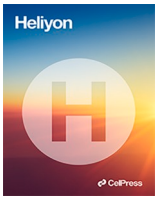
Heliyon 10 (2024) e35407

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A novel deep learning framework for rolling bearing fault diagnosis enhancement using VAE-augmented CNN model

Yu Wang, Dexiong Li, Lei Li, Runde Sun, Shuqing Wang *

Department of Electrical Engineering, Shijiazhuang Institute of Railway Technology, Shijiazhuang, 050041, China

ARTICLE INFO

Keywords:
Rolling bearing
Fault diagnosis
Variational autoencoder
Deep learning
Convolutional neural networks

ABSTRACT

In the context of burgeoning industrial advancement, there is an increasing trend towards the integration of intelligence and precision in mechanical equipment. Central to the functionality of such equipment is the rolling bearing, whose operational integrity significantly impacts the overall performance of the machinery. This underscores the imperative for reliable fault diagnosis mechanisms in the continuous monitoring of rolling bearing conditions within industrial production environments. Vibration signals are primarily used for fault diagnosis in mechanical equipment because they provide comprehensive information about the equipment's condition. However, fault data often contain high noise levels, high-frequency variations, and irregularities, along with a significant amount of redundant information, like duplication, overlap, and unnecessary information during signal transmission. These characteristics present considerable challenges for effective fault feature extraction and diagnosis, reducing the accuracy and reliability of traditional fault detection methods. This research introduces an innovative fault diagnosis methodology for rolling bearings using deep convolutional neural networks (CNNs) enhanced with variational autoencoders (VAEs). This deep learning approach aims to precisely identify and classify faults by extracting detailed vibration signal features. The VAE enhances noise robustness, while the CNN improves signal data expressiveness, addressing issues like gradient vanishing and explosion. The model employs the reparameterization trick for unsupervised learning of latent features and further trains with the CNN. The system incorporates adaptive threshold methods, the "3/5" strategy, and Dropout methods. The diagnosis accuracy of the VAE-CNN model for different fault types at different rotational speeds typically reaches more than 90 %, and it achieves a generally acceptable diagnosis result. Meanwhile, the VAE-CNN augmented fault diagnosis model, after experimental validation in various dimensions, can achieve more satisfactory diagnosis results for various fault types compared to several representative deep neural network models without VAE augmentation, significantly improving the accuracy and robustness of rolling bearing fault diagnosis.

1. Introduction

In the realm of contemporary industrial production, mechanical equipment serves as a foundational element, vital for the operational success of modern enterprises. The enhancement of mechanical equipment through the integration of artificial intelligence, Internet of Things, and similar technologies, has led to significant advancements in terms of refinement, systematization, intelligence,

* Corresponding author.
E-mail address: wanghb2311@163.com (S. Wang).

https://doi.org/10.1016/j.heliyon.2024.e35407
Received 13 March 2024; Received in revised form 17 June 2024; Accepted 29 July 2024
Available online 30 July 2024

and automation [1]. Among these components, rolling bearings, a quintessential element of mechanical transmission, find extensive utilization across various industrial and transportation machineries. However, these bearings are susceptible to numerous failure modes including fatigue damage, cracks, and wear, particularly under prolonged high load conditions. Such failures not only escalate maintenance costs and extend equipment downtime but also heighten the risk of accidents. Consequently, the investigation into rolling bearing defect diagnostics emerges as a crucial field, offering valuable insights for prompt identification of aberrations in mechanical equipment, thereby safeguarding operational stability and augmenting business productivity [2]. Contemporary fault detection methodologies predominantly focus on scrutinizing monitoring signals from malfunctioning equipment, analyzing time or frequency characteristics to detect variations in amplitude and frequency over time [3]. Traditional analysis techniques, such as Fast Fourier transform [4], empirical mode decomposition (EMD) [5], wavelet transform [6], and variational mode decomposition [7], have demonstrated efficacy in extracting fault features of rolling bearings to a certain degree. However, these techniques encounter challenges in real-world industrial settings due to the complex operational conditions of bearings. The presence of background noise and concurrent faults further complicates the vibration signals, making the diagnosis and analysis of bearing faults more difficult.

In light of these complexities, the adoption of deep neural network-based techniques for rolling bearing fault diagnosis has gained momentum, propelled by the advancements in deep learning [8–10]. Deep learning, in contrast to conventional shallow learning approaches, facilitates autonomous, adaptive extraction of high-dimensional features from input signals, thereby reducing reliance on expert knowledge and enhancing the precision and efficacy of fault detection. Deep learning models have thus emerged as a favored approach in defect diagnostics. Recent literature illustrates various innovations in this domain: a dual-input fault diagnosis model utilizing CNN and LSTM networks has been proposed for enhanced noise immunity under diverse noise and load conditions [11]; an improved CNN-SVM methodology for rapid motor bearing fault diagnosis [12]; and an enhanced LeNet-5 model tailored for scenarios with incomplete bearing failure samples [13].

A comprehensive review of the literature underscores the remarkable success of deep learning in detecting rolling bearing faults. However, these achievements often require extensive neural network training and large annotated datasets, a process known as supervised learning. This requirement can be restrictive in practical diagnosis scenarios, where obtaining substantial fault data or a sufficient number of labeled training samples may be impractical due to the unique characteristics of some mechanical devices [14]. To address these limitations, data augmentation has emerged as a potential solution, but current approaches still fall short of meeting the extensive training data needs for effective intelligent fault detection.

Recent advancements in Variational Autoencoders (VAE) offer a promising avenue in this regard [15–17]. VAEs, through their ability to capture data distributions via neural networks and generate unseen samples, excel in extracting deep abstract features from original input data. The integration of VAEs with deep convolutional neural network models, forming hybrid models, has shown potential in rolling bearing defect detection. This research introduces a VAE-enhanced deep convolutional neural network model, the VAE-CNN model, combining the strengths of CNN's signal data representation and VAE's noise resilience. The foundation of this study is the VAE's capability for realistic data generation based on semi-supervised classification. Meanwhile, novel techniques are integrated into the model training process: an adaptive threshold method for personalized warning levels, a "3/5" method to minimize misclassification, and a Dropout method to combat overfitting. The diagnosis accuracy of the VAE-CNN model for different fault types at different rotational speeds typically reaches more than 90 %, and it achieves a generally acceptable diagnosis result. Therefor, the practical implementation of cloud-based collaboration for rolling bearing fault diagnosis showcases the VAE-CNN model's enhanced robustness, applicability, and diagnostic accuracy.

2. Related theories

This section systematically introduces and expounds the research basis of this paper, including intelligent fault diagnosis based on deep learning and variational autoencoder theories, so as to provide reference for the subsequent model construction.

