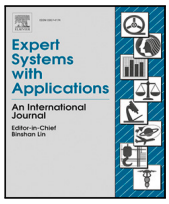
Expert Systems With Applications 277 (2025) 127248

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



ISONet: Reforming 1DCNN for aero-engine system inter-shaft bearing fault diagnosis via input spatial over-parameterization

Qian Xiang [a], Xiaodan Wang [b,*], Yafei Song [b], Lei Lei [c]

[a] Laboratory of Intelligent Control, PLA Rocket Force University of Engineering, Xi'an, 710025, China
[b] College of Air and Missile Defense, Air Force Engineering University, Xi'an, 710051, China
[c] College of Information and Navigation, Air Force Engineering University, Xi'an, 710077, China

ARTICLE INFO

Dataset link: https://github.com/HouLeiHIT/HIT-dataset

Keywords:
Aero-engine system
Fault diagnosis
One-dimensional convolutional neural networks
Over-parameterization
Preconditioning scheme

ABSTRACT

Data-driven neural networks have risen as avant-garde approaches to fault diagnosis. However, recent studies have indicated that traditional one-dimensional convolutional neural networks (1DCNNs) exhibit inadequate performance in diagnosing faults within the aero-engine system's inter-shaft bearings. Addressing this, we introduces a novel 1DCNN variant, termed ISONet, which incorporates over-parameterization in the input spatial dimension to enhance the training landscape of 1DCNNs. Theoretically, we prove that this over-parameterization is akin to implicitly integrating a specific preconditioning scheme from a dynamic differential perspective. This preconditioning facilitates optimization movement along previously traversed directions, effectively acting as an acceleration procedure that amalgamates momentum with adaptive learning rates. Empirical validation confirms the irreplaceability of this implicit acceleration mechanism, demonstrating its capacity to further augment the convergence speed and stability of training, even when advanced gradient descent optimizers with explicit momentum and adaptive learning rates are employed. During the training phase, ISONet does not significantly increase training time, and it can be folded into a standard 1DCNN during the testing phase. The efficacy of the proposed ISONet architecture is substantiated through empirical testing utilizing real-world vibration data from an aero-engine test rig. Comparative analyses indicate that ISONet surpasses current state-of-the-art deep learning models, particularly under various limited-sample conditions.

1. Introduction

The aero-engine, as the heart of modern aviation, plays a pivotal role in ensuring flight safety and efficiency. The inter-shaft bearing, an integral component within the aero-engine system, is critical for maintaining smooth operation and transmitting power between rotating shafts. However, due to the extreme operating conditions, including high speeds, temperatures, and loads, these bearings are susceptible to various faults that can lead to catastrophic failures if not detected promptly (Berghout, Bentrcia, Lim, & Benbouzid, 2023; Fei et al., 2024; Kang, Cao, Hou, & Chen, 2022; Wang et al., 2024). Therefore, the development of advanced diagnostic techniques for the accurate detection and classification of inter-shaft bearing faults is of paramount importance.

Traditional fault diagnosis methods often rely on signal processing techniques such as Fast Fourier Transform analysis (Wang et al., 2024) or statistical features extraction combined with machine learning algorithms (Hou, Zhao et al., 2022a; Tian et al., 2020; Tian, Zhang, Zhang, Ai, & Wang, 2023). While these approaches have been effective to some extent, they are limited by their reliance on handcrafted features and may not generalize well to new data or under varying operational conditions. The application of deep learning (DL) in diagnosing faults in aircraft engine bearings holds tremendous promise and vast potential. Leveraging the computational prowess of DL transforms the approach to detecting and pinpointing issues in these systems, thereby aiding in the rapid and precise analysis of a significant amount of data. Ultimately, DL has the capability to bolster aircraft safety, dependability, and performance by facilitating proactive maintenance and forestalling disastrous malfunctions (Berghout et al., 2023). For instance, Hou, Wang et al. (2022b) proposed a Siamese multiscale residual feature fusion network (SMSRFFN) tailored for aero-engine bearing fault diagnosis under small-sample conditions, SMSRFFN enhanced fault feature extraction, fusion, and identification, demonstrating superior diagnostic accuracy compared to state-of-the-art small-sample diagnostic techniques. Zuo, Zhang, Zhang, Luo, and Liu (2021) introduced a spiking neural network tailored for bearing fault diagnosis in rotating machinery. The method, which incorporates time into its model, mimics

* Corresponding author.
E-mail addresses: qianxljp@126.com (Q. Xiang), afeu_wang@163.com (X. Wang), yafei_song@163.com (Y. Song), wendyandpaopao@163.com (L. Lei).

https://doi.org/10.1016/j.eswa.2025.127248
Received 13 July 2024; Received in revised form 13 February 2025; Accepted 10 March 2025
Available online 17 March 2025

natural neural networks and demonstrates promising accuracy in diagnosing bearing faults. Berghout et al. (2023) proposed a comprehensive DL approach for diagnosing faults in aircraft engine bearings, addressing data complexity and drift through structured preprocessing and LSTM-based adaptive learning. Fu et al. (2023) proposed EdgeCog, a real-time failure diagnosis method designed for lightweight embedded kernels. EdgeCog deployed an optimized DL model on a miniaturized microcontroller platform, utilizing an attention mechanism for self-adaptive feature optimization. Through knowledge distillation and quantization, EdgeCog achieved high real-time performance while enhancing data security and diagnosis efficiency, making it suitable for cost-effective edge computing in distributed scenarios.

In recent years, convolutional neural networks (CNNs), particularly one-dimensional CNNs (1DCNNs), have emerged as powerful tools for fault diagnosis in mechanical systems (Che, Zhang, Wang, & Xiong, 2024; Guo, Yang, Li, Dai, & Huang, 2024; Sun et al., 2024). These networks have shown exceptional capabilities in automatically extracting relevant features from raw data, making them suitable for complex signal processing tasks. However, the performance of 1DCNNs in aero-engine fault diagnosis is relatively low, especially concerning the intricacies associated with inter-shaft bearing faults (Hou et al., 2023).

The optimization of neural network training has long been a central theme in enhancing the performance of these systems. In the context of 1DCNNs, 1D convolution (1D-Conv) layers are the fundamental components. Building upon this foundation, we propose to enhance a 1D-Conv layer by incorporating additional parameters, specifically within the input spatial dimension of the 1D-Conv, a process we term input spatial over-parameterization, resulting in an input spatial over-parameterized 1D-Conv (1D-ISOConv). This augmentation introduces a set of learnable parameters through the composition of two tensors, thereby constituting over-parameterization. A significant advantage of input spatial over-parameterization is that the multi-layer composite linear operations, introduced by the over-parameterization, can be condensed into a compact single-layer representation during the testing phase. Consequently, only a single layer is operational during inference, minimizing computations to an extent equivalent to that of a standard 1D-Conv layer.

On the basis of 1D-ISOConv, we have constructed the ISONet to address the current shortcomings of 1DCNNs in diagnosing faults within the aero-engine system's inter-shaft bearings. More importantly, we have provided a theoretical explanation for the optimization mechanism of input spatial over-parameterization from the perspectives of matrix computation and differential dynamics, offering solid theoretical support for the superior performance of ISONet. Utilizing real-world measured data, we have demonstrated the advanced capabilities of ISONet, especially under conditions of limited training samples, for aero-engine system inter-shaft bearing fault diagnosis. The experimental results highlight the transformative potential of ISONet as a tool for bolstering the reliability and efficacy of aero-engine diagnostic systems.

The main contributions of this paper are as follows:

1. We proposed a novel 1D-Conv, *a.k.a.*, 1D-ISOConv, which enhances the standard 1D-Conv in the input spatial dimensions. Specifically, 1D-ISOConv is over-parameterized in the training phase, which significantly improves the training process. However, during the testing phase, 1D-ISOConv is equivalent to 1D-Conv without any additional complexity. This innovative design enables the proposed 1D-ISOConv to achieve superior performance in aero-engine inter-shaft bearing fault diagnosis tasks while maintaining computational efficiency.
2. Based on 1D-ISOConv, we proposed a new variant of 1DCNN, namely ISONet, and explore its applicability in aero-engine system inter-shaft bearing fault diagnosis. By leveraging the unique strengths of ISONet, we have demonstrated its effectiveness in accurately diagnosing bearing faults.
3. We theoretically explored the rationale of the novel 1D-ISOConv, which introduces an over-parameterized structure that implicitly confers momentum acceleration and adaptive learning rate properties to the training process. We have demonstrated its capacity to further augment the convergence speed and training stability when advanced gradient descent optimizers with explicit momentum and adaptive learning rates are employed.
4. We validate the superiority of the proposed ISONet using real-world vibration data collected from an aero-engine test rig (Hou et al., 2023). The results are compared with those obtained from state-of-the-art DL models to demonstrate the superiority of ISONet under various limited-sample conditions. Additionally, we have also tested the impact of various optimizers and activation functions on ISONet.

The organization of this paper is structured as follows: Section 2 offers a retrospective review of the existing scholarly works. In Section 3, we delve into the detailed architecture and properties of 1D-ISOConv and ISONet, elucidating their unique features and functionalities. Section 4 is devoted to the experimental setup, where we outline the conditions and parameters used in our experiments. Furthermore, we present the experimental results, discussing their significance and implications. Finally, in Section 5, we summarize the key findings and contributions of this paper, highlighting the importance of our work and its potential future search direction.

2. Related works

2.1. Aero-engine system inter-shaft bearing fault diagnosis

In the scholarly discourse of aero-engine system inter-shaft bearing fault diagnosis, the meticulous examination of vibration signals is recognized as an essential analytical approach. The foundational work by Yang, Zhang, and Chen (2022) emphasizes the pivotal role of high-speed rotor dynamics, utilizing symmetrical bearing-rotor models coupled with envelope analysis to decipher the underlying dynamics, which has been substantiated through empirical studies. Gao, Yuan, Liu, Cao, and Sun (2024) delve into the "paroxysmal impulse vibration" observed in inter-shaft bearings, presenting predictive models that harmonize experimental findings with simulation data, thereby deepening the comprehension of the underlying vibration mechanisms. Yu, Fang, Chen, and Cong (2023) introduce a multifaceted algorithm, the CCFWT-SVD-Katz, integrating wavelet transform, singular value decomposition, and Katz fractal dimension analysis, among others, to achieve signal denoising, enhancement, and the extraction of fault features.

In addition to the examination of vibration signals, the domain of aero-engine system inter-shaft bearing fault diagnosis has also witnessed advancements based on machine learning methodologies. Hou, Zhao et al. (2022a) introduce a pioneering feature extraction strategy, adept at mitigating the impact of noise through the application of Laplace wavelets, orthogonal matching pursuit, and sparse representation theory, effectively identifying the transient shock components indicative of bearing faults. Zhang, Ji, Huang, and Lou (2021) present a method based on canonical correlation analysis for the extraction of specific fault signals from multi-channel observations, particularly for the complex task of diagnosing compound faults in aero-engine spindle bearings, enhancing the convergence and reliability of the diagnosis process. Yu, Pan, Meng, and Chen (2021) enhance the diagnosis of compound faults in rolling bearings through the combined application of Intrinsic Time-scale Decomposition and Singular Value Decomposition, followed by Hilbert spectrum envelope analysis, which offers improved noise control and precision in characteristic frequency extraction.

