subtracting the mean of the original data and then dividing the result by the original standard deviation.

4.3. Experimental Results

Having successfully trained the VAE based model, the reconstructed bearing vibration signal should closely resemble the original signal, as shown in Fig. 5. Although a perfect reconstruction (reconstruction error = 0) might cause a loss of generality and degrade the generative feature of VAE, a reasonably close reconstruction with a small error can indicate that the VAE has achieved a balance between re-
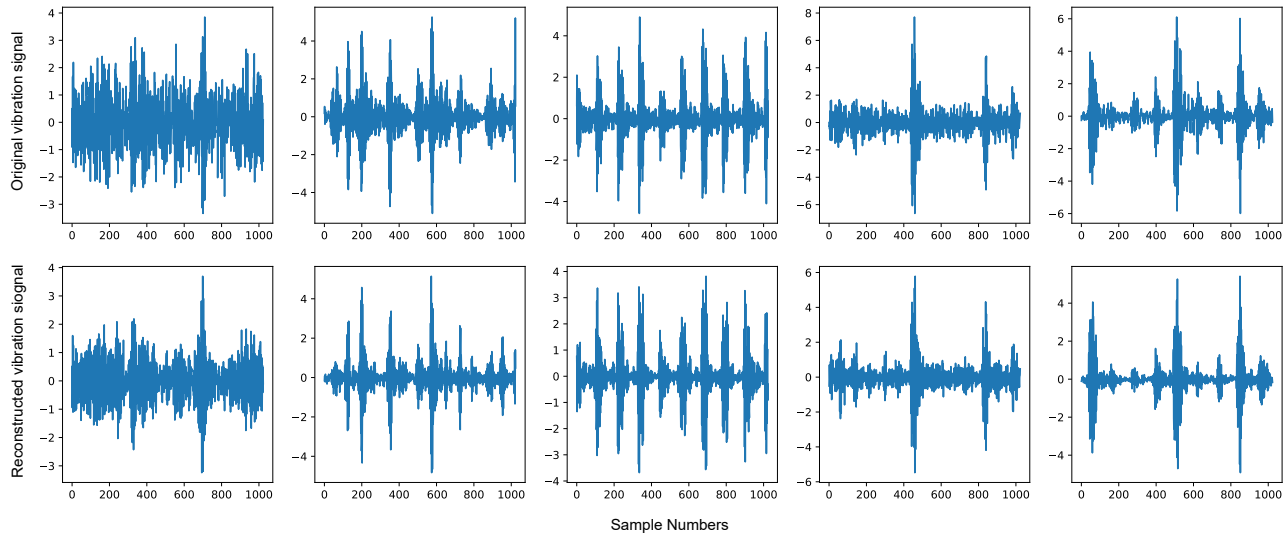
Figure 5. Comparison of the original and the reconstructed bearing vibration signals after training the VAE M1 model (top row: the original signal; bottom row: the reconstructed signal).

construction and generation, which is crucial for exploiting the generative feature of the algorithm.

Besides implementing the variational autoencoder based deep generative M1 and M2 model to perform bearing fault diagnostics with semi-supervised learning, other popular semi-supervised learning methods, such as PCA and the traditional autoencoder, as well as mainstream supervised learning paradigms such as CNN, are also implemented using the same training and test set.

4.3.1. CWRU DATASET

For the CWRU dataset, a total of 10 rounds of semi-supervised experiments are performed on randomly shuffled training and test sets. Only a fraction of labels of the 129,000 training samples will be visible to different algorithms to construct a semi-supervised setting for bearing anomaly detection, where 8 case studies are carried out with 10, 50, 100, 245, 516, 860, 1,075, and 2,150 labels, representing roughly 0.08%, 0.39%, 0.78%, 1.9%, 4%, 6.67%, 8.33%, and 16.67% of the training data length, respectively.

The average accuracy and standard deviation for different algorithms after performing 10 rounds of experiments are presented in Tab. 3, in which the latent-feature discriminative model (M1) performs better than other benchmark unsupervised models such as PCA and autoencoder, demonstrating the effectiveness of the latent space in providing robust features that allow for easier classification. It is worthwhile to notice that initially the VAE based M1 model compares favorably against CNN til the number of labeled samples $N = 860$, after which it becomes slightly inferior, which is in contradiction to their implementation results

on the MNIST dataset provided in (Kingma et al., 2014). One interpretation of this deviation might be due to the 1-D time-series CWRU dataset that has more explicit feature representations to be directly captured by supervised and fine-tuned structures such as CNN, as thus it may only require around 2,000 labeled training data segments to achieve a satisfactory result.