2.1. Intelligent fault diagnosis based on deep learning

Due to their effective feature extraction, deep neural networks have been employed extensively in intelligent defect diagnosis

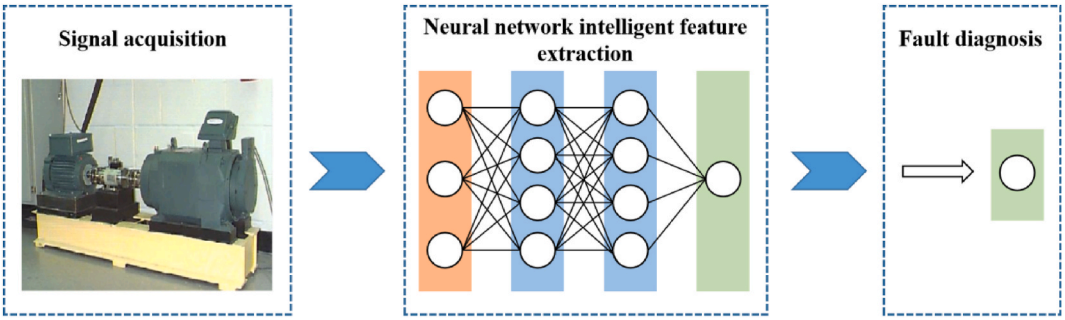


Fig. 1. Intelligent fault diagnosis process based on neural network.

scenarios [18]. They are currently used in one of two ways to diagnose faults. The first method makes use of the original one-dimensional vibration signal and divides the data into temporal signals of a specific length using sliding window interception. This method creates a one-dimensional deep neural network that can learn fault diagnosis techniques. The second is to discuss how deep neural networks are applied in image processing. Data pre-processing involves cutting and stacking or time-frequency transforming one-dimensional vibration signals into a two-dimensional matrix, and using this two-dimensional matrix to train a two-dimensional deep neural network to recognize faults.

The following steps are involved in intelligent defect diagnosis using deep neural networks: Signal acquisition is the process of gathering vibration signals, which include information about the condition of the equipment, using sensors that are often located on the apparatus. Deep neural networks are utilized for feature extraction. Fault diagnosis is the process of identifying a problem with a piece of equipment based on features that the neural network has extracted. In Fig. 1, the specific procedure is displayed.

The intelligent fault diagnosis process depicted in Fig. 1 is articulated in an academic context as follows: The methodology initiates with the transmission of the input vibration signal through the neural network, a pivotal step in the fault diagnosis procedure. Throughout this phase, the network, functioning as a nonlinear classifier, is responsible for extracting features from the vibration signal and subsequently generating an output that predicts the fault. Post this initial diagnosis, the network's output, which embodies the diagnostic conclusion, is juxtaposed with the actual label of the signal. The discrepancy between these two elements is then retroactively fed into the neural network. This feedback mechanism is instrumental in adjusting the network's parameters. This iterative process persists throughout the training phase until the model attains an optimal state, defined by the minimization of the divergence between the network's output and the true label to a level beneath the predetermined ideal threshold for training. Distinct from traditional fault diagnosis methodologies, which rely heavily on expert experience, the deep learning-based intelligent fault diagnosis approach excels in autonomously uncovering and leveraging latent, valuable information within the vibration signal. Its integration of feature extraction and refinement within the fault diagnosis process markedly enhances diagnostic efficacy [19].

Contemporary research in this domain has witnessed the successful application of various network models, evolved from the foundational feedforward neural networks, in defect diagnosis scenarios [20–22]. This study contributes to this evolving field by developing a model that amalgamates a deep convolutional neural network with a variational autoencoder, specifically tailored for the detection of defects in rolling bearings. This integration not only leverages the strengths of both neural network architectures but also addresses the complexities inherent in accurately diagnosing rolling bearing defects.

2.2. Variational autoencoder

Many generative models were present before deep learning, but the majority of them were challenging to describe and model. Deep learning's introduction has aided specialists and academics in finding solutions to these issues. Variational autoencoders (VAE) [23] and generative adversarial networks (GAN) [24] are examples of generative models based on deep learning concepts. The VAE, a deep learning generative model based on variational theories, is adopted in this study. Its concepts and guiding principles are briefly discussed below.

In 2012, Kingma et al. presented the generative model VAE, which has since gained popularity as one of the most effective techniques for unsupervised learning of complicated distributions. It can decode low-dimensional potential space representations of high-dimensional data to produce fresh samples from random vectors in the potential space [25]. Compared to conventional autoencoders, VAE offer enhanced interpretability and a superior ability to capture the complexities of data distribution. The core principle of VAE involves encoding the original data into the mean and variance within the latent space. Subsequently, the latent vector is treated as noise, generated by sampling from a prior distribution. Fig. 2 displays the network structure model of VAE.

Combining network structure in Fig. 2, the precise VAE calculation procedure is as follows:

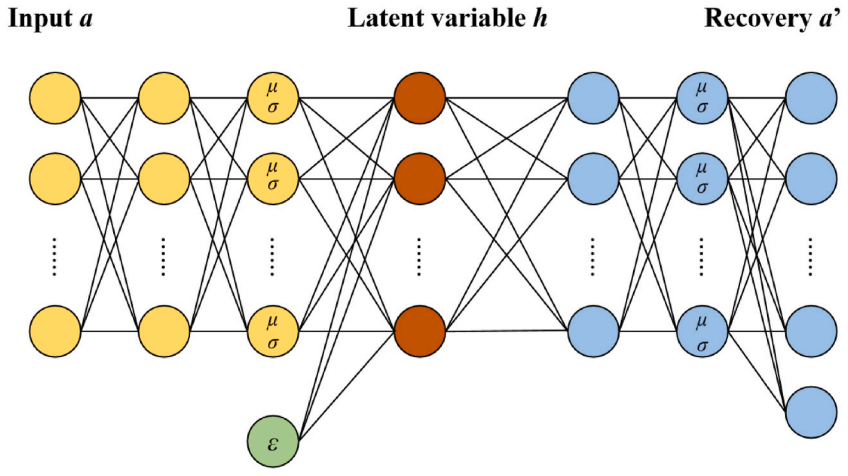


Fig. 2. Network structure model of VAE.

Assume that a is the input data and h is the implied variable. a is then produced by h. The generative model $g(a|h)$, which can be viewed as the self-encoder's decoder, converts h to a. The recognition model is a to h, and the encoder in the self-encoder is similar to $r(h|a)$. Formula (1) is utilized to model the observed sample $g(a)$:

$$g(a) = \int_h g(a,h,P)E_h[g(a|h,P)] \tag{1}$$

P stands for the collection of parameters, while $g(a,h,P)$ stands for the combined probability distribution of the observed sample a and the hidden variable h. In order to approximate the distribution $g(a)$ of a, the conditional probability $g(a|h)$ is introduced in Formula (1). The conditional probability of a with regard to h is represented by $g(a|h,P)$. The fundamental principle of VAE is to calculate $g(a)$ based on a and sample to obtain the hidden variable h that is most likely to produce it. By creating a new function, $r(h|a)$, one can conditionally describe the hidden variable h's distribution in relation to the input observation sample. VAE builds $r(h|a)$ using the variational inference principle.