Besides, methodologies grounded in DL have further propelled the progress in diagnosing inter-shaft bearing faults. Wang, Li et al. (2022b) advance the state of the art with the development of the

Improved Spiking Neural Network (SNN), demonstrating superior diagnostic efficacy in comparison to traditional SNNs for the detection of inter-shaft bearing faults. Liu, Chen, Cheng, Wei, and Wang (2022) propose a data-driven method for predicting the remaining useful life of aero-engine bearings, employing a deep CNN to extract evolutionary features from normal-stage vibration data, coupled with a particle filter algorithm for tracking and predicting life, which outperforms RMS values in terms of prediction accuracy and stability. Hou, Wang et al. (2022b) propose the Siamese Multiscale Residual Feature Fusion Network, a novel technique for fault diagnosis under small-sample conditions, overcoming traditional limitations by employing multiscale residual feature extraction and attention mechanisms for improved diagnostic performance. Wang, Xu et al. (2022a) propose a fault diagnosis method for planetary gearboxes that employs time–frequency representation and deep reinforcement learning, framing fault identification as a sequential decision-making process and achieving diagnostic accuracy exceeding 99.5% across various operating conditions. Liu, Chen, Hao, and Pan (2023) explore the integration of LSTM networks with deep CNNs, crafting a hybrid model that surpasses the predictive accuracy of individual deep CNNs and LSTMs.

Recently, Hou et al. (2023) introduce the HIT benchmark dataset for aero-engine inter-shaft bearing fault diagnosis, which, due to its complexity and realism, poses a significant challenge to existing diagnostic methods and advances the frontier of predictive maintenance in aircraft operations. The use of CNN, LSTM, and TST on the HIT dataset by Hou et al. (2023) yielded average accuracies of 83.13%, 85.41%, and 71.07%, respectively. Subsequently, Berghout (Berghout et al., 2023) improved upon these results by employing a structured data preprocessing strategy and LSTM models with recursive weights initialization, achieving an average accuracy of 92.03% on the HIT dataset. The ongoing pursuit of higher diagnostic accuracy in the field of aero-engine system inter-shaft bearing fault diagnosis necessitates the further exploration of innovative approaches. In this paper, we propose a novel input spatial over-parameterized 1DCNN, demonstrating an average classification accuracy of over 99% on the HIT dataset with the same data split strategy.

2.2. Recent advances on 1DCNNs

In the evolving landscape of artificial intelligence, 1DCNNs have been the beneficiaries of a plethora of enhancements, extending their reach and efficacy across a myriad of domains. This section endeavors to provide a succinct overview, cataloging the principal strategies that have been pivotal in advancing the state of 1DCNNs. These strategies encompass refinements to the Conv operations themselves, the introduction of novel architectural elements such as attention mechanisms and residual structures, and employ novel training strategies.

One of the key advancements in the domain has been the transformation of traditional Conv operations. MSGF-1DCNN (Xiang, Wang, Lai, Song, Li et al., 2022) is a testament to the innovation in feature extraction. By employing multi-scale group 1D-Conv, this model efficiently extracts features across various scales, integrating them through point-wise convolution. MSGF-1DCNN has been demonstrated to enhance recognition accuracy while significantly reducing model parameters, making it over 2.4 times more efficient than standard 1DCNNs. Addressing the computational complexity head-on, the GFAC-1DCNN (Xiang et al., 2023a) incorporates group convolution and a linear fusion layer. This not only enhances feature integration and recognition accuracy but also enriches the model's capability to capture hierarchical features through the inclusion of layer-wise auxiliary classifiers. This has led to an improvement in overall testing accuracy and recall rates.

As for new 1DCNN architecture design, GCK-MSSC (Xiao, Zhao, Zhou, Ou, & Huang, 2024) stands out with its departure from traditional convolutional kernels, replacing them with gate convolutional kernels for dynamic weight adjustment. The integration of a one-dimensional global attention mechanism further bolsters the model's robustness and accuracy in fault detection tasks. DSFCNN (Xiang et al., 2024) has made significant strides in alleviating the computational burden associated with traditional CNNs. Coupled with a sample-fitting and class-distinguishability loss, this model enhances recognition accuracy under limited-sample conditions. The innovative quadruplet DSFCNN refines this approach during training, while a parameter quantity shifting-fitting performance coordinate system offers valuable insights into the model's efficiency and effectiveness. GGRU-1DCNN-AdaBN (Sun et al., 2024) integrates an improved gap-gated recurrent unit with 1DCNNs and adaptive batch normalization, enhancing feature extraction and model adaptability under variable conditions. The incorporation of a global average pooling layer and the fusion of multi-type features through 1DCNN improve diagnostic accuracy while reducing training parameters. Dong, Zhao, and Cui (2024) proposed a hybrid model that integrates a 1D residual network, which adeptly mitigates gradient-related issues. This model is further augmented with an attention mechanism that adeptly highlights critical fault characteristics. The addition of a bidirectional gated recurrent unit for temporal feature extraction has culminated in a model that offers enhanced fault classification accuracy and stability. Xiang et al. (2021) have taken a step further in enhancing 1DCNNs for recognition tasks by incorporating an aggregation–perception–recalibration block. This dynamic channel recalibration, coupled with the application of the global best leading artificial bee colony algorithm for network pruning, has significantly improved the efficiency of the network. This approach has commendably reduced the computational load without undermining the recognition accuracy.

Training strategies are also important for 1DCNNs. Che et al. (2024) introduces an interpretable multi-domain meta-transfer learning approach that enhances 1DCNNs by integrating time–frequency domain features and a fusion hierarchical class activation mapping, which improves the interpretability and diagnostic accuracy under variable operating conditions with few-shot samples.

While the aforementioned works have significantly enhanced the performance of 1DCNNs in specific tasks, there remain several issues that warrant attention. Firstly, the theoretical understanding of the deep optimization mechanisms within the 1D-Conv layers is not well-established. This obscurity can impede further advancements, as the lack of clarity on how these optimizations influence the network's learning process may limit the ability to refine and improve upon these techniques. Secondly, the methodologies introduced in these works can be relatively complex, involving multiple technical steps that may not be easily interpretable or replicable, especially for researchers and practitioners new to the field. In this paper, we propose a novel approach that addresses these issues by introducing over-parameterization of the input spatial dimensions in the 1D-Conv layers. This technique is both simple and effective, offering a marked improvement in the performance of 1DCNNs for aero-engine system inter-shaft bearing fault diagnosis. Moreover, our theoretical analysis reveals that this technique offers implicit acceleration based on momentum and adaptive learning rates. This finding is significant as it provides a deeper understanding of how the model learns.

3. Method

3.1. Architecture overview

To efficiently extract key features from vibration signals, the ISONet is proposed. As illustrated in Fig. 1, the comprehensive framework of ISONet reveals its intricate structure, divisible into seven distinct components. Initially, the input layer receives the raw vibration signals. This is followed by the 1D-Conv layer, responsible for extracting hierarchical features from the input data. Subsequently, the batch normalization (BN) (Ioffe & Szegedy, 2015) layer stabilizes the learning

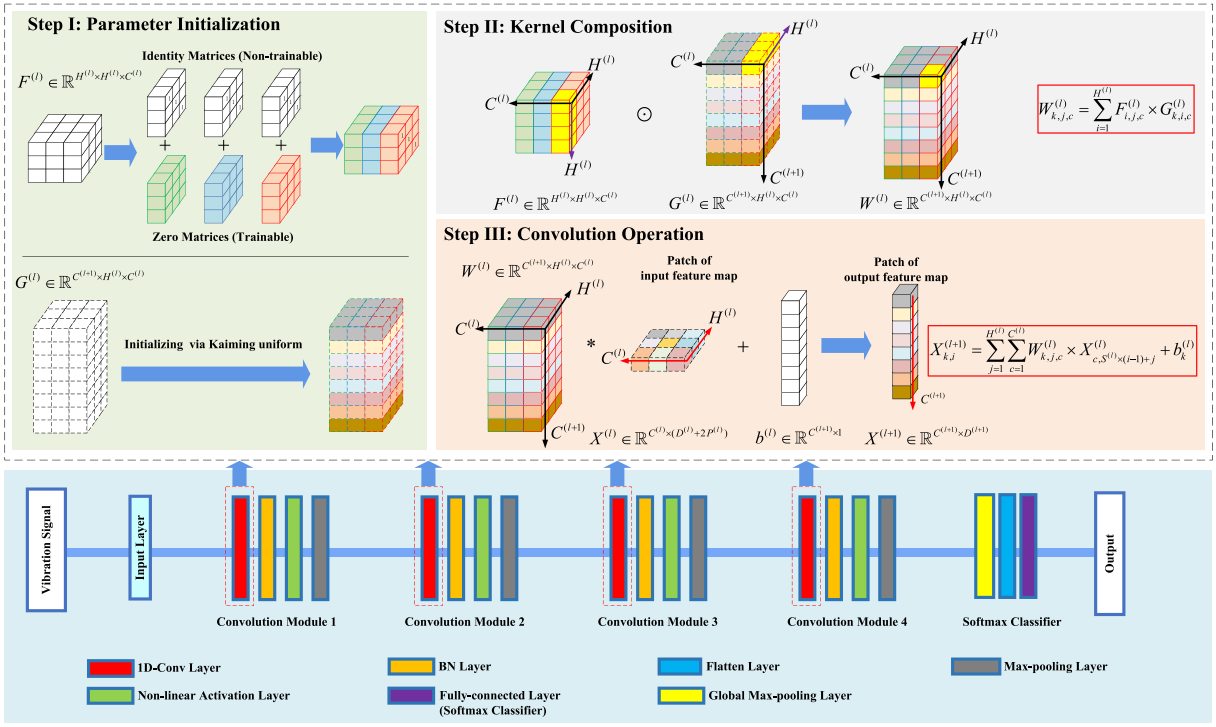


Fig. 1. The architecture of ISONet.

process and mitigates internal covariate shift. The non-linear activation layer introduces non-linearity, enabling the network to learn complex representations. The pooling layer then downsamples the feature maps, reducing dimensionality and computational complexity. Next, the flatten layer transforms the multi-dimensional tensor into a one-dimensional array, facilitating the transition to the final layers. Lastly, the Softmax classifier performs target classification, outputting probabilities for each class. Notably, the output of the Softmax classifier can also serve as the final embedded features, encapsulating the essential characteristics of the input vibration signals. The Conv module, comprising the 1D-Conv layer, BN layer, non-linear activation layer, and pooling layer, is connected successively to perform feature extraction efficiently. In contrast to the standard convolution operation typically used in a 1D-Conv layer, we introduce the concept of input spatial over-parameterized convolution in our 1D-Conv layer. This novel approach differs significantly from the conventional methods.

3.2. Components of ISONet

3.2.1. Input layer

The DL-based fault diagnosis methods generally include two major phases-the training phase and the testing phase. The training phase is the process of tuning the parameters based on the input data, while the testing phase is the process of predicting the input sample label using the trained model. These two phases require the dimensionality of the input data to be consistently the same size, so the vibration signal vector $x \in \mathbb{R}^{D^{(1)}}$ is transformed to a two dimension feature map tensor $X^{(1)} \in \mathbb{R}^{C^{(1)} \times D^{(1)}}$, where $D^{(1)}$ denotes the length of $X^{(1)}$ and $C^{(1)} = 1$ denotes the channel number of $X^{(1)}$, thus the input layer is fed with the shape-fixed original data tensor $X^{(1)}$.