However, by combining features learnt in the M1 model with a classification mechanism directly in the same model, as in the conditional generative model (M2), we are able to get much better result than simply using a SVM classifier as in the M1 model. Specifically, it only needs 4% of the training data to have labels to reach a fault classification accuracy of around 95%, while the best value attainable from the rest of the algorithms is only 57.06%, demonstrating a nearly 40% increase in accuracy with very consistent performance, as its standard deviation is as low as 0.88%.

4.3.2. IMS DATASET

Similarly, 10 rounds of semi-supervised experiments are performed on data retrieved from the IMS dataset, and 10 case studies are carried out by labeling the last 10, 40, 100, 200, 400, 800, 1,000, 2,000, 4,000, and 8,000 data segments of the training set, representing roughly 0.06%, 0.25%, 0.63%, 1.25%, 2.5%, 5%, 6.25%, 12.5%, 25%, and 50% of the training data, respectively.

The average accuracy and standard deviation for the same 5 algorithms after performing 10 rounds of experiments are presented in Tab. 4. The latent-feature discriminative model (M1) performs better than PCA but has an almost equivalent performance as an autoencoder, which indicates

Table 3. Experimental results of semi-supervised classification on CWRU bearing dataset with limited labels

N	PCA+SVM	Autoencoder	CNN	VAE M1	VAE M2
10	17.10 ($\pm$ 1.25)	27.72 ($\pm$ 1.28)	21.33 ($\pm$ 2.52)	23.04 ($\pm$ 1.32)	23.71 ($\pm$ 1.67)
50	26.82 ($\pm$ 1.98)	33.40 ($\pm$ 1.24)	30.74 ($\pm$ 2.88)	32.93 ($\pm$ 2.01)	40.57 ($\pm$ 2.91)
100	32.79 ($\pm$ 1.99)	37.96 ($\pm$ 0.65)	34.50 ($\pm$ 2.16)	36.91 ($\pm$ 1.37)	60.04 ($\pm$ 3.57)
300	46.88 ($\pm$ 2.10)	44.06 ($\pm$ 2.61)	60.28 ($\pm$ 3.42)	47.03 ($\pm$ 1.22)	87.63 ($\pm$ 2.80)
516	50.56 ($\pm$ 1.69)	50.88 ($\pm$ 2.03)	75.41 ($\pm$ 2.74)	57.06 ($\pm$ 1.76)	94.16 ($\pm$ 1.66)
860	64.48 ($\pm$ 1.88)	58.89 ($\pm$ 1.81)	87.39 ($\pm$ 0.93)	67.19 ($\pm$ 1.70)	96.77 ($\pm$ 0.38)
1075	65.32 ($\pm$ 2.05)	62.83 ($\pm$ 1.61)	91.07 ($\pm$ 1.46)	71.97 ($\pm$ 1.40)	97.86 ($\pm$ 0.51)
2150	78.91 ($\pm$ 1.35)	77.09 ($\pm$ 0.98)	97.19 ($\pm$ 0.99)	86.59 ($\pm$ 1.43)	98.06 ($\pm$ 0.88)

Table 4. Experimental results of semi-supervised classification on IMS bearing dataset with limited labels

N	PCA+SVM	Autoencoder	CNN	VAE M1	VAE M2
10	56.74 ($\pm$ 0.44)	54.10 ($\pm$ 1.58)	53.92 ($\pm$ 2.77)	32.93 ($\pm$ 3.81)	52.79 ($\pm$ 1.98)
40	61.60 ($\pm$ 0.63)	64.54 ($\pm$ 2.07)	59.08 ($\pm$ 2.68)	72.01 ($\pm$ 1.91)	66.27 ($\pm$ 8.31)
100	67.22 ($\pm$ 1.15)	67.07 ($\pm$ 1.49)	62.93 ($\pm$ 4.10)	76.61 ($\pm$ 1.48)	71.15 ($\pm$ 6.24)
200	70.31 ($\pm$ 0.49)	73.42 ($\pm$ 0.94)	68.64 ($\pm$ 5.40)	78.74 ($\pm$ 1.25)	76.54 ($\pm$ 3.58)
400	75.38 ($\pm$ 0.90)	78.42 ($\pm$ 1.17)	74.20 ($\pm$ 3.17)	81.66 ($\pm$ 1.02)	82.78 ($\pm$ 2.21)
800	77.85 ($\pm$ 0.62)	84.81 ($\pm$ 0.78)	78.73 ($\pm$ 2.98)	85.03 ($\pm$ 1.15)	88.45 ($\pm$ 1.71)
1000	78.19 ($\pm$ 0.59)	85.83 ($\pm$ 0.85)	81.29 ($\pm$ 4.18)	86.61 ($\pm$ 1.27)	89.66 ($\pm$ 1.54)
2000	78.50 ($\pm$ 0.30)	86.61 ($\pm$ 0.77)	86.62 ($\pm$ 4.11)	87.20 ($\pm$ 1.18)	90.87 ($\pm$ 1.97)
4000	78.96 ($\pm$ 0.72)	83.72 ($\pm$ 0.89)	87.74 ($\pm$ 0.54)	85.14 ($\pm$ 0.96)	92.01 ($\pm$ 0.92)
8000	79.06 ($\pm$ 0.65)	84.00 ($\pm$ 1.21)	81.56 ($\pm$ 2.79)	85.36 ($\pm$ 1.17)	88.11 ($\pm$ 3.47)