Maximizing the expectation of likelihood is the goal. The lower bound ELBO function $L(r)$ of the evidence of the variational, as stated in Formula (2), is the objective function for optimization in variational reasoning.

$$L(r) = E_r\{\ln[g(a|h,P)] - KL[r(h|a,P)\|g(h,P)]\} \tag{2}$$

where the KL scatter is used to measure the distance between two distributions. The second term in the expectation on the right-hand side of Formula (2)$KL[r(h|a,P)\|g(h,P)]$ represents the encoding of a into the hidden variable h.

Assuming that the observed sample a obeys a Gaussian distribution of $N[\mu(a,P),\sigma(a,P)]$ and the hidden variable h obeys a Gaussian distribution of N(0, 1), the rightmost KL distance in Formula (2) can be simplified and expressed as Formula (3):

$$KL\{N[m(a,P),\sigma(a,P)]\|N(0,1)\} = \frac{1}{2}tr[\sigma(a)] + \{m(a)^T[\mu(a) - k - \log\det\sigma(a)]\} \tag{3}$$

The integral of the posterior probability is intractable for high-dimensional or complex distributions, necessitating the use of sampling methods to estimate the likelihood function. Consequently, this research employs the Markov Chain Monte Carlo (MCMC) sampling approach [26] to address this challenge. Gradient back propagation cannot be used for the sampling operation in the heavy parametric processing, thus the sampling action is relocated to the output layer operation, where it can be combined with $e \sim N(0,1)$ sampling from $N[\mu(a),\sigma(a)]$ to obtain the mean value of $r(h|a)$ $\mu(a)$ and the covariance matrix $\sigma(a)$, and compute $h = \mu(a) + \sigma^{1/2}(a) \times e$. Then the objective function is transformed into Formula (4):

$$L(r) = E_r\{\ln[g(a|h) = \mu(a) + \sigma^{1/2}(a) \times e] - KL[r(h|a,P)\|g(h,P)]\} \tag{4}$$

The backward propagation allows for the derivation of the gradient in Formula (4).

While the traditional VAE model uses the decoder to create new samples, this paper employs the encoder to train the entire network using a gradient descent back propagation approach [27]. The hidden variable space is rebuilt to extract higher-level features from the original data, avoiding the issues of dimensional catastrophe and significant computational effort. The mean and variance of h are solved in accordance with a.

Compared to other generative models, such as GAN, VAE has been demonstrated to have more stability and interpretability. Additionally, it can be employed for operations like data compression and feature extraction because to its compressive coding

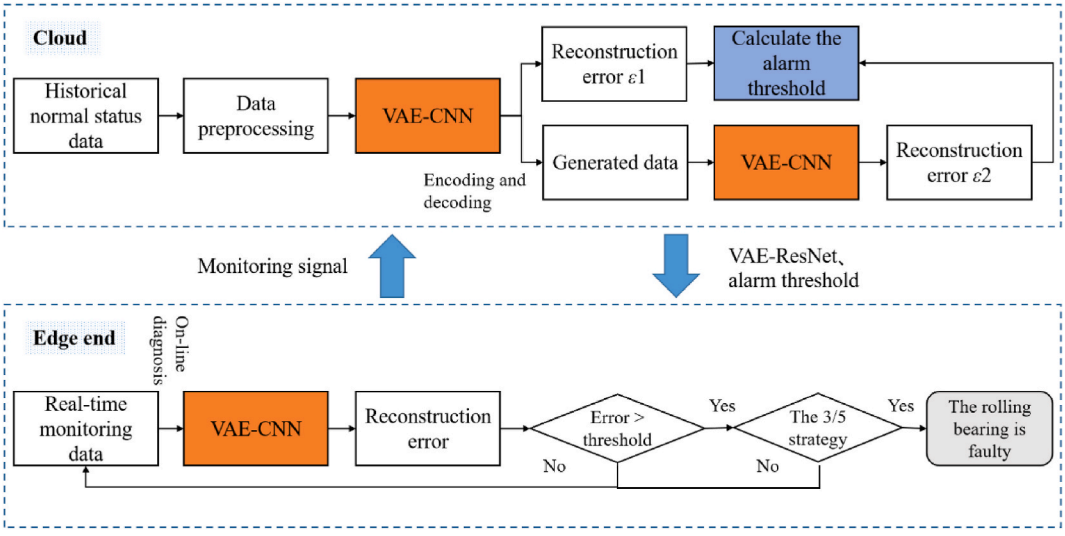


Fig. 3. General framework of the algorithm.

capacity. As a result, the VAE utilized in this study can be used to produce accurate data based on semi-supervised classification, improving the robustness and precision of deep neural network models for diagnosing rolling bearing faults.

3. Deep neural network model with VAE enhancement

Aiming at the challenges faced by existing fault diagnosis methods in rolling bearing fault diagnosis, this paper proposes a deep convolutional neural network fault diagnosis method based on variational autoencoder enhancement on the basis of deep learning technology, hoping to adaptively extract the depth features of vibration signals and accurately identify and judge the fault categories of rolling bearing. In this section, the general framework of the algorithm, the construction of VAE-CNN fault diagnosis model and model training are discussed in detail.

3.1. General framework of the algorithm

More and more academics are calling for the deployment of data-driven defect detection techniques to edge devices in a cooperative cloud platform as edge computing [28], cloud computing [29], artificial intelligence [30], and other technologies improve. Based on this, this research studies integrating VAE with CNN and suggests a VAE-CNN model as an intelligent fault diagnostic model to increase the effectiveness and accuracy of diagnosing rolling bearing faults. Fig. 3 displays the algorithm's overall architecture.

The cloud and the edge end make up the two main components of the rolling bearing problem diagnosis method suggested in this paper, as shown in Fig. 3. The monitoring of sensing data from rolling bearing equipment at the edge, followed by data upload for storage and analysis in the cloud, facilitates collaboration between the cloud and edge. For the purpose of identifying rolling bearing faults, the VAE-CNN intelligent fault detection model is trained in the cloud and subsequently deployed to the edge.

3.2. Construction of VAE-CNN fault diagnosis model

The CNN model [31] and VAE model's benefits are combined in the fault diagnosis model. On the one hand, higher dimensional timing features in vibration signals are thoroughly studied by using CNN to construct time series reliance on vibration signals [32]. However, VAE, a new kind of generative model, can compress the input data into a low-dimensional potential space representation and produce new data samples from it while preserving the characteristics of the original data. As a result, it is currently being used to identify rolling bearing faults and has shown promising results. The VAE-CNN model is employed in this study to model the rolling bearing health state vibration signal by variational inference, learn the distribution of the rolling bearing health state vibration signal, and map it to the hidden space, considerably enhancing the model's resilience. Fig. 4 depicts the structure of the hybrid VAE-CNN model.