3.2.2. 1D-Conv layer

The 1D-Conv layer performs convolution operations between the input feature maps and convolution kernel of fixed window size to automatically extract features, and generally the size of the convolution kernel is much smaller than that of the input feature map, so the convolution kernel moves over the input feature maps along the length in a certain stride, and the convolution kernel is shared among all the strides to achieve parameter sharing. Consider there are $\mathcal{L}$ Conv modules in ISONet, and the tensor $X^{(l)} \in \mathbb{R}^{C^{(l)} \times D^{(l)}}$ denotes the multi-channel input feature map of the lth ($l \in \{1, 2, \ldots, \mathcal{L}\}$) standard 1D-Conv layer, each channel corresponds to a 1D feature map, where $D^{(l)}$ and $C^{(l)}$ respectively denote the length of a single 1D feature map and the channel number of feature maps. The input feature map of the first 1D-Conv layer is the original data tensor, so the input feature map of the first 1D-Conv has just one channel as mentioned above. The 1D-Conv layer transforms the multi-channel input feature maps $X^{(l)}$ into multi-channel output feature maps $X^{(l+1)} \in \mathbb{R}^{C^{(l+1)} \times D^{(l+1)}}$, where $D^{(l+1)}$ and $C^{(l+1)}$ denote the length and channel number of the feature maps after convolution operation respectively. Let $H^{(l)}$ be the window size of the convolution kernel in the lth standard 1D-Conv layer and tensor $W^{(l)} \in \mathbb{R}^{C^{(l+1)} \times H^{(l)} \times C^{(l)}}$ be the lth standard 1D-Conv kernel tensor, since each 1D-Conv filter performs convolution operation with all the input feature maps, the 1D feature maps generated by the kth ($k \in \{1, 2, \ldots, C^{l+1}\}$) 1D-Conv filter $W_k^{(l)}$ is

$$X_k^{(l+1)} = W_k^{(l)} * X^{(l)} + b_k^{(l)} \tag{1}$$

where the operator $*$ denotes the convolution operation, $b^{(l)} \in \mathbb{R}^{C^{(l+1)} \times 1}$, $b_k^{(l)}$ denotes the bias parameter of the kth 1D-Conv kernel. Considering that the input feature maps of each layer may be padded with zeros, let the number of zeros padded on one side of each input feature map of the lth layer be $P^{(l)}$, the length of the input feature maps turns into $D^{(l)} + 2P^{(l)}$, i.e., $X^{(l)} \in \mathbb{R}^{C^{(l)} \times (D^{(l)} + 2P^{(l)})}$, and then the convolution operation is performed with the convolution kernel of a stride $S^{(l)}$, finally, the value of the output feature map in the kth layer at the position i ($i \in \{1, 2, \ldots, D^{(l+1)}\}$) is

$$X_{k,i}^{(l+1)} = \sum_{j=1}^{H^{(l)}} \sum_{c=1}^{C^{(l)}} W_{k,j,c}^{(l)} \times X_{c, S^{(l)} \times (i-1)+j}^{(l)} + b_k^{(l)} \tag{2}$$

where $i \in \{1, 2, \ldots, D^{(l+1)}\}$, $k \in \{1, 2, \ldots, C^{(l+1)}\}$. The length of the output feature map can be calculated as

$$D^{(l+1)} = \lfloor \frac{D^{(l)} + 2P^{(l)} - H^{(l)}}{S^{(l)}} \rfloor + 1 \tag{3}$$

The 1D-Conv process is illustrated in Fig. 1, Step III. In Section 3.3, we provide a detailed introduction to the input spatial over-parameterized 1D-Conv (1D-ISOConv) operation. This operation extends the traditional 1D-Conv by incorporating additional parameters only in the training phase.

3.2.3. Flatten layer and softmax classifier

The flatten layer arranges the output feature map of the global max-pooling layer $X^{*(\mathcal{L})} \in \mathbb{R}^{C^{(\mathcal{L})} \times 1}$ into a one-dimensional vector $u \in \mathbb{R}^{C^{(\mathcal{L})}}$ as the input to the Softmax classifier, which is $u^{(k)}$ for the kth sample. The Softmax classifier can be considered as a special fully-connected layer activated by the Softmax function, whose number of neurons is equal to $\mathcal{Q}$, the total number of target classes to be classified. Let the parameters of the Softmax classifier be $\hat{\theta} = \{\hat{W}, \hat{b}\}$, and denote the vibration signal sample as the pair $(x^{(k)}, y^{(k)})$, where $x^{(k)}$ and $y^{(k)}$ denote the vibration signal and the corresponding ground-truth label of the kth sample respectively, then the output of the qth neuron of the Softmax classifier represents the probability of predicting the sample $x^{(k)}$ as the qth class, which is

$$P(\bar{y}^{(k)} = q | x^{(k)}; \theta) = \frac{\exp(u^{(k)T} \cdot \hat{W}_q + \hat{b}_q)}{\sum_{j=1}^{\mathcal{Q}} \exp(u^{(k)T} \cdot \hat{W}_j + \hat{b}_j)} \tag{4}$$

$$u^{(k)} = Flatten(\phi_{\theta^*}(x^{(k)})) \tag{5}$$

where $\phi_{\theta^*}(\cdot)$ denotes the multi-dimensional features extracted by all layers before the Softmax classifier, θ^* is the parameters of all feature extraction layers except the Softmax classifier, $Flatten(\cdot)$ denotes the flatten operation, and $\theta = \{\theta^*, \hat{\theta}\}$ denotes all trainable parameters of the ISONet. The prediction process of the sample $x^{(k)}$ is to maximize the posterior probability,

$$y^{*(k)} = \underset{q \in \{1,2,\ldots,\mathcal{Q}\}}{\arg\max} \; P(\bar{y}^{(k)} = q | x^{(k)}; \theta) \tag{6}$$

From the above, it can be seen that the whole neural network is equivalent to implementing a mapping function f_θ parameterized by θ from the sample space $\mathbb{R}^{C^{(1)} \times 1}$ to the feature space $\mathbb{R}^{\mathcal{Q}}$,

$$x^{(k)} \in \mathbb{R}^{C^{(1)} \times 1} \xrightarrow{f_\theta} \tilde{y}^{(k)} \in \mathbb{R}^{\mathcal{Q}} \tag{7}$$

where the vector $\tilde{y}^{(k)}$ can be regarded as the embedded features (Xiang et al., 2024), which is

$$\tilde{y}^{(k)} = \begin{bmatrix} P(\bar{y}^{(k)} = 1 | x^{(k)}; \theta) \\ P(\bar{y}^{(k)} = 2 | x^{(k)}; \theta) \\ \vdots \\ P(\bar{y}^{(k)} = \mathcal{Q} | x^{(k)}; \theta) \end{bmatrix} \tag{8}$$

where $\bar{y}^{(k)}$ denotes the predicted label.

The ground-truth label $y^{(k)}$ of the sample $x^{(k)}$ can be represented by its unique one-hot vector $\hat{y}^{(k)} = [\hat{y}_1^{(k)}, \hat{y}_2^{(k)}, \ldots, \hat{y}_{\mathcal{Q}}^{(k)}]^T$, where

$$\hat{y}_i^{(k)} = \begin{cases} 1, i = y^{(k)} \\ 0, i \neq y^{(k)} \end{cases} \tag{9}$$

3.3. Design of 1D-ISOConv

Training optimization stands as an essential mechanism within neural networks. We introduce the concept of input spatial over-parameterization with a trilateral rationale: 1) to accelerate the training convergence of 1DCNNs designed for diagnosing inter-shaft bearing faults in aero-engine systems; 2) to ensure straightforward implementability, implying that the methodology proposed should eschew procedural complexity; and 3) to augment model efficacy without a substantial augmentation in parameter count and computational expense.

Fig. 2 depicts the structure of a 1D-Conv kernel, which includes dimensions for input channels, output channels, and the spatial extent of convolution. The 1D-Conv kernel's input channel count corresponds to the depth of the feature map channels, facilitating convolution operations with a dedicated set of filters for each input channel. The quantity of output channels dictates the channel count of the ensuing feature map, as articulated in Eq. (3), with the kernel sizes ascertaining the dimensions of the output feature map.

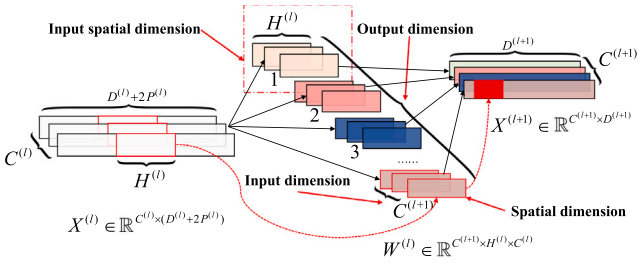


Fig. 2. Illustration of a standard 1D-Conv where the input spatial dimension is chosen as the dimension for over-parameterization. The blocks marked in red illustrate the positions corresponding to an element in the output feature map X_{l+1} in relation to the input feature map X_l and the 1D-Conv kernel W_l, thereby providing an example of a 1D-Conv operation. Referencing Eq. (2) will facilitate a better understanding.

The decision to over-parameterize in the input and spatial dimensions was deliberate, considering the trade-offs between model complexity and performance enhancement. As illustrated in Fig. 2, over-parameterization in the output and spatial dimensions, or in the input and output dimensions, would necessitate the construction of two matrices, thereby increasing execution complexity and making it less convenient to leverage the properties of matrix operations for interpretation. Moreover, over-parameterizing in all three dimensions simultaneously would result in a parameter count equivalent to double that of the standard convolution, significantly increasing computational cost and model complexity. Over-parameterization along the input channel and convolutional extent dimensions enables each output channel to share an over-parameterized tensor, curtailing the proliferation of parameters. Simultaneously, the transformation of each input channel, as a contiguous tensor in memory, into a matrix format is conducive to efficient computation through the application of sophisticated matrix computation algorithms. Therefore, the chosen approach of over-parameterizing in the input and spatial dimensions offers a more straightforward implementation and a better balance between model efficacy and computational efficiency, while also facilitating a clearer explanation of the over-parameterization mechanism.

To adapt to tasks involving 1D signal inputs, especially in the field of fault diagnosis, we propose ISONets with 1D-ISOConvs. Conventionally, as illustrated in Fig. 1, Step III, the standard convolution kernel $W^{(l)} \in \mathbb{R}^{C^{(l+1)} \times H^{(l)} \times C^{(l)}}$ is initialized using the Kaiming normalization method (He, Zhang, Ren, & Sun, 2015). However, in contrast to this standard practice, we employ a composition of two tensors in 1D-ISOConv, $F^{(l)} \in \mathbb{R}^{H^{(l)} \times H^{(l)} \times C^{(l)}}$ and $G^{(l)} \in \mathbb{R}^{C^{(l+1)} \times H^{(l)} \times C^{(l)}}$, to synthesize $W^{(l)}$ through a correlation tracingcorrelation tracing operation $\odot$, thus

$$W^{(l)} = F^{(l)} \odot G^{(l)} \tag{10}$$

Specifically, tensor $G^{(l)}$ shares the same shape as $W^{(l)}$, while tensor $F^{(l)}$ has the same number of input channels as tensor $G^{(l)}$. We refer to tensor $F^{(l)}$ as the over-parameterizing kernel because all additional parameters compared to the standard convolution operation originate from $F^{(l)}$.

As depicted in Fig. 1, Step I, tensors $G^{(l)}$ and $F^{(l)}$ undergo different initialization methods. Specifically, $G^{(l)}$ is initialized using the Kaiming normalization method, which is consistent with the initialization approach for the standard convolution kernel $W^{(l)}$. In contrast, for the over-parameterizing kernel $F^{(l)}$, we initialize each $H^{(l)} \times H^{(l)}$ matrix corresponding to every input channel as a zero matrix and then add the identity matrix to it. The zero matrix serves as trainable parameters, whereas the identity matrix remains frozen during the training process.