that the discriminative latent space of VAE has no obvious advantage over the encoded feature space of vanilla autoencoders. However, the M1 model also outperforms the supervised learning algorithm CNN with a limited number of labeled data segments from $N = 40$ to $N = 1,000$ by approximately 5% to 15% with much smaller variance, demonstrating the benefit of incorporating the vast majority of unlabeled data in the training process.

Similar to the comparison result obtained using the CWRU dataset, the classifier performance of the deep generative VAE M2 model outperforms other 4 algorithms, showcasing the strength of the integrated training process of the VAE model and its built-in classifier. An important observation can be made that when the number of labeled data segments increases from $N = 4,000$ to $N = 8,000$, the accuracy of CNN decreases by more than 6%, while VAE M2 also suffers a 4% loss. This can be largely attributed to the fact that the healthy data have been mistakenly labeled as faulty ones, although this decrease in accuracy was expected to happen while transitioning from $N = 1,000$ to $N = 2,000$, since there are only 1,000 faulty data segments before the anticipated bearing degradation starting point based on AEC as shown in Table 2. This inconsistency might be caused by the accuracy of these estimated degradation starting points, since an early indication of the "Subset 1 Bearing 3" fault using the MAS-Kurtosis method is projected to be 2,320

data segments (9,280 for all 4 classes) earlier than the currently employed AEC results. Therefore, the ground-truth degradation point may lie somewhere in the middle between these two results.

However, the experimental results obtained in the current setting are still able to support the earlier claims that taking advantage of the unlabeled data can effectively promote the classifier performance using the proposed VAE based semi-supervised learning methods, especially the deep generative M2 model. Additionally, it is also shown that inaccurate labeling can impair the accuracy of supervised learning algorithms in realistic applications, if the labels are assigned without a good level of confidence.

5. Conclusion

In this paper, we applied two semi-supervised models based on the generative feature of VAE for bearing anomaly detection. The new algorithms has shown to greatly outperform state-of-art unsupervised and supervised learning algorithms. For the CWRU dataset, the improvement of the VAE M2 model with 4% of labeled training data obtained a roughly 40% accuracy enhancement. Considering the fact that the bearing defects are manually initiated in this dataset and are thus inconsistent with real-world scenarios that involve naturally evolved defects, we've also adopted the

IMS dataset to validate the proposed semi-supervised VAE models. The results have shown that inaccurate labeling can degrade the classifier performance of the mainstream supervised learning algorithms, while incorporating semi-supervised models and keeping many of the miss-labeled data *unlabeled* can be an effectively method improve the classifier performance.

References

M. Blodt, P. Granjon, B. Raison, and G. Rostaing. 2008. Models for Bearing Damage Detection in Induction Motors Using Stator Current Monitoring. *IEEE Transactions on Industrial Electronics* 55, 4 (April 2008), 1813–1822.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349* (2015).

Xinan Chen, Zhipeng Wang, Zhe Zhang, Limin Jia, and Yong Qin. 2018. A Semi-Supervised Approach to Bearing Fault Diagnosis under Variable Conditions towards Imbalanced Unlabeled Data. *Sensors* 18, 7 (2018), 2097.

Case Western Reserve University (CWRU). 2005. Bearing Data Center. [Online]. Available: https://csegroups.case.edu/bearingdatacenter/pages/project-history. (2005). Accessed: Oct. 2019.

Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan R Salakhutdinov. 2017. Good semi-supervised learning that requires a bad gan. In *Advances in neural information processing systems*. 6510–6520.

A. L. Ellefsen, E. Bjørlykhaug, V. Æsøy, and H. Zhang. 2019. An Unsupervised Reconstruction-Based Fault Detection Algorithm for Maritime Components. *IEEE Access* 7 (2019), 16101–16109.

John Geweke. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society* (1989), 1317–1339.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

J. Harmouche, C. Delpha, and D. Diallo. 2015. Improved Fault Diagnosis of Ball Bearings Based on the Global Spectrum of Vibration Signals. *IEEE Transactions on Energy Conversion* 30, 1 (March 2015), 376–383.

Ramin M Hasani, Guodong Wang, and Radu Grosu. 2017. An automated auto-encoder correlation-based health-monitoring and prognostic method for machine bearings. *arXiv preprint arXiv:1703.06272* (2017).