The input layer, the encoder, the hidden layer, the decoder, and the output layer are the essential components of the VAE-CNN model, as shown in Fig. 4. The input layer is responsible for the segmentation of the received signal. The encoder is made up of several CNN memory units, which output the mean and variance of the coding vectors after converting the input data into coding vectors in the potential space. The two-dimensional latent distribution is sampled from by the hidden layer, which then outputs the sampled compressed features to the decoder. A three-dimensional sequence is produced by the decoder after it decodes the compressed characteristics. In aim to recover the original time series, the created sequence is concatenated back into the final output layer.

3.3. Model training

(1) Calculation of warning thresholds

The VAE-CNN driver model still functions essentially as a encoder and decoder, encoding and decoding the input data before outputting the results of the reconstruction. The degree of discrepancy between the input data and the reconstructed output results is measured in this paper using the Mean Square Error (MSE) [33], and its calculation formula is shown in Formula (5).

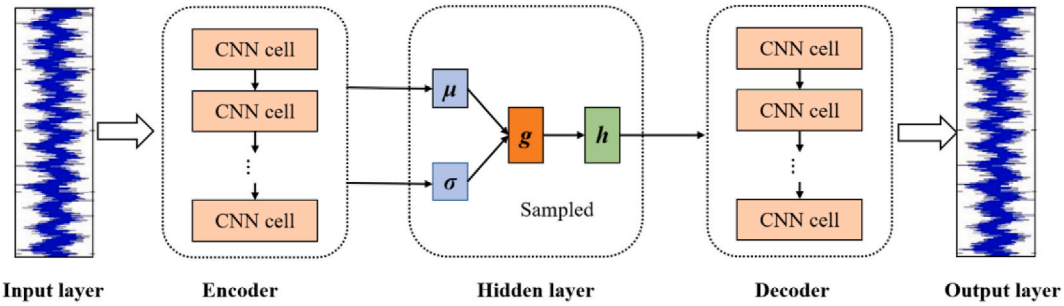


Fig. 4. VAE-CNN enhanced fault diagnosis model.

$$\mathrm{MSE}(\mathrm{Y},\widetilde{\mathrm{Y}}) = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(\widetilde{y}^{i} - y^{i}\right)^{2} \tag{5}$$

Where n represents the total number of samples. y^i represents the original input data of sample i, and $\widetilde{y}^i$ represents the reconstructed sample data. In the context of intelligent fault diagnosis for rolling bearings, the Mean Squared Error (MSE) serves as a critical metric. If the MSE surpasses a pre-established warning threshold, the rolling bearing is deemed defective, prompting an immediate activation of the alarm system. Conversely, if the MSE remains below this threshold, the rolling bearing is considered to be functioning normally, necessitating no immediate action. This determination hinges on the disparity between the original input data and the reconstructed output, as well as the value of the warning threshold. The setting of the warning threshold, typically predefined by operators in most contemporary fault diagnosis methodologies, is crucial. However, the reliance on empirically determined thresholds can be problematic in complex industrial environments. Diverse intrinsic and extrinsic factors render a uniform warning threshold inapplicable across different mechanical systems. An excessively high threshold may delay the activation of an alarm until the rolling bearing failure has advanced significantly, potentially leading to equipment damage or severe accidents. Conversely, a threshold set too low might trigger false alarms, causing unnecessary equipment downtime and hindering production efficiency. To circumvent these challenges, the present study employs an adaptive threshold method, which tailors the warning level for each rolling bearing [34].

The VAE-CNN model's training phase involves calculating the reconstruction error, as denoted by Formula (5). However, due to the paucity of normal samples in hypothetical scenarios, the count of samples for calculating the reconstruction error is limited. Relying solely on these samples for threshold calculation could lead to inaccuracies. The VAE model, a generative network, facilitates the generation of additional data, thereby augmenting the sample pool. The VAE-CNN model, grounded in the VAE network architecture, similarly harnesses this capability to create more data samples. These generated samples, assumed to mimic the healthy state of the rolling bearings, enhance the robustness of the model's learning about normal operational conditions.

The enhanced generative capabilities of the VAE-CNN model facilitate the incorporation of both original and newly generated normal state data in the determination of warning thresholds. This expanded dataset enables more precise and reliable adaptive learning of these thresholds, substantially improving the system's ability to issue timely alerts at the onset of rolling bearing failures. Fig. 5 delineates the detailed calculation process for determining these early warning thresholds, illustrating the methodological sophistication of this approach.

According to Fig. 5: More samples with healthy states are first generated using the trained VAE-CNN augmented model. The newly generated samples are then fed back into the VAE-CNN model, and the reconstructed samples a' are then retrieved after encoding and decoding. Formula (5) is used to determine the MSE between a and a', and the reconstruction error $\varepsilon 2$ is used to denote this error. At last, the mean μ and standard deviation σ of the combination of the reconstruction error $\varepsilon 1$ and the reconstruction error $\varepsilon 2$ are calculated, where the reconstruction error $\varepsilon 1$ is the MSE generated by training the VAE-CNN model. In accordance with the well-known Lajda criterion, the warning threshold in this paper can be placed at + 3. Since the VAE-CNN enhancement model has a one-to-one relationship with the rolling bearing, the warning threshold calculated here also has a one-to-one relationship with the rolling bearing, thus achieving the purpose of adaptivity.

(2) "3/5" strategy

In a typical system, there is a strong possibility that a disruption may trigger the vibration signal to violently jitter and then soon restore to smoothness, resulting in false alarms. This is because industrial systems are easily disturbed by environmental background

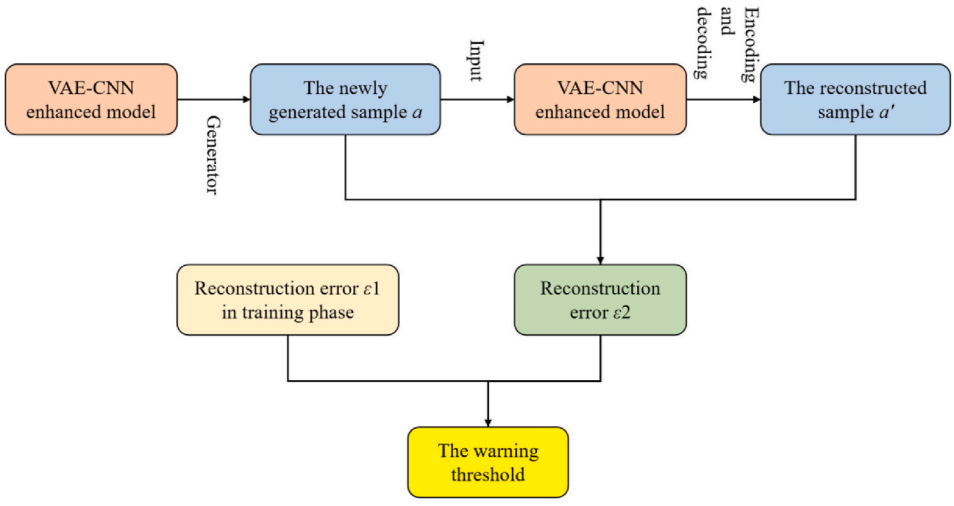


Fig. 5. Example diagram of the calculation process of early warning threshold.

noise and other components in a real industrial environment. This paper employs the "3/5" technique to enhance the accuracy and stability of fault identification and decrease the incidence of false alarms brought on by signal jitters. When three defects appear in five straight diagnoses, this technique suggests that a rolling bearing is actually broken.