This initialization strategy ensures that at the onset of training, the ISONet behaves essentially the same as a standard 1DCNN. Consequently, the over-parameterization does not initially alter the operation of the standard 1DCNN.

The kernel composition is achieved through the element-wise multiplication and accumulation of $F^{(l)}$ and $G^{(l)}$ across each input channel. Essentially, this kernel composition operates in a input spatial manner, meaning that the composition is performed independently for each input channel. As depicted in Fig. 1, Step II, the detail process of kernel composition in Eq. (10) is formulated as shown in Eq. (11).

$$W^{(l)}_{k,j,c} = \sum_{i=1}^{H^{(l)}} F^{(l)}_{i,j,c} \times G^{(l)}_{k,i,c} \tag{11}$$

where $k \in \{1, 2, \dots, C^{(l+1)}\}$, $c \in \{1, 2, \dots, C^{(l)}\}$, $j \in \{1, 2, \dots, H^{(l)}\}$.

3.4. Theoretical interpretation of input spatial over-parameterization

To elucidate the theoretical insights of input spatial over-parameterization, we first present the training optimization objective of ISONet, namely, the loss function. The cross-entropy loss function is commonly employed for training deep neural networks, including the ISONet used for aero-engine system inter-shaft bearing fault diagnosis. However, training the model with all the training samples simultaneously can be computationally demanding and memory-intensive, especially for large datasets and networks with many parameters. Therefore, the parameters are typically updated using the mini-batch stochastic gradient descent algorithm, which involves randomly selecting a mini-batch set of samples $\mathscr{B}$ at each iteration. The cross-entropy loss function on each mini-batch of samples can be expressed by Eq. (12).

$$\mathcal{J}(\mathscr{B}; \theta) = -\frac{1}{|\mathscr{B}|} \sum_{x^{(k)} \in \mathscr{B}} \sum_{q=1}^{Q} \hat{y}^{(k)}_q \log P(\bar{y}^{(k)} = q | x^{(k)}; \theta) \tag{12}$$

The overarching objective is to iteratively converge upon the optimal parameter set θ^* that minimizes the loss function as articulated by Eq. (13). During the tth iteration ($t \in \{1, 2, \dots, \mathcal{T}\}$), where $\mathcal{T}$ denotes the total number of mini-batches, the stochastic gradient descent update rule, presented in Eq. (14), involves adjusting the parameters based on the learning rate η and the gradient $g^{(t)}$, which is derived from Eq. (15).

$$\theta^* = \arg\min_{\theta} \mathcal{J}(\mathscr{B}; \theta) \tag{13}$$

$$\theta^{(t+1)} = \theta^{(t)} - \eta d^{(t)} \tag{14}$$

$$d^{(t)} = \nabla_{\theta^{(t)}} \mathcal{J}(\mathscr{B}; \theta^{(t)}) \tag{15}$$

Due to the formal nature of Conv operations as tensor operations, which currently cannot be derived using mathematical theory, we equivalently transform the tensor operations involved in the input spatial over-parameterization process into corresponding matrix/vector operations. We then leverage the properties of matrix/vector operations to conduct theoretical proofs.

Consider an element z of the Conv layer output $X^{(l+1)}$ in Eq. (2), denoted as $z = X^{(l+1)}_{k,i}$, where i indexes over the spatial dimensions $D^{(l+1)}$ and k indexes over the channels $C^{(l+1)}$. As depicted in Fig. 3, we trace the range of Conv weight parameters and the corresponding range of the input feature map $X^{(l)}$ that contribute to the computation of z. The relevant portions of the weight tensors F, G, and W, as well as the input feature map tensor $X^{(l)}$, are denoted by $\mathscr{F}$, g, w, and x, respectively. Note that the Conv operation involves the computation of weights between input channels and the interaction of input feature maps across those channels, we represent the process of the Conv operation to obtain z in vector form, such that $\mathscr{F} \in \mathbb{R}^{(C^{(l)} \times H^{(l)}) \times (C^{(l)} \times H)}$, $g \in \mathbb{R}^{(C^{(l)} \times H^{(l)}) \times 1}$, $\mathbf{x} \in \mathbb{R}^{(C^{(l)} \times H^{(l)}) \times 1}$, $\mathbf{w} \in \mathbb{R}^{(C^{(l)} \times H^{(l)}) \times 1}$. Fig. 3 detailedly illustrates how to rearrange $\mathscr{F}$, g, w, and x to meet the computational requirements. Therefore, the computation of z can be expressed as $z = (\mathscr{F} \cdot g)^T \cdot x$, where $w = \mathscr{F} \cdot g$ and thus $z = w^T \cdot x$.

We have made a minor adjustment to the notation of the loss function, employing $\mathcal{J}(z)$ to denote the forward propagation process induced by z. Then, the gradients of $\mathscr{F}$, g, and w with respect to a loss function $\mathcal{J}$ induced by z are given by the following set of equations:

$$\begin{aligned} \nabla_w &= \frac{\partial \mathcal{J}(w^T x)}{\partial z} \cdot x, \\ \nabla_{\mathscr{F}} &= \frac{\partial \mathcal{J}((\mathscr{F} \cdot g)^T x)}{\partial z} \cdot g \cdot x^T, \\ \nabla_g &= \frac{\partial \mathcal{J}((\mathscr{F} \cdot g)^T x)}{\partial z} \cdot \mathscr{F}^T \cdot x. \end{aligned} \tag{16}$$

According to the parameter update rule outlined in Eq. (14), the updates for $\mathscr{F}$ and g are performed as follows:

$$\begin{aligned} \mathscr{F}^{(t+1)} &= \mathscr{F}^{(t)} - \eta \nabla_{\mathscr{F}^{(t)}}, \\ g^{(t+1)} &= g^{(t)} - \eta \nabla_{g^{(t)}}. \end{aligned} \tag{17}$$

The dynamics of the underlying parameter $w = \mathscr{F} \cdot g$ are then described by:

$$\begin{aligned} w^{(t+1)} &= \mathscr{F}^{(t+1)} \cdot g^{(t+1)} \\ &= (\mathscr{F}^{(t)} - \eta \nabla_{\mathscr{F}^{(t)}}) \cdot (g^{(t)} - \eta \nabla_{g^{(t)}}) \\ &= \mathscr{F}^{(t)} \cdot g^{(t)} - \eta \mathscr{F}^{(t)} \cdot \nabla_{g^{(t)}} - \eta \nabla_{\mathscr{F}^{(t)}} \cdot g^{(t)} + \mathcal{O}(\eta^2) \\ &= w^{(t)} - \eta (\mathscr{F}^{(t)})^2 \nabla_{w^{(t)}} - \frac{\eta \nabla_{\mathscr{F}^{(t)}}}{\mathscr{F}^{(t)}} w^{(t)} + \mathcal{O}(\eta^2), \end{aligned} \tag{18}$$

where η is assumed to be small and the higher-order term $\mathcal{O}(\eta^2)$ is neglected. By defining $\rho^{(t)} = \eta (\mathscr{F}^{(t)})^2 \in \mathbb{R}^{(C^{(l)} \times H^{(l)}) \times (C^{(l)} \times H^{(l)})}$ and $\xi^{(t)} = \frac{\eta \nabla_{\mathscr{F}^{(t)}}}{\mathscr{F}^{(t)}} \in \mathbb{R}^{(C^{(l)} \times H^{(l)}) \times (C^{(l)} \times H^{(l)})}$, the update rule for w is simplified to:

$$w^{(t+1)} = w^{(t)} - \rho^{(t)} \nabla_{w^{(t)}} - \xi^{(t)} w^{(t)}. \tag{19}$$

Given that $\mathscr{F}$ and g are initialized near zero, w will also initialize near zero. This implies that at each iteration t, $w^{(t)}$ is a weighted sum of past gradients. Consequently, there exist coefficients $\mu^{(t,\tau)} \in \mathbb{R}^{(C^{(l)} \times H^{(l)}) \times H^{(l)}}$ such that the update rule for w can be expressed as a gradient descent process with an adaptive momentum (Arora, Cohen, & Hazan, 2018; Kingma & Ba, 2015; Zhuang et al., 2020) term:

$$w^{(t+1)} = w^{(t)} - \rho^{(t)} \nabla_{w^{(t)}} - \sum_{\tau=1}^{t-1} \mu^{(t,\tau)} \nabla_{w^{(\tau)}}. \tag{20}$$

This reveals that the dynamics of the underlying parameter w are equivalent to those of a gradient descent algorithm with both a time-varying learning rate $\rho^{(t)}$ and adaptive momentum coefficients $\mu^{(t,\tau)}$. This observation provides insight into how input spatial over-parameterization can implicitly introduce acceleration mechanisms within the optimization process.

In the subsequent analysis, we demonstrate that input spatial over-parameterization adheres to a specific preconditioning scheme for gradient descent on a loss function $\mathcal{J}$, independent of the particular forms of $\mathscr{F}^{(t)}$ and $g^{(t)}$. The following analysis also unfolds from more complex network training scenarios, such as considering weight decay. We shall elucidate the underlying continuity from the perspective of differential equations, focusing on the trainable parameters $\mathscr{F}^{(t)}$ and $g^{(t)}$. The gradient descent updates are articulated as follows:

$$\begin{aligned} \mathscr{F}^{(t+1)} &= (1 - \eta\lambda)\mathscr{F}^{(t)} - \eta \frac{\partial \mathcal{J}(\mathscr{F}^{(t)} \cdot g^{(t)})}{\partial \mathscr{F}}, \\ g^{(t+1)} &= (1 - \eta\lambda) g^{(t)} - \eta \frac{\partial \mathcal{J}(\mathscr{F}^{(t)} \cdot g^{(t)})}{\partial g}, \end{aligned} \tag{21}$$

where λ is an optional weight decay factor.