IEEE. 1985. Report of Large Motor Reliability Survey of Industrial and Commercial Installations, Part I. *IEEE Transactions on Industry Applications* IA-21, 4 (July 1985), 853–864.

F. Immovilli, A. Bellini, R. Rubini, and C. Tassoni. 2010. Diagnosis of Bearing Faults in Induction Machines by Vibration or Current Signals: A Critical Comparison. *IEEE Transactions on Industry Applications* 46, 4 (July 2010), 1350–1359.

JEMA. 2000. On recommended interval of updating induction motors (in Japanese). (2000).

M. Kang, J. Kim, and J. Kim. 2015. An FPGA-Based Multicore System for Real-Time Bearing Fault Diagnosis Using Ultrasampling Rate AE Signals. *IEEE Transactions on Industrial Electronics* 62, 4 (April 2015), 2319–2329.

Brian Keng. 2017. Semi-supervised Learning with Variational Autoencoders. [Online]. Available: http://bjlkeng.github.io/posts/semi-supervised-learning-with-variational-autoencoders/. (2017). Accessed: Oct. 2019.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*. 3581–3589.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

J Lee, H Qiu, G Yu, and J Lin. 2007. Rexnord technical services. *Bearing data set* (2007).

W. Li, S. Zhang, and G. He. 2013. Semisupervised Distance-Preserving Self-Organizing Map for Machine-Defect Detection and Classification. *IEEE Transactions on Instrumentation and Measurement* 62, 5 (May 2013), 869–879.

Chenyu Liu and Konstantinos Gryllias. 2019. A semi-supervised Support Vector Data Description-based fault detection method for rolling element bearings based on Cyclic Spectral Coherence.

Han Liu, Jianzhong Zhou, Yanhe Xu, Yang Zheng, Xuanlin Peng, and Wei Jiang. 2018. Unsupervised fault diagnosis of rolling bearings using a deep neural network based on generative adversarial networks. *Neurocomputing* 315 (2018), 412–424.

Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, Lawrence Carin, et al. 2019. Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing. *arXiv preprint arXiv:1903.10145* (2019).

A. Ming, W. Zhang, Z. Qin, and F. Chu. 2015. Dual-Impulse Response Model for the Acoustic Emission Produced by a Spall and the Size Evaluation in Rolling Element Bearings. *IEEE Transactions on Industrial Electronics* 62, 10 (Oct 2015), 6606–6615.

Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. In *Advances in neural information processing systems*. 3546–3554.

R. Razavi-Far, E. Hallaji, M. Farajzadeh-Zanjani, and M. Saif. 2019a. A Semi-Supervised Diagnostic Framework Based on the Surface Estimation of Faulty Distributions. *IEEE Transactions on Industrial Informatics* 15, 3 (March 2019), 1277–1286.

R. Razavi-Far, E. Hallaji, M. Farajzadeh-Zanjani, M. Saif, S. H. Kia, H. Henao, and G. Capolino. 2019b. Information Fusion and Semi-Supervised Deep Learning Scheme for Diagnosing Gear Faults in Induction Machine Systems. *IEEE Transactions on Industrial Electronics* 66, 8 (Aug 2019), 6331–6342.

Gabriel San Martin, Enrique López Droguett, Viviane Meruane, and Márcio das Chagas Moura. 2018. Deep variational auto-encoders: A promising tool for dimensionality reduction and ball bearing elements fault diagnosis. *Structural Health Monitoring* (2018), 1092–1128.

S. R. Saufi, Z. A. B. Ahmad, M. S. Leong, and M. H. Lim. 2019. Challenges and Opportunities of Deep Learning Models for Machinery Fault Detection and Diagnosis: A Review. *IEEE Access* 7 (2019), 122644–122662.

R. R. Schoen, T. G. Habetler, F. Kamran, and R. G. Bartfield. 1995. Motor bearing damage detection using stator current monitoring. *IEEE Transactions on Industry Applications* 31, 6 (Nov 1995), 1274–1279.

David Benjamin Verstraete, Enrique López Droguett, Viviana Meruane, Mohammad Modarres, and Andrés Ferrada. 2019. Deep semi-supervised generative adversarial fault diagnostics of rolling element bearings. *Structural Health Monitoring* (2019), 1–22.

Shen Zhang, Shibo Zhang, Bingnan Wang, and Thomas G Habetler. 2019. Machine Learning and Deep Learning Algorithms for Bearing Fault Diagnostics-A Comprehensive Review. *arXiv preprint arXiv:1901.08247* (2019).

M. Zhao, B. Li, J. Qi, and Y. Ding. 2017. Semi-supervised classification for rolling fault diagnosis via robust sparse and low-rank model. In *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*. 1062–1067.