The "3/5" strategy's specific implementation concept is as follows: first, construct a sliding window of size 5. The model The algorithm then would check the situation there and save the outcomes of each diagnosis. If there are three failures out of the five diagnostic results, the rolling bearing is deemed to have a real failure, and the system will sound an alarm right away. Otherwise, the first window's diagnosis result will be deleted, the second window's diagnosis result will be stored in the last window, and so on.

(3) Dropout

Large-scale datasets are crucial to the success of deep learning. Large datasets inherently contain more noisy data, which might hinder the training of deep learning models and render their ability to fit data insufficient. Although the new drive model itself has certain robustness to noise, considering that the vibration signal of real industrial production environment is more complex and the deep neural network model is prone to overfitting problem with more training parameters and data noise, the VAE-CNN model is enhanced by introducing the Dropout method to prevent the model from overfitting.

By having the ability to haphazardly alter the network topology of the model itself, the Dropout approach can be used to avoid model overfitting [35]. The main concept is that the neurons in each layer of the network are deactivated with a certain probability during each iteration of the model training process, and the deactivated neurons in the model are not involved in forward and backward propagation, meaning that the weight parameters of the deactivated neurons are not updated during this iteration of training. The deactivated neurons are not involved in this training, but since the model will often undergo multiple iterations of training, they might be included in the subsequent training. The network structure of the model changes during each iteration of training because the neurons in each layer of the network are deactivated with a certain probability, even though the model's parameters are generally the same. As a result, this method can improve the model's ability to generalize and strengthen the trained model.

4. Experimental data and testing

Using the publicly accessible CWRU bearing failure dataset, experiments are carried out to compare various bearing vibration signal dimensions. The experimental findings are then assessed to ascertain the efficacy of the suggested method for the diagnosis of rolling bearing failure signals.

4.1. Experimental data and environment

In this study, simulation trials were conducted using CWRU bearing data [36]. The drive-side data utilized in the trials were gathered intentionally using the EDM technique to generate flaws on the bearing's inner ring, outer ring, and rolling element. The tests mainly used fault data of 0.007 inch, 0.014 inch, and 0.021 inch. The motor load range is from 1 HP to 3 HP, and the test bearing model is SKF 6205. Because the deep learning approach can efficiently handle complex data without manual feature extraction when dealing with multidimensional data, 10 types of data were selected for the experiment, including 9 types of fault data at different positions and speeds and 1 vibration signal under normal conditions. Table 1 displays the fundamental data from the tests. In the experiment, the rotational speeds are set at 1772, 1750, and 1730 rpm, with a sampling frequency of 12 kHz for the bearing data. A moving window of length 1024 is employed to capture the samples, resulting in each sample containing 1024 data points. The dataset is partitioned into training and test sets with a ratio of 6:4. To make it easier for the network model to be trained, these 10 types of data are named L1 through L10.

The experimental environment was configured with a 12th Gen Intel(R) Core(TM) i5-12400F 2.50 GHz processor and NVIDIA GeForce GTX 1650 graphics card, and the experiments were conducted under Python 3.7 and Pytorch 1.9.1.

Table 1
CWRU bearing dataset.

Label	Fault type	Fault size/inches	Load
L1	R07	0.007	1,2,3
L2	R14	0.014	1,2,3
L3	R21	0.021	1,2,3
L4	I07	0.007	1,2,3
L5	I14	0.014	1,2,3
L6	I21	0.021	1,2,3
L7	O07	0.007	1,2,3
L8	O14	0.014	1,2,3
L9	O21	0.021	1,2,3
L10	Normal	/	1,2,3

4.2. Experimental results and analysis

Experiment 1. Performance testing of different failure types

Comparative experiments employing bearing failure data were carried out using various methods and speeds to examine the effectiveness of the suggested network. The trial outcomes are depicted in Fig. 6.

The diagnosis accuracy of the VAE-CNN model developed in this paper for different fault types at different rotational speeds typically reaches more than 90 %, and it achieves a generally acceptable diagnosis result, as shown by the diagnosis results in Fig. 6(a) to 6(c). This demonstrates that the new model can be utilized under different rotational speed conditions and has strong robustness and application in identifying rolling bearing faults. This enhancement arises from the synergistic integration of the CNN and VAE models, leveraging their respective strengths. The CNN model excels in extracting deep temporal features from the vibration signals and establishing temporal dependencies within these signals. Concurrently, the VAE model preserves the intrinsic properties of the original data while compressing the input into a low-dimensional latent space representation. Additionally, the VAE can generate new data samples from this latent space, thereby ensuring the model's diagnostic accuracy.

Three additional deep neural network models, CNN, AlexNet [37] and ResNet [38] were chosen for comparison studies in order to evaluate the diagnostic performance of the proposed technique on rolling bearing fault data. The outcomes are displayed in Fig. 7.

Fig. 7 shows that at various rotational speeds, the fault diagnostic accuracy of three typical deep neural networks, CNN, AlexNet, and ResNet, is comparable. However, the VAE-CNN model dramatically enhances the identification of rolling bearing faults and has a significantly greater diagnostic accuracy than the other three models. The new model, which does not rely on the labeled dataset or the expert's empirical knowledge and can learn the deep features in the vibration signal directly, combines the benefits of the CNN's good expression of signal data with VAE's sensitivity of data noise. This could be the underlying cause of the analysis.