Under the assumption of a small learning rate, the aforementioned discrete updates can be transformed into the following differential equations, where t represents a continuous time index, and $\dot{\mathscr{F}}(t)$ and

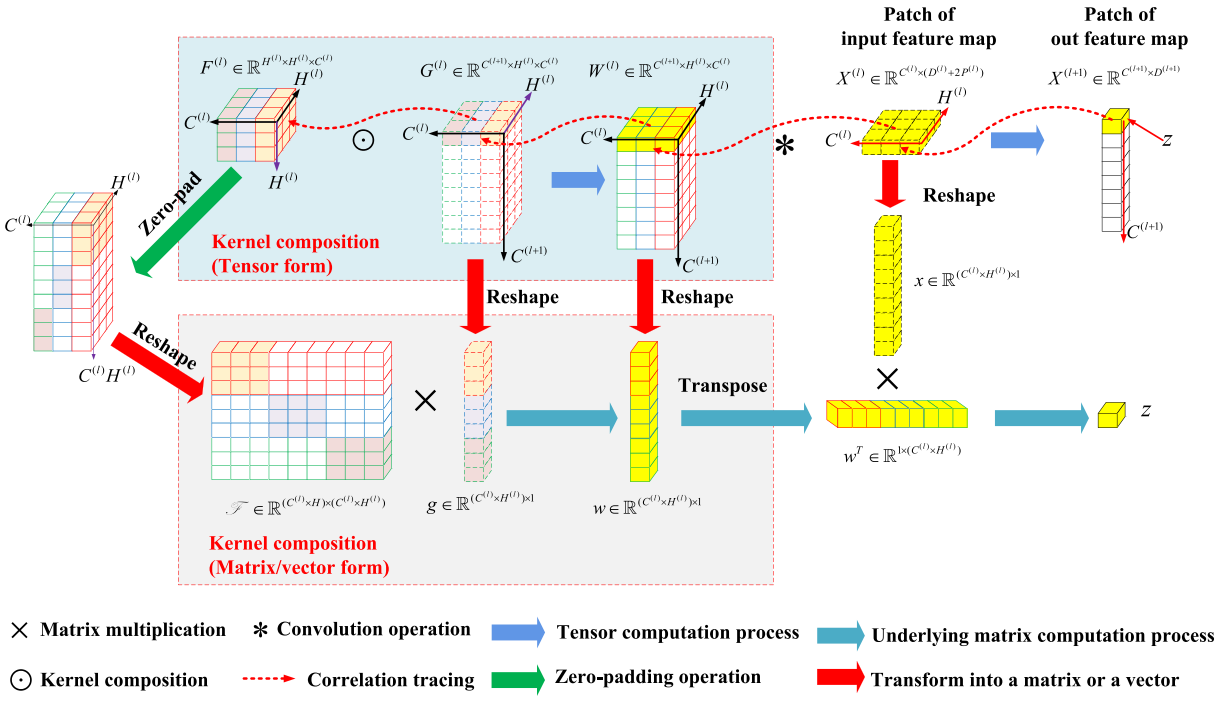


Fig. 3. Diagram of tensor operation process of 1D-ISOConv and its underlying matrix/vector operations. The blocks filled with color represent the weights/inputs related to the computation of z. Referencing Eq. (11) will facilitate a better understanding.

$\dot{g}(t)$ denote the derivatives of $\mathscr{F}(t)$ and $g(t)$ with respect to time, respectively:

$$\begin{aligned}\dot{\mathscr{F}}(t) &= -\eta\lambda\mathscr{F}(t) - \eta\frac{\partial\mathcal{J}(\mathscr{F}(t)\cdot g(t))}{\partial\mathscr{F}},\\ \dot{g}(t) &= -\eta\lambda g(t) - \eta\frac{\partial\mathcal{J}(\mathscr{F}(t)\cdot g(t))}{\partial g}.\end{aligned} \tag{22}$$

As the learning rate diminishes, the trajectories of discrete optimization algorithms converge towards smoother curves, which are well-captured by continuous-time differential equations. This convergence facilitates the application of established theories associated with differential equations (Su, Boyd, & Candès, 2014; Wibisono, Wilson, & Jordan, 2016).

To begin with, we present Lemma 1, a critical tool that delineates the circumstances under which a pair of differentiable functions will display congruent properties across a designated interval. This is predicated on the existence of a particular correlation between the derivatives of the functions and their respective values.

Lemma 1. *(Arora et al., 2018) Consider a connected subset I of the real numbers $\mathbb{R}$, and let $f, g : I \to \mathbb{R}$ be differentiable functions defined thereon. Suppose that for some non-negative constant $\alpha \geq 0$, the following condition is satisfied for all t within I :*

$$\dot{f}(t) + \alpha \cdot f(t) = \dot{g}(t) + \alpha \cdot g(t)$$

If f and g are observed to take identical values at some point t_0 within I, whether it be an interior point or a boundary, it is hereby asserted that f and g are indistinguishable throughout the interval, that is, it must be the case that $f(t) = g(t)$ for all $t \in I$.

This lemma provides a rigorous foundation for understanding the implications of a differential relationship between two functions, highlighting the uniqueness of their solutions under given initial conditions. With the continuous formulation established, we proceed to express the dynamics of w as postulated by Theorem. 1.

Theorem 1. *Assuming the dynamics of continuous gradient descent as described in Eq. (22), and given that the initial values at time t_0 satisfy the condition:*

$$\mathscr{F}^T(t_0)\mathscr{F}(t_0) = g(t_0)g^T(t_0) \tag{23}$$

Then, the vector-form parameter w is governed by the following differential equation

$$\begin{aligned}\dot{w}(t) = &-2\eta\lambda w(t)\\ &- \eta\|w(t)\|_2\left(\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w} + Proj_w\left(\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}\right)\right)\end{aligned} \tag{24}$$

where the notation $\|\cdot\|_2$ denotes the Euclidean norm. Furthermore, the operator $Proj_w\{\cdot\}$, for $w \in \mathbb{R}^{(C^{(l)}\times H^{(l)})\times 1}$, is characterized as the projection operator onto the direction of w. This is mathematically formalized as $Proj_w : \mathbb{R}^{(C^{(l)}\times H^{(l)})\times 1} \to \mathbb{R}^{(C^{(l)}\times H^{(l)})\times 1}$. The definition of $Proj_w\{V\}$ is encapsulated by the conditional expression:

$$Proj_w\{V\} = \begin{cases}\frac{w}{\|w\|_2}V^\top\cdot\frac{w}{\|w\|_2} & \text{if } w \neq 0,\\ 0 & \text{if } w = 0.\end{cases} \tag{25}$$

This operator serves to map any vector V in the $\mathbb{R}^{(C^{(l)}\times H^{(l)})\times 1}$ space onto the direction of w, with the operation being trivially zero when w is the null vector. This formulation ensures that the projection is aligned with the vector w, thereby maintaining the geometric integrity of the operation within the specified dimensional space.

Proof. Given the premise that $w = \mathscr{F}\cdot g$, we derive the following expressions for the partial derivatives of the loss function $\mathcal{J}$ with respect to $\mathscr{F}$ and g:

$$\begin{aligned}\frac{\partial\mathcal{J}(\mathscr{F}\cdot g)}{\partial\mathscr{F}} &= \frac{\mathrm{d}\mathcal{J}(w)}{\mathrm{d}w}g^T,\\ \frac{\partial\mathcal{J}(\mathscr{F}\cdot g)}{\partial g} &= \mathscr{F}^T\frac{\mathrm{d}\mathcal{J}(w)}{\mathrm{d}w}.\end{aligned} \tag{26}$$

Substituting these into Eq. (22) yields the dynamical system:

$$\begin{aligned}\dot{\mathscr{F}}(t) &= -\eta\lambda\mathscr{F}(t) - \eta\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}g^T(t),\\ \dot{g}(t) &= -\eta\lambda g(t) - \eta\mathscr{F}^T(t)\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}.\end{aligned} \tag{27}$$

By premultiplying the first equation of Eq. (27) by $\mathscr{F}^T(t)$ and postmultiplying the second by $g^T(t)$, we obtain:

$$\begin{aligned}\mathscr{F}^T(t)\dot{\mathscr{F}}(t) &= -\eta\lambda\mathscr{F}^T(t)\mathscr{F}(t) - \eta\mathscr{F}^T(t)\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}g^T(t),\\ \dot{g}(t)g^T(t) &= -\eta\lambda g(t)g^T(t) - \eta\mathscr{F}^T(t)\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}g^T(t).\end{aligned} \tag{28}$$

From Eq. (28), it is evident that

$$\mathscr{F}^T(t)\dot{\mathscr{F}}(t) + \eta\lambda\mathscr{F}^T(t)\mathscr{F}(t) = \dot{g}(t)g^T(t) + \eta\lambda g(t)g^T(t). \tag{29}$$

Transposing Eq. (29) provides

$$\dot{\mathscr{F}}^T(t)\mathscr{F}(t) + \eta\lambda\mathscr{F}^T(t)\mathscr{F}(t) = g(t)\dot{g}^T(t) + \eta\lambda g(t)g^T(t). \tag{30}$$

Adding Eqs. (28) and (29), we arrive at

$$\begin{aligned}&\mathscr{F}^T(t)\dot{\mathscr{F}}(t) + \dot{\mathscr{F}}^T(t)\mathscr{F}(t) + 2\eta\lambda\mathscr{F}^T(t)\mathscr{F}(t)\\ &= \dot{g}(t)g^T(t) + g(t)\dot{g}^T(t) + 2\eta\lambda g(t)g^T(t).\end{aligned} \tag{31}$$

Let $a(t) = \mathscr{F}(t)\mathscr{F}^T(t)$, $a'(t) = \mathscr{F}^T(t)\mathscr{F}(t)$, $b(t) = g(t)g^T(t)$, and $b'(t) = g^T(t)g(t)$. According to Eq. (31), we obtain

$$\dot{a}'(t) + 2\eta\lambda a'(t) = \dot{b}(t) + 2\eta\lambda b(t). \tag{32}$$

Given the assumption that $a'(t_0) = b(t_0)$, invoking Lemma. 1, we conclude that $a'(t) = b(t)$, hence

$$\mathscr{F}^T(t)\mathscr{F}(t) = g(t)g^T(t). \tag{33}$$

Considering $\mathscr{F}(t)$ and $g(t)$ as fixed matrices, and performing singular value decomposition on $\mathscr{F}(t)$ and $g(t)$ respectively, we deduce:

$$\begin{aligned}g(t) &= U_g\Sigma_g V_g^T,\\ \mathscr{F}(t) &= U_{\mathscr{F}}\Sigma_{\mathscr{F}}V_{\mathscr{F}}^T.\end{aligned} \tag{34}$$

where $U_g, U_{\mathscr{F}}, V_g$, and $V_{\mathscr{F}}$ are all orthogonal matrices, and Σ_g and $\Sigma_{\mathscr{F}}$ are diagonal matrices with non-negative singular values on their diagonals, Eq. (33) reveals an insightful relationship:

$$V_{\mathscr{F}}\Sigma_{\mathscr{F}}^T\Sigma_{\mathscr{F}}V_{\mathscr{F}}^T = U_g\Sigma_g\Sigma_g^T U_g^T. \tag{35}$$

The two sides of the aforementioned equation represent orthogonal eigenvalue decompositions of the identical matrix. Consequently, the square-diagonal matrices $\Sigma_{\mathscr{F}}^T\Sigma_{\mathscr{F}}$ and $\Sigma_g\Sigma_g^T$ are identical, leading to the formulation:

$$\Sigma_{\mathscr{F}}^T\Sigma_{\mathscr{F}} = \Sigma_g\Sigma_g^T = \mathrm{diag}(\rho). \tag{36}$$

Furthermore, there exists a real number o such that:

$$U_g = V_{\mathscr{F}}\mathrm{diag}(o). \tag{37}$$

This leads to the following expressions for $g(t)$ and $\mathscr{F}(t)$, respectively:

$$\begin{aligned}g(t) &= U_g\Sigma_g V_g^T = V_{\mathscr{F}}\mathrm{diag}(o)\mathrm{diag}(\sqrt{\rho})V_g^T,\\ \mathscr{F}(t) &= U_{\mathscr{F}}\Sigma_{\mathscr{F}}V_{\mathscr{F}}^T = U_{\mathscr{F}}\mathrm{diag}(\sqrt{\rho})V_{\mathscr{F}}^T.\end{aligned} \tag{38}$$

Given $w = \mathscr{F}\cdot g$, it follows that:

$$\begin{aligned}w(t)w^T(t) &= \mathscr{F}(t)g(t)g^T(t)\mathscr{F}^T(t) = U_{\mathscr{F}}\mathrm{diag}(\rho^2)U_{\mathscr{F}}^T,\\ w^T(t)w(t) &= g^T(t)\mathscr{F}^T(t)\mathscr{F}(t)g(t) = V_g\mathrm{diag}(\rho^2)V_g^T.\end{aligned} \tag{39}$$

According to Eq. (38), for g we have:

$$\begin{aligned}g^T(t)g(t) &= V_g diag(\rho)V_g^T\\ &= \left[w^T(t)w(t)\right]^{\frac{1}{2}}.\end{aligned} \tag{40}$$

And according to Eq. (38), for $\mathscr{F}$ we have:

$$\begin{aligned}\mathscr{F}(t)\mathscr{F}^T(t) &= U_{\mathscr{F}}diag(\rho)U_{\mathscr{F}}^T\\ &= \left[w(t)w^T(t)\right]^{\frac{1}{2}}.\end{aligned} \tag{41}$$

For w, the derivative with respect to time t is given by:

$$\begin{aligned}\dot{w}(t) &= \dot{\mathscr{F}}(t)g(t) + \mathscr{F}(t)\dot{g}(t)\\ &= -2\eta\lambda\mathscr{F}(t)g(t) - \eta\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}g^T(t)g(t)\\ &\quad - \eta\mathscr{F}(t)\mathscr{F}^T(t)\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}\\ &= -2\eta\lambda w(t) - \eta\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}\left[w^T(t)w(t)\right]^{\frac{1}{2}}\\ &\quad - \eta\left[w(t)w^T(t)\right]^{\frac{1}{2}}\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}.\end{aligned} \tag{42}$$

Considering the property that

$$\left[w(t)w^T(t)\right]^{\frac{1}{2}} = \|w(t)\|_2\left(\frac{w(t)}{\|w(t)\|_2}\right)\left(\frac{w(t)}{\|w(t)\|_2}\right)^T. \tag{43}$$

From Eq. (42), we ultimately arrive at the following expression for the time derivative of w:

$$\begin{aligned}\dot{w}(t) &= -2\eta\lambda w(t)\\ &\quad - \eta\|w(t)\|_2\left(\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w} + \mathrm{Proj}_w\left(\frac{\mathrm{d}\mathcal{J}(w(t))}{\mathrm{d}w}\right)\right)\end{aligned} \tag{44}$$

This derivation elucidates the dynamical behavior of $w(t)$ under the influence of the optimization process, incorporating the effects of weight decay, gradient descent, and the projection onto the subspace spanned by the eigenvectors associated with the largest singular values of $w(t)$. □

Deriving an update rule for the variable w from the continuous dynamics outlined in Eq. (24) of Theorem. 1, we engage in the process of reverting to discrete time, and obtain

$$\begin{aligned}w^{(t+1)} &= (1-2\eta\lambda)w^{(t)}\\ &\quad - \eta\|w^{(t)}\|_2\left(\frac{\mathrm{d}\mathcal{J}(w^{(t)})}{\mathrm{d}w} + \mathrm{Proj}_{w^{(t)}}\left(\frac{\mathrm{d}\mathcal{J}(w^{(t)})}{\mathrm{d}w}\right)\right)\end{aligned} \tag{45}$$

This reveals that the evolution of w is subject to a differential equation that encapsulates the effects of preconditioning within the gradient descent framework, thereby offering insights into the accelerated optimization process afforded by over-parameterization. The impact of overparameterization on the gradient descent process, as elucidated in our scholarly work, manifests in a dual capacity. Initially, it engenders an adaptive learning rate schedule. This is achieved through the incorporation of a multiplicative factor $\eta\|w^{(t)}\|_2$, which dynamically modulates the learning rate based on the magnitude of the parameter vector w at iteration t. Secondly, overparameterization augments the projection of the gradient onto the direction of w, thereby reinforcing the alignment of the update step with the trajectory established by prior iterations. It is imperative to recognize that w is perceived not solely as the parameter to be optimized but also as a representation of the cumulative movement within the optimization landscape, given that the initialization is presumed to be in proximity to the origin. In this context, the adaptive learning rate schedule serves as a mechanism of increasing confidence; as the optimization progresses and diverges from the initial state, the step sizes are correspondingly enlarged.

Moreover, the amplification of the gradient projection can be construed as a form of momentum, which perpetuates the optimization along the direction that has been endorsed by the preceding updates. This momentum is instrumental in propelling the optimization process forward, thereby enhancing the likelihood of swift convergence towards the optimal solution. These synergistic effects, as delineated in Section 4.5, are posited to possess the inherent capacity to expedite the convergence of the optimization algorithm, thereby offering a compelling rationale for the strategic utilization of overparameterization in the pursuit of enhanced algorithmic performance.

3.5. Recap of 1D-ISOConv properties

3.5.1. 1D-ISOConv is over-parameterized in the training phase to accelerate training process

During the training phase, the 1D-ISOConv operation is over-parameterized due to the introduction of the over-parameterizing kernel $F^{(l)}$. As mentioned earlier, all additional parameters compared to the standard convolution operation originate from $F^{(l)}$, which is $H^{(l)}\times H^{(l)}\times C^{(l)}$. The 1D-ISOConv layer, through its unique initialization and kernel composition, offers an implicit yet profound enhancement to the optimization process:

- **Momentum-like behavior** (Zhuang et al., 2020): The trainable zero matrices in F serve as an accumulator of gradients, analogous to the velocity term in momentum methods. This initialization strategy imparts an inertial component to the learning process, where the update at each step is influenced by the accumulated gradients, thereby simulating a momentum-like effect.
- **Adaptive learning rate mechanism** (Arora et al., 2018; Kingma & Ba, 2015; Luo, Xiong, Liu, & Sun, 2019; Zhuang et al., 2020): The interaction between F and G during the kernel composition operation inherently modulates the learning rate for different input channels. The trainable elements in F adaptively scale the gradients contributed by G, akin to an adaptive learning rate scheme that adjusts the learning rate based on the historical gradient information.

This recomposition encapsulates the momentum-like behavior and adaptive learning rate properties within the 1D-ISOConv framework. The over-parameterization, a hallmark of 1D-ISOConv, enriches the optimization landscape, potentially leading to faster convergence and improved generalization.

During the training phase, the over-parameterizing kernel $F^{(l)}$ in 1D-ISOConv is initialized with zero and identity matrices, which ensures that the network behaves essentially the same as a standard 1DCNN at the onset of training. As training progresses, the trainable zero matrices in $F^{(l)}$ allow for accelerated learning, while the frozen identity matrices ensure that the over-parameterization does not initially alter the operation of the standard 1DCNN. This combination of trainable and frozen parameters in $F^{(l)}$ enables the network to learn faster without compromising its initial behavior.

3.5.2. 1D-ISOConv is equivalent to 1D-Conv in the testing phase without any additional complexity

During the testing phase, the 1D-ISOConv operation is equivalent to a standard 1D-Conv operation. This is because the over-parameterizing kernel $F^{(l)}$ is no longer trainable, and the identity matrix added to each $H^{(l)}\times H^{(l)}$ matrix during initialization remains frozen. As a result, the kernel composition operation in 1D-ISOConv reduces to a standard convolution operation, with the kernel $W^{(l)}$ being equal to the composition of $F^{(l)}$ and $G^{(l)}$. By performing the kernel composition operation beforehand and saving the resulting kernel, the 1D-ISOConv operation can be viewed as having the same complexity as the 1D-Conv operation during the testing phase. This is because the kernel composition operation is performed only once, and the resulting kernel can be used for all subsequent convolution operations. As a result, the 1D-ISOConv operation can be implemented efficiently and accurately for tasks involving one-dimensional signals, such as fault diagnosis.

This equivalence between 1D-ISOConv and 1D-Conv in the testing phase ensures that the learned features are consistent and that the performance of the network is maintained. It also allows for the use of pre-trained standard 1D-Conv networks for tasks involving one-dimensional signals, as the learned features can be directly transferred to the ISONet.

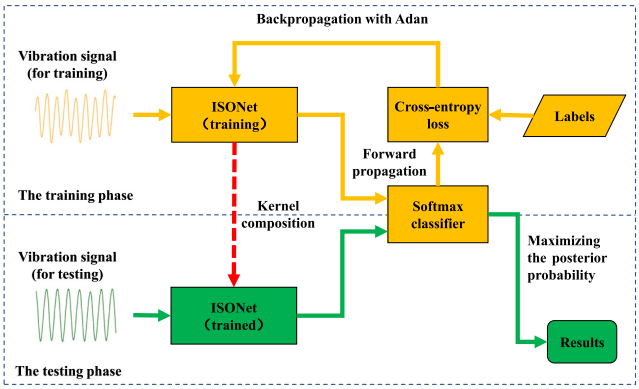


Fig. 4. The flowchart of aero-engine inter-shaft bearing fault diagnosis based on ISONet.

In summary, the design of 1D-ISOConv effectively meets the three criteria outlined in Section 3.3. Firstly, the introduction of the over-parameterized kernel in the training phase significantly accelerates the convergence of 1DCNNs by incorporating momentum-like behavior and an adaptive learning rate mechanism. Secondly, the implementation of 1D-ISOConv is straightforward, as it does not require complex procedural steps and can be easily integrated into existing 1DCNN architectures. Lastly, the enhancement in model efficacy is achieved without a substantial increase in parameters and computational cost, as the over-parameterization is limited to the training phase and does not affect the testing phase, where 1D-ISOConv is equivalent to a standard 1D-Conv. This design ensures that 1D-ISOConv is both efficient and effective for the task of aero-engine system inter-shaft bearing fault diagnosis.

3.6. Model training with Adan optimizer

The Adan optimizer (Xie, Zhou, Li, Lin, & Yan, 2022) is an advanced and recently proposed optimization algorithm that harnesses the synergy between the moving average concept and the Nesterov acceleration technique. It adeptly approximates the first-order and second-order moments to facilitate an informed parameter update strategy. Prior research has elucidated that the Nesterov acceleration, which exploits the gradient at an extrapolated point of the current solution, theoretically outpaces the convergence rate of traditional heavy ball methods (Nemirovskii & Nesterov, 1985). This strategic "future glimpse" potentially enhances the robustness of deep neural networks by leveraging comprehensive trajectory information. Consequently, the Adan optimizer is employed in the training regime of the ISONet for the fault diagnosis of aero-engine inter-shaft bearing systems.

In the context of ISONet's training, the Adan optimizer undertakes the task of iteratively refining the network's parameters using Eq. (12), which encapsulates the cross-entropy loss. As shown in Fig. 4, the methodology for cultivating the ISONet for aero-engine system inter-shaft bearing fault diagnosis unfolds over several structured steps:

3.6.1. Step 1: Constructing ISONet

Initially, a one-dimensional configuration array $C = [C^{(1)}, C^{(2)}, \ldots, C^{(\mathcal{L})}]$ delineates the number of output channels within each convolutional module, signifying the model's architectural blueprint. Subsequently, the specifics of the network structure, including layer count, kernel dimensions, strides, and padding, are meticulously defined alongside the selection of pooling and activation functions as illustrated in Fig. 1.

3.6.2. Step 2: ISONet initialization

As delineated in Section 3.3, the initialization of $G^{(l)}$ adheres to the principle of Kaiming normalization (He et al., 2015). For the overparameterized kernel $F^{(l)}$, an $H^{(l)} \times H^{(l)}$ matrix associated with every input channel is initialized as a zero matrix, added by an identity matrix to maintain certainty during the training phase.

3.6.3. Step 3: Forward propagation

The input data undergoes propagation through the network, with each convolution module generating an output informed by convolution operations followed by pooling and activation functions. The discrepancy between the actual and predicted labels is then quantified using the cross-entropy loss function specified in Eq. (12).