5. Experiment 2: experimental test under noise-added environment

Noise from vibration and component-to-component friction is unavoidable in the environment where rolling bearings operate. However, adding noise might make vibration data invalid, which can reduce the precision of fault identification. To imitate noise and other disturbances during the experiment, white noise is introduced to the raw data that was collected. The CWRU data set is subjected

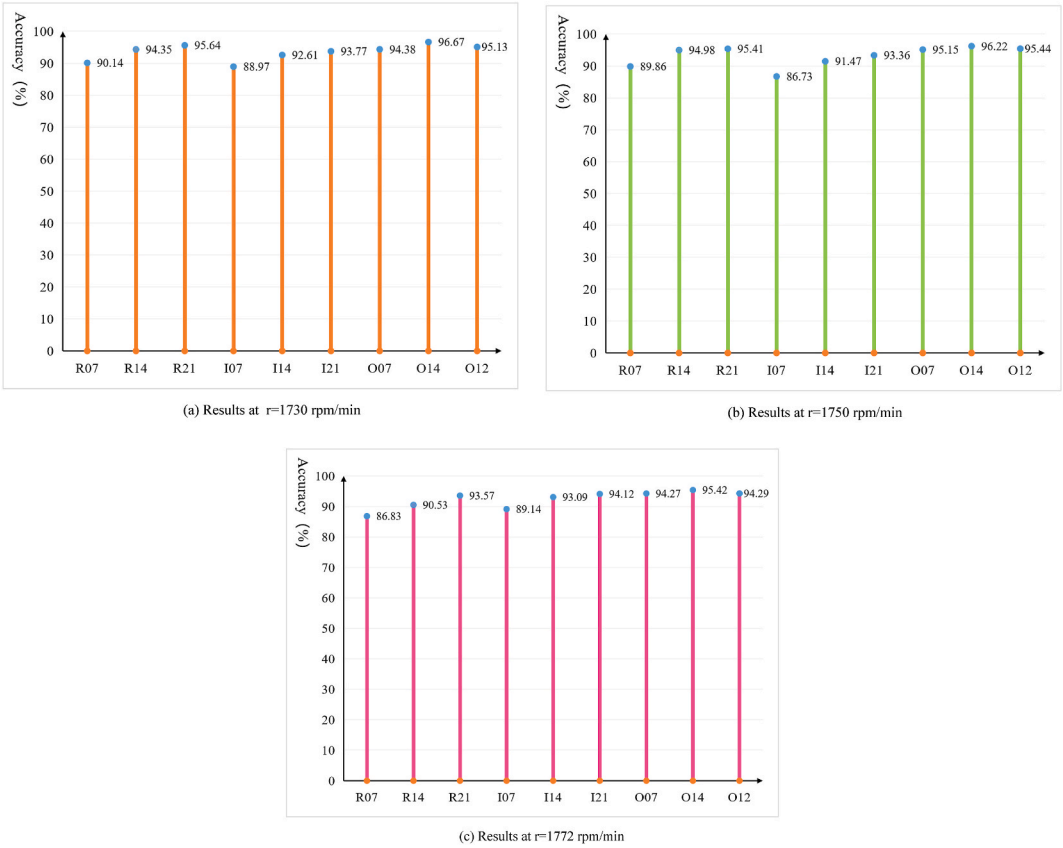


Fig. 6. Diagnostic results at different speeds.

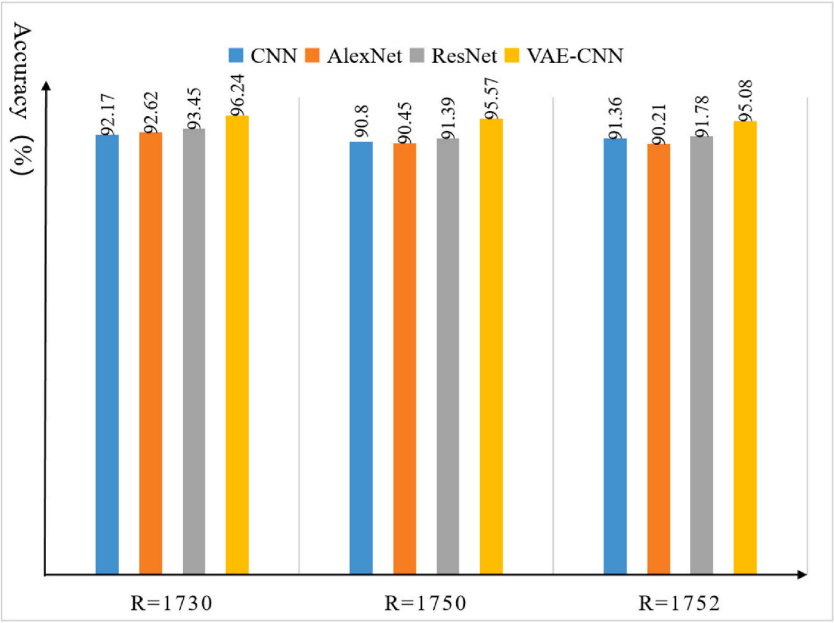


Fig. 7. Bearing fault diagnosis performance at different speeds.

to several procedures and noise levels of −3, −1, 1, 3, and 5 dB are applied. Using the average accuracy reported in Fig. 8.

The accuracy of the bearing diagnosis based on the VAE-CNN model is higher than the other approaches in various noise situations, as can be observed in Fig. 8. This aspect demonstrates the great noise immunity of the suggested technique. This is primarily because the suggested method gets better diagnostic results and exhibits greater noise immunity even under various noise environments by making full use of the VAE model's superior robustness to noise compared to other deep neural network models.

The VAE-CNN augmented fault diagnosis model, after experimental validation in various dimensions, can achieve more satisfactory diagnosis results for various fault types compared to several representative deep neural network models without VAE augmentation, and is therefore applicable to a variety of fault diagnosis tasks. The augmented model also solves the shortcomings of conventional fault diagnosis models, such as weak vibration signal feature extraction, reliance on large-scale datasets with labels and expertise, and greatly enhances the applicability and robustness of the model, which can diagnose bearing faults under high noise and multiple

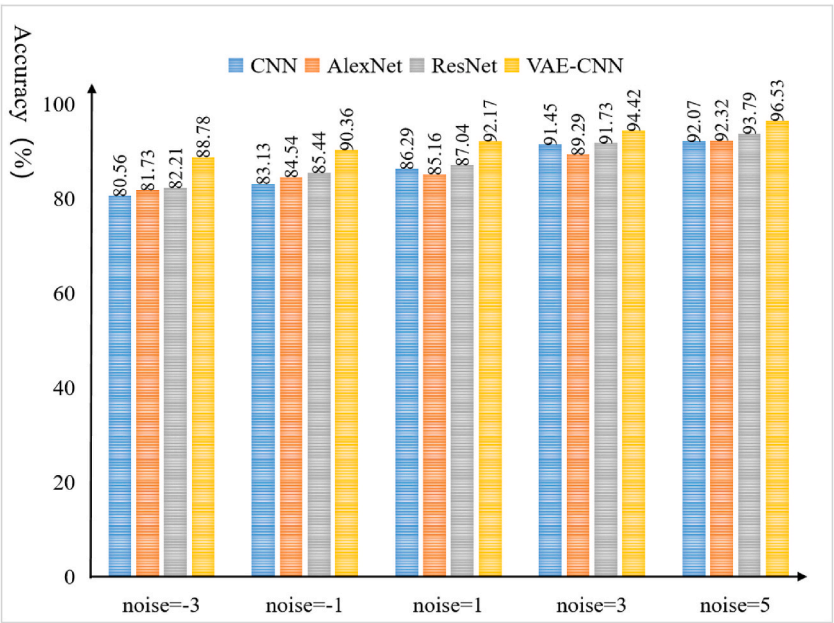


Fig. 8. Bearing fault diagnosis performance under different noise conditions.

redundancies.