3.6.4. Step 4: Backpropagation with Adan

The Adan algorithm, coupled with the chain rule, facilitates the computation of gradients for each trainable parameter as indicated by Eq. (15). These gradients are instrumental in updating the parameters via Eq. (14) with the Adan optimizer (Xie et al., 2022) providing adaptive tuning of the learning rate and update direction. The entire process is reiterated for each mini-batch until all training samples have contributed to the parameter refinement.

3.6.5. Step 5:training per mini-batch

The processes detailed in Steps 3 and 4 are executed cyclically, encompassing $\mathcal{T}$ mini-batches of samples. The model, post-training, is deployed to predict outcomes for the test dataset using Eq. (6), with performance metrics such as accuracy and loss being duly recorded.

3.6.6. Step 6: Epoch-wise training

The training regimen spans across Y epochs, with each epoch comprising a complete cycle over all training samples. Post each training epoch, the model parameters that achieve the highest testing accuracy are preserved.

Through these methodical steps, the ISONet can be effectively trained and refined using the Adan optimizer, culminating in precise vibration signal recognition. The nuances of each step, including network specifications, initialization protocols, and learning rate adaptability, can be fine-tuned to further enhance the efficacy of the network.

4. Experiments and results

4.1. Measured dataset

The HIT (Harbin Institute of Technology) (Hou et al., 2023) dataset is a significant contribution to the field of aero-engine fault diagnosis, particularly focusing on inter-shaft bearing faults. The dataset was meticulously developed through a series of experiments conducted on an aero-engine test rig (as shown in Fig. 5) that replicates the structure and function of a real aero-engine. This rig was driven by two motors to simulate operational conditions and was equipped with a lubricating system to maintain the integrity of the engine components during testing.

Prior to the commencement of the tests, the aero-engine underwent a process of disassembly and reassembly in accordance with specified procedures. During this phase, three inter-shaft bearings were carefully replaced with ones that had been intentionally damaged through wire cutting on both the inner and outer rings to create artificial faults. The aero-engine test was then executed across 28 distinct groups of rotor speeds, encompassing both high and low pressure scenarios, with rotational speeds ranging from 1000 to 6000 rpm. To capture comprehensive vibration data, six measurement points were strategically arranged as shown in Fig. 5). Two displacement sensors were positioned to detect displacement vibration signals from the low-pressure rotor, while four acceleration sensors were placed to record acceleration vibration signals from the casing.

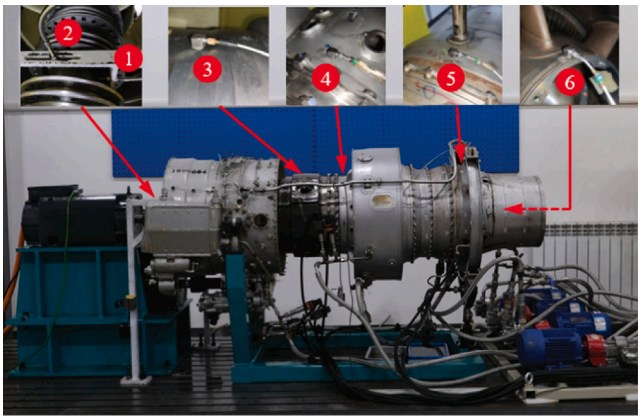


Fig. 5. The real aero-engine test rig and the positions of the 6 testing points. This figure is reproduced from Hou et al. (2023).

Throughout the experimentation, data was collected at a consistent sampling frequency of 25 kHz. This rigorous methodology resulted in a robust dataset comprising 2412 sets of vibration signals. The HIT dataset encompasses three distinct states of the inter-shaft bearings: healthy condition, inner ring fault, and outer ring fault. Each dataset entry is a comprehensive record, including both displacement and acceleration signals sourced from the six measurement points, alongside the speed ratio and labels indicating the fault state. The dataset, originally in .mat format and sourced from Hou et al. (2023) accessible at https://github.com/HouLeiHIT/HIT-dataset, is randomly divided into a training subset comprising 2412 instances and a test subset encompassing 9648 instances. This disproportionate distribution, with a relatively small training set and a large test set, poses a considerable challenge for the model in the context of aero-engine inter-shaft bearing fault diagnosis.

4.2. Experimental settings

The hardware and software environment employed for the experiments comprises the following specifications: The programming language utilized is Python 3.11.5. Deep learning models were constructed using the open-source framework PyTorch 2.1.0 (Paszke et al., 2019), integrated with CUDA V12.4.131 to enhance training speed. The model training was conducted on a system equipped with 8 Nvidia Tesla T4 GPUs, each boasting 16 GB of memory. The operating system in use is Ubuntu 22.04.1 LTS, supported by an Intel® Xeon® Gold 6230R CPU clocked at 2.10 GHz with a total of 104 cores. Additionally, the system is equipped with 128 GB of RAM.

In the experimental process, it is impractical to exhaustively investigate the impact of all hyper-parameters on the ISONet model due to the high computational cost and time requirements. Therefore, a common approach is to fix the less important hyper-parameters and focus on investigating the impact of the more critical hyper-parameters. In this paper, we fixed several hyper-parameters to reduce the search space and computational cost. Specifically, we fixed the number of convolution modules $\mathcal{L} = 4$, the learning rate $\eta = 5 \times 10^{-4}$, the window size of all pooling layers $H^{*(l)} = 3$, the stride of all pooling layers $S^{*(l)} = 3$, the padding of all convolution kernels $P^{(l)} = 0$ and the stride of all convolution kernels $S^{(l)} = 3$. To ensure a fair comparison, all models were trained for a fixed number of epochs $Y = 200$. Fixing these hyper-parameters allowed us to focus on investigating the impact of the more critical hyper-parameters, such as the number of channels, kernel size, and batch size, on the ISONet model's performance. By doing so, we were able to gain insights into the optimal hyper-parameter configuration for the ISONet model under different scenarios.

Table 1
Comparison of 1DCNN and ISONet performances across varying channel configurations and batch sizes. The highest accuracies (%) for each row are highlighted in bold.

#Channels	$H^{(l)}$	1DCNN					ISONet (the proposed)				
		$\|\mathscr{B}\| = 8$	$\|\mathscr{B}\| = 16$	$\|\mathscr{B}\| = 32$	$\|\mathscr{B}\| = 64$	$\|\mathscr{B}\| = 128$	$\|\mathscr{B}\| = 8$	$\|\mathscr{B}\| = 16$	$\|\mathscr{B}\| = 32$	$\|\mathscr{B}\| = 64$	$\|\mathscr{B}\| = 128$
C_1	3	87.03	87.75	86.67	87.95	87.20	84.11	88.22	87.89	**88.00**	87.05
	5	91.17	90.85	91.23	87.53	89.90	**92.20**	90.81	91.57	89.22	89.55
	7	94.55	92.56	92.91	94.36	92.49	**95.51**	93.20	93.21	94.82	92.75
	9	96.90	94.93	95.53	95.96	95.11	**96.91**	94.98	95.62	96.03	95.13
C_2	3	90.44	90.74	90.17	90.80	90.50	90.73	90.27	**91.01**	90.80	90.10
	5	92.90	94.97	94.02	94.37	94.31	93.07	**95.11**	94.04	93.79	94.29
	7	94.27	95.74	96.52	96.67	95.16	94.67	94.76	**97.12**	96.67	94.67
	9	98.09	96.32	98.00	98.20	97.32	**98.26**	95.84	97.77	97.73	97.36
C_3	3	91.85	91.48	92.18	92.07	91.43	91.49	91.88	91.95	**92.55**	90.65
	5	95.92	95.53	94.07	95.90	95.86	96.24	95.79	94.71	95.07	**96.25**
	7	97.89	97.77	96.44	97.20	96.17	**97.93**	97.89	97.09	97.33	96.49
	9	98.74	98.33	97.95	98.30	98.47	98.60	98.51	98.05	98.45	**98.80**
C_4	3	94.61	93.39	93.49	94.43	94.43	**94.86**	94.17	94.58	94.16	94.35
	5	95.37	96.26	96.59	97.17	96.34	95.79	96.75	**97.33**	97.05	96.97
	7	97.14	97.81	98.21	98.03	98.02	97.66	97.93	97.66	97.92	**98.14**
	9	98.60	98.13	98.42	98.62	98.70	98.75	98.23	98.80	98.49	**98.84**
C_5	3	94.41	94.36	94.42	95.67	95.48	94.22	94.40	94.21	94.34	**95.90**
	5	96.41	96.80	97.28	97.26	97.31	96.59	97.20	97.31	**97.50**	97.41
	7	96.95	98.04	97.67	98.06	98.03	97.92	97.84	**98.14**	98.13	98.06
	9	98.46	97.70	98.41	98.55	98.52	98.73	97.90	**98.92**	98.77	98.71

4.3. Metrics

Accuracy, precision, recall, and F1-Score are used metrics for evaluating the performance of ISONet. Accuracy provides an overall measure of the model's performance, while precision and recall provide more specific information about the model's performance with respect to the positive class. F1-Score provides a balanced measure of precision and recall, and is often used when both precision and recall are important.

Accuracy is a performance metric for classification tasks that measures the proportion of correct predictions made by a model. It is calculated as the number of correct predictions divided by the total number of predictions. The formula for accuracy is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{46}$$

where true positives (TP) is the number of correct positive predictions, true negatives (TN) is the number of correct negative predictions, false positives (FP) is the number of incorrect positive predictions, and false negatives (FN) is the number of incorrect negative predictions.

Precision is a performance metric that measures the proportion of correct positive predictions made by a model out of all the positive predictions it made. It is calculated as the number of true positives divided by the sum of true positives and false positives. The formula for precision is:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{47}$$

Recall is a performance metric that measures the proportion of correct positive predictions made by a model out of all the actual positive samples. It is calculated as the number of true positives divided by the sum of true positives and false negatives. The formula for recall is:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{48}$$

F1-Score is a performance metric that combines precision and recall into a single metric. It is calculated as the harmonic mean of precision and recall. The formula for F1-Score is:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{49}$$

4.4. Experiments of hyper-parameter selection & ablation studies

In this section, we conducted a comprehensive evaluation of two CNN architectures, namely 1DCNN and ISONet, across various experimental settings. The primary focus was to assess the impact of different hyper-parameters on the models' performance, specifically exploring the influence of the number of channels (#Channels), including $C_1 = [4, 8, 16, 32]$, $C_2 = [8, 16, 32, 64]$, $C_3 = [16, 32, 64, 128]$, $C_4 = [32, 64, 128, 256]$ and $C_5 = [64, 128, 256, 512]$, as well as the kernel sizes $H^{(l)} \in \{3, 5, 7, 9\}$. To ensure a rigorous analysis, we tested the models using multiple batch sizes $|\mathscr{B}|$ ranging from {8, 16, 32, 64, 128}. The results, presented in Table 1, demonstrate the accuracy achieved by both 1DCNN and ISONet under different configurations. Notably, the highest accuracies for each row are highlighted in bold, facilitating a direct comparison between the two models. This experimental design allows us to draw insights into the optimal configurations for each model, considering factors such as the complexity of the network (determined by #Channels and kernel size $H^{(l)}$) and the computational efficiency (affected by $|\mathscr{B}|$). By analyzing these results, we aim to identify the best practices for tuning these hyper-parameters to maximize the performance of ISONet in practical applications.