6. Conclusion

This manuscript elaborates on the development of a Variational Autoencoder-Convolutional Neural Network (VAE-CNN) model, designed for the fault diagnosis of rolling bearings. This innovative model aims to address the limitations commonly encountered in data-driven fault diagnosis models, particularly their deficiencies in extracting features from vibration signals and their dependency on large-scale, labeled datasets and expert knowledge. By amalgamating the CNN model's superior capacity for representing vibration signal data with the VAE model's robustness to data noise, the proposed VAE-CNN model excels in scenarios where only a limited amount of observational data is available at the initial stages of rolling bearing operation. It is uniquely capable of directly learning the deep characteristics of vibration signals, independent of labeled datasets or empirical expertise from specialists. The VAE-CNN model achieves over 90 % accuracy for diagnosing different fault types at various speeds. Experimental validation using the CWRU fault dataset has demonstrated the model's commendable performance across various metrics, highlighting its robustness and practical applicability in real-world settings.

The advent of artificial intelligence heralds the inevitable progression towards intelligent diagnosis of rolling bearing faults. However, the approach presented in this study is not without its limitations. Firstly, the implementation of VAE necessitates careful selection of appropriate prior distributions and loss functions, alongside the fine-tuning of hyperparameters, to enhance the model's performance. Consequently, there is a need for further development in the algorithms and theoretical underpinnings of VAE to establish a more efficacious, reliable, and broadly applicable generative model. Secondly, the experimental validation in this study primarily utilized laboratory datasets, which may not fully replicate the complexities of actual industrial conditions. Thus, the fidelity of the bearing failure data from real industrial environments remains an open question, necessitating future research to validate the proposed strategy using data derived from genuine operational conditions. Lastly, the current VAE-CNN model can only be used for intelligent diagnosis of a single type or category of rolling bearing faults, indicating that the model's practicality needs further improvement. Future research could explore the development of a joint fault diagnosis model for different types of bearings.

Data availability statement

The data is available from the corresponding author upon request.

Funding statement

The research in this article has not received funding or project support.

CRediT authorship contribution statement

Yu Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software. **Dexiong Li:** Visualization, Software, Resources, Project administration. **Lei Li:** Supervision, Project administration, Methodology, Data curation, Conceptualization. **Runde Sun:** Writing – original draft, Visualization, Software, Resources. **Shuqing Wang:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: huqing Wang reports administrative support, article publishing charges, and writing assistance were provided by Department of Electrical Engineering, Shijiazhuang Institute of Railway Technology. Shuqing Wang reports a relationship with Department of Electrical Engineering, Shijiazhuang Institute of Railway Technology that includes: employment, non-financial support, and travel reimbursement. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Javaid, A. Haleem, R.P. Singh, et al., Artificial intelligence applications for Industry 4.0: a literature-based study, Journal of Industrial Integration and Management 7 (1) (2022) 83–111, https://doi.org/10.1142/S2424862221300040.

[2] Wentao Qiu, Bing Wang, Xiong Hu, Rolling bearing fault diagnosis based on RQA with STD and WOA-SVM, Heliyon 10 (4) (2024) e26141, https://doi.org/10.1016/j.heliyon.2024.e26141.

[3] M. Mansouri, R. Fezai, M. Trabelsi, et al., Fault diagnosis of wind energy conversion systems using Gaussian process regression-based multi-class Random Forest, IFAC-PapersOnLine 55 (6) (2022) 127–132, https://doi.org/10.1016/j.ifacol.2022.07.117.

[4] N. Shabbir, B. Asad, M. Jawad, et al., Spectrum analysis for condition monitoring and fault diagnosis of ventilation motor: a case study, Energies 14 (7) (2021) 1–16, https://doi.org/10.3390/en14072001.

[5] R. Bodile, T. Rao, Adaptive filtering of electrocardiogram signal using hybrid empirical mode decomposition-Jaya algorithm, J. Circ. Syst. Comput. 30 (12) (2021) 2150209.1–212150209, https://doi.org/10.1142/S0218126621502091.

[6] A. Yilmaz, G. Bayrak, A new signal processing-based islanding detection method using pyramidal algorithm with undecimated wavelet transform for distributed generators of hydrogen energy, Int. J. Hydrogen Energy 47 (45) (2022) 19821–19836, https://doi.org/10.1016/j.ijhydene.2022.03.114.

[7] H. Habbouche, Y. Amirat, T. Benkedjouh, et al., Bearing fault event-triggered diagnosis using a variational mode decomposition-based machine learning approach, IEEE Trans. Energy Convers. 37 (1) (2022) 466–474, https://doi.org/10.1109/TEC.2021.3085909.
[8] Y. Huangfu, E. Seddik, S. Habibi, et al., Fault detection and diagnosis of engine spark plugs using deep learning techniques, SAE International Journal of Engines 15 (4) (2022) 515–525, https://doi.org/10.4271/03-15-04-0027.
[9] R. Baessler, T. Baessler, M. Kley, Classification of load and rotational speed at wire-race bearings using Convolutional Neural Networks with vibration spectrograms, Technisches Messen: Sensoren, Gerate, Systeme 89 (5) (2022) 352–362, https://doi.org/10.1515/teme-2021-0143.
[10] I. Moumene, N. Ouelaa, Gears and bearings combined faults detection using optimized wavelet packet transform and pattern recognition neural networks, Int. J. Adv. Des. Manuf. Technol. 120 (7/8) (2022) 4335–4354, https://doi.org/10.1007/s00170-022-08792-2.
[11] H. Nakamura, K. Asano, S. Usuda, et al., A diagnosis method of bearing and stator fault in motor using rotating sound based on deep learning, Energies 14 (5) (2021) 1–15, https://doi.org/10.3390/en14051319.
[12] P. Kumar, A.S. Hati, Convolutional neural network with batch normalization for fault detection in squirrel cage induction motor, IET Electr. Power Appl. 15 (1) (2021) 39–50, https://doi.org/10.1049/elp2.12005.
[13] A. Choudhary, T. Mian, S. Fatima, Convolutional neural network-based bearing fault diagnosis of rotating machine using thermal images, Measurement 176 (4) (2021) 109196, https://doi.org/10.1049/elp2.12005.
[14] O. Gueltekin, E. Cinar, K. Oezkan, et al., Multisensory data fusion-based deep learning approach for fault diagnosis of an industrial autonomous transfer vehicle, Expert Syst. Appl. 200 (Aug) (2022) 117055.1–11117055, https://doi.org/10.1016/j.eswa.2022.117055.
[15] W. Kou, D.A. Carlson, A.J. Baumann, et al., A deep-learning-based unsupervised model on esophageal manometry using variational autoencoder, Artif. Intell. Med. 112 (Feb) (2021) 102006, https://doi.org/10.1016/j.artmed.2020.102006.
[16] S.J. Bang, M.J. Kang, M.G. Lee, et al., STO-CVAE: state transition-oriented conditional variational autoencoder for data augmentation in disability classification, Complex & Intelligent Systems 10 (3) (2024) 4201–4222, https://doi.org/10.1007/s40747-024-01370-x.
[17] D.D. Chakladar, S. Datta, P. Roy, et al., Cognitive workload estimation using variational autoencoder and attention-based deep model, IEEE Transactions on Cognitive and Developmental Systems 15 (2023) 581–590, https://doi.org/10.1109/TCDS.2022.3163020.
[18] J. Lee, M. Kim, U.K. Jin, et al., Asymmetric inter-intra domain alignments (AIIDA) method for intelligent fault diagnosis of rotating machinery, Reliab. Eng. Syst. Saf. 218 (Feb. Pt. B) (2022) 108186.1–10108186, https://doi.org/10.1016/j.ress.2021.108186.
[19] A. Althubaiti, Fault diagnosis and health management of bearings in rotating equipment based on vibration analysis - a review, Journal of Vibroengineering 24 (1) (2021) 46–74, https://doi.org/10.21595/jve.2021.22100.
[20] M. Gupta, R. Wadhvani, A. Rasool, A real-time adaptive model for bearing fault classification and remaining useful life estimation using deep neural network, Knowl. Base Syst. 259 (Jan.10) (2023) 1–22, https://doi.org/10.1016/j.knosys.2022.110070.
[21] B.G. Basher, A. Ghanem, S. Abulanwar, et al., Fault classification and localization in microgrids: leveraging discrete wavelet transform and multi-machine learning techniques considering single point measurements, Elec. Power Syst. Res. 231 (2024), https://doi.org/10.1016/j.epsr.2024.110362.
[22] V.R. Kumar, P.A. Jeyanthy, R. Kesavamoorthy, Optimization-assisted CNN model for fault classification and site location in transmission lines, Int. J. Image Graph. 24 (1) (2024) 24500–24508, https://doi.org/10.1142/S0219467824500086.
[23] S.J. Bang, M.J. Kang, M.G. Lee, et al., STO-CVAE: state transition-oriented conditional variational autoencoder for data augmentation in disability classification, Complex & Intelligent Systems 10 (3) (2024) 4201–4222, https://doi.org/10.1007/s40747-024-01370-x.
[24] K. Yan, X. Chen, X. Zhou, et al., Physical model informed fault detection and diagnosis of air handling units based on transformer generative adversarial network, IEEE Trans. Ind. Inf. 19 (2) (2023) 2192–2199, https://doi.org/10.1109/TII.2022.3193733.
[25] J.I. Monroe, V.K. Shen, Learning efficient, collective Monte Carlo moves with variational autoencoders, J. Chem. Theor. Comput.: JCTC 18 (6) (2022) 3622–3636, https://doi.org/10.1021/acs.jctc.2c00110.
[26] B.P.M. Laevens, F.P. Pijpers, H.J. Boonstra, et al., A Markov chain Monte Carlo approach for the estimation of photovoltaic system parameters, Sol. Energy 265 (Nov) (2023) 1–20, https://doi.org/10.1016/j.solener.2023.112132.
[27] I. Abuqaddom, B. Mahafzah, H. Faris, Oriented stochastic loss descent algorithm to train very deep multi-layer neural networks without vanishing gradients, Knowl. Base Syst. 230 (2021) 107391, https://doi.org/10.1016/j.knosys.2021.107391.
[28] T. Lee, Research trend on edge computing based on keyword frequency, centrality analysis and social network analysis: focusing on United States, United Kingdom, South Korea, Journal of the Korea Contents Association (2023), https://doi.org/10.5392/jkca.2023.23.03.076.
[29] C. Jyoti, Z. Efpraxia, Understanding and exploring the value co-creation of cloud computing innovation using resource based value theory: an interpretive case study, J. Bus. Res. (2023), https://doi.org/10.1016/j.jbusres.2023.113970.
[30] J. Luo, M. Pan, K. Mo, et al., Emerging role of artificial intelligence in diagnosis, classification and clinical management of glioma, Semin. Cancer Biol. 91 (2023) 110–123, https://doi.org/10.1016/j.semcancer.2023.03.006.
[31] M. Zhou, N. Kazemi, P. Musilek, Distribution grid fault classification and localization using convolutional neural networks, Smart Grids and Sustainable Energy 9 (1) (2024), https://doi.org/10.1007/s40866-024-00205-5.
[32] D.R. Wijaya, R. Sarno, E. Zulaika, DWTLSTM for electronic nose signal processing in beef quality monitoring, Sensors and Actuators B Chemical 326 (2021) 128931, https://doi.org/10.1016/j.snb.2020.128931.
[33] P.N. Malleswari, C.H. Bindu, K.S. Prasad, An improved denoising of electrocardiogram signals based on wavelet thresholding, Journal of Biomimetics Biomaterials and Biomedical Engineering 51 (2021) 117–129.
[34] K.S. Schairer, D.B. Putterman, D.H. Keefe, et al., Automated adaptive wideband acoustic reflex threshold estimation in normal-hearing adults, Ear Hear. 43 (2) (2022) 370–378, https://doi.org/10.1097/AUD.0000000000001102.
[35] A. Lemay, K. Hoebel, C.P. Bridge, et al., Improving the repeatability of deep learning models with Monte Carlo dropout 174 (2022) 1–11, https://doi.org/10.48550/arXiv.2202.07562.
[36] J. Hendriks, P. Dumond, D.A. Knox, Towards better benchmarking using the CWRU bearing fault dataset, Mech. Syst. Signal Process. 169 (2) (2022) 108732, https://doi.org/10.1016/j.ymssp.2021.108732.
[37] S. Kollem, K.R. Reddy, C.R. Prasad, et al., AlexNet-NDTL: classification of MRI brain tumor images using modified AlexNet with deep transfer learning and Lipschitz-based data augmentation, Int. J. Imag. Syst. Technol. 33 (4) (2023) 1306–1322, https://doi.org/10.1002/ima.22870.
[38] S. Asel, K. Mo, K. Andrii, et al., Hybrid quantum ResNet for car classification and its hyperparameter optimization, Quantum Machine Intelligence 5 (2) (2023) 38–54, https://doi.org/10.1007/s42484-023-00123-2. DOI.