Moreover, the direct comparison between the 1DCNN and ISONet architectures also serves as an ablation study, given that the ISONet fundamentally incorporates an additional matrix, denoted as $F^{(l)}$, beyond the structure of the 1DCNN. This comparative analysis allows for an examination of the impact of the supplementary matrix on the performance of the CNN models. By integrating matrix $F^{(l)}$ into the 1DCNN framework to form the ISONet, the ablation study aims to elucidate the contributions and efficacy of this additional component in enhancing the model's capabilities.

Table 1 reveals several key insights regarding the performance of the proposed architecture under different hyper-parameter settings.

- Firstly, we observe that the performance of both 1DCNN and ISONet tends to improve with an increase in #Channels and $H^{(l)}$. This trend is consistent across different batch sizes $|\mathscr{B}|$, indicating that a more complex model with a larger capacity is better able to capture the underlying patterns in the data. However, it is important to note that increasing the model complexity also increases the computational requirements and the risk of overfitting, especially when the dataset size is limited.
- Upon examination of the results, it is evident that the performance of both models is influenced by $|\mathscr{B}|$. In essence, the performance of the ISONet is relatively stable across different batch sizes, especially when using larger #Channels and $H^{(l)}$. Although there are specific variations in the numerical values, the overall trend indicates that the performance does not substantially degrade or improve with changing batch sizes. This

stability can be attributed to the over-parameterized nature of ISONet, which allows it to learn more complex patterns and generalize better across different batch sizes. In contrast, 1DCNN exhibits a more significant variation in performance with changing batch sizes under the same conditions of #Channels and $H^{(l)}$. The model's performance may improve or degrade substantially depending on the specific $|\mathscr{B}|$ used. This sensitivity to $|\mathscr{B}|$ can be attributed to the standard convolution kernel used in the 1DCNN, which may not be able to learn complex patterns as effectively as the over-parameterized kernel used in the ISONet

- The ISONet, which is an over-parameterized convolution that adds a parameter tensor to each input channel of the standard convolution kernel, demonstrates superior performance than 1DCNN. Notably, for C_5, the ISONet model achieves the highest accuracy of 98.92% with $H^{(l)} = 9$ and $|\mathscr{B}| = 32$, outperforming the 1DCNN model's highest accuracy of 98.55%. Similarly, for C_1, C_2, C_3, and C_4, the ISONet model achieves the highest accuracy with specific configurations. This suggests that the additional parameters introduced by the input spatial over-parameterization techniques in ISONet effectively enhance the model's representational power.

In conclusion, ISONet demonstrates superior performance over 1DCNN in various configurations, with the optimal hyper-parameter configuration dependent on the specific channel configuration. The results highlight the potential of over-parameterized convolutions in improving model performance. ISONet model is also more robust to changes in $|\mathscr{B}|$ under conditions of larger convolution kernels and channel numbers. This robustness is an essential factor to consider when selecting a model for a specific task, as it can lead to more consistent performance across different batch sizes. However, further research is needed to investigate the underlying mechanisms that contribute to the observed performance trends and to develop strategies for selecting the optimal $|\mathscr{B}|$ for different models and tasks.

4.5. Comparing with different optimizers

To assess the optimization efficacy of the Nesterov-accelerated optimizer Adan for training the ISONet, we conducted a comparative analysis with other popular optimizers including SGD (Bottou, 2010), Adadelta (Zeiler, 2012), Adamax (Kingma & Ba, 2015), AdaBound (Luo et al., 2019), AdaBelief (Zhuang et al., 2020), and CAME (Luo et al., 2023). The ISONet model was configured with a channel count of C_5, a kernel size $H^{(l)} = 9$, and a batch size $|\mathscr{B}| = 32$. For a fair comparison, all optimizers were trained for 200 epochs. The evolution of accuracy during the training process is presented in Fig. 6. This figure offers a visual representation of how each optimizer performs in terms of convergence speed, stability, and overall performance. By comparing the curves, we can gain insights into the relative strengths and weaknesses of each optimizer for the specific task of training the ISONet.

We observed that the Adan optimizer exhibited a faster convergence rate on the training dataset compared to other optimizers. Notably, the final accuracy on the training set achieved by Adan was identical to that of CAME, Adamax, AdaBound, and AdaBelief, all converging to 100%. More importantly, on the test set, Adan demonstrated both a higher accuracy asymptote and a lower loss function asymptote. These results strongly suggest the unique advantages of the Adan optimizer in terms of both generalization and optimization efficiency. Consequently, the Adan optimizer appears to be a highly suitable choice for the Aero-engine System Inter-shaft Bearing Fault Diagnosis task based on the ISONet model.

To demonstrate the implicit acceleration effects of incorporating input spatial over-parameterization, we have depicted the variation in the test set loss functions during the training process of the ISONet and 1DCNN under different optimizer conditions. As illustrated in Fig. 7, the fluctuations in the loss function of the ISONet are significantly less pronounced than those of the 1DCNN, and it converges at a faster rate. This observation indicates that the ISONet possesses superior stability compared to the 1DCNN. Consequently, the introduction of input spatial over-parameterization, which implicitly incorporates momentum and adaptive learning rate strategies, is found to be beneficial in enhancing the stability of training and accelerating convergence.

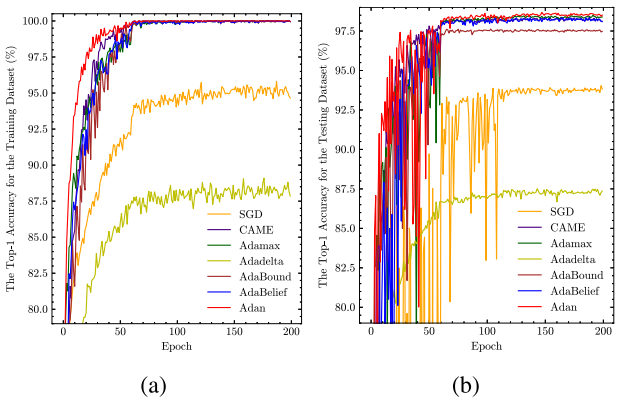


Fig. 6. The accuracy curves of ISONet with different optimizers. (a) The training accuracy curves, (b) The testing accuracy curves.

Particularly noteworthy in Fig. 7 is that optimizers such as Adamax, AdaBelief, CAME, and Adam inherently possess explicit momentum and adaptive learning rate mechanisms. However, upon the additional implicit introduction of momentum and adaptive learning rates through input spatial over-parameterization, a marked improvement in performance is observed. Therefore, we can empirically conclude that the implicit momentum and adaptive learning rates introduced by input spatial over-parameterization offer distinct advantages over existing explicit mechanisms, rendering them irreplaceable in this context.

This comparative analysis within the scope of our research underscores the unique contribution of input spatial over-parameterization in optimizing the training process, suggesting that it may provide a novel and effective approach to enhancing the performance of neural network training algorithms.

4.6. Comparing with different activation functions

To further investigate the impact of different activation functions on the performance of the ISONet, we conducted an experiment with identical hyper-parameter settings to the previous study. Keeping the model architecture fixed and using the same training approach with the Adan optimizer and consistent hyper-parameters, we replaced the Mish activation function in the ISONet with various well-known activation functions: TanhExp (Liu & Di, 2021), Serf (Nag et al., 2023), SELU (Klambauer et al., 2017), RReLU (Xu et al., 2015), ReLU6 (Howard et al., 2017), ReLU (Nair & Hinton, 2010), PReLU (He et al., 2015), GELU (Hendrycks & Gimpel, 2016), and AReLU (Chen et al., 2020).

The experimental results are summarized in Table 2. To facilitate a comprehensive analysis of the performance of each activation function across different metrics, we visualized the experimental outcomes using a radar chart, as presented in Fig. 8.

From Fig. 8 and Table 2, it is evident that while the Mish activation function performs relatively poorer in terms of the Recall for the health and outer ring fault states compared to the other activation functions, it exhibits a clear advantage for the inner ring fault and overall weighted averaged Recall. In terms of the Precision, the Mish activation function demonstrates its superiority primarily in the overall weighted averaged Precision, although it may not achieve the best classification performance for specific fault states.

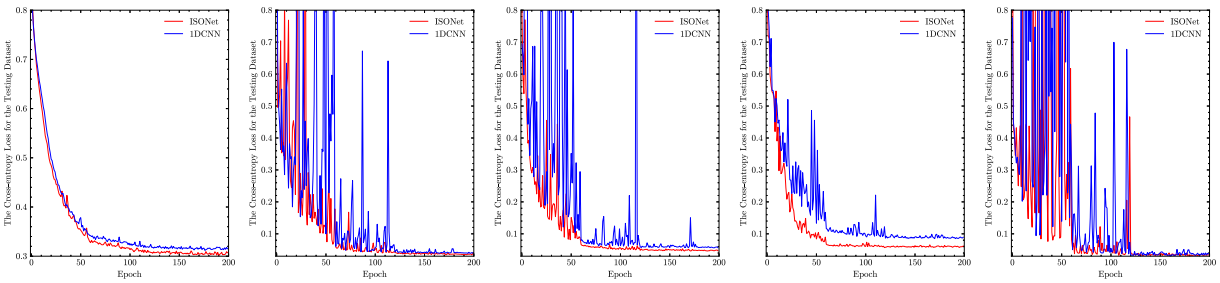


Fig. 7. The loss curves of ISONet and 1DCNN with different optimizers. (a) Adadelta, (b) Adamax, (c) AdaBelief, (d) CAME, and (e) Adan.

Table 2
Results on different activation functions. The highest metrics (%) for each row are highlighted in bold.

Metrics	Mish (Misra, 2020)	TanhExp (Liu & Di, 2021)	Serf (Nag, Bhattacharyya, Mukherjee, & Kundu, 2023)	SELU (Klambauer, Unterthiner, Mayr, & Hochreiter, 2017)	RReLU (Xu, Wang, Chen, & Li, 2015)	ReLU6 (Howard et al., 2017)	ReLU (Nair & Hinton, 2010)	PReLU (He et al., 2015)	GELU (Hendrycks & Gimpel, 2016)	AReLU (Chen, Li, & Xu, 2020)
Precision (Health)	98.849	98.691	98.233	98.288	**98.864**	97.511	98.294	98.229	98.151	98.487
Precision (Inner Ring Fault)	98.842	98.521	98.838	**98.935**	98.416	97.958	97.260	98.156	97.887	98.691
Precision (Outer Ring Fault)	99.263	98.927	99.146	99.487	97.312	97.691	98.230	**99.541**	99.310	98.755
Precision (Weighted)	**98.924**	98.664	98.656	98.782	98.387	97.731	97.850	98.443	98.257	98.622
Recall (Health)	99.030	98.821	99.057	**99.292**	98.035	98.559	98.139	98.821	98.742	98.926
Recall (Inner Ring Fault)	**99.529**	99.132	99.132	99.107	98.636	97.545	98.586	99.008	98.810	99.082
Recall (Outer Ring Fault)	97.333	97.278	96.722	96.944	**98.556**	96.389	95.556	96.333	95.944	96.944
Recall (Weighted)	**98.922**	98.663	98.653	98.777	98.383	97.730	97.844	98.435	98.248	98.621
F1-Score (Health)	**98.940**	98.756	98.643	98.788	98.447	98.032	98.217	98.524	98.445	98.706
F1-Score (Inner Ring Fault)	**99.184**	98.826	98.985	99.021	98.526	97.751	97.918	98.580	98.346	98.886
F1-Score (Outer Ring Fault)	**98.289**	98.095	97.919	98.199	97.930	97.036	96.874	97.911	97.598	97.841
F1-Score (Weighted)	**98.921**	98.662	98.651	98.775	98.384	97.729	97.842	98.433	98.246	98.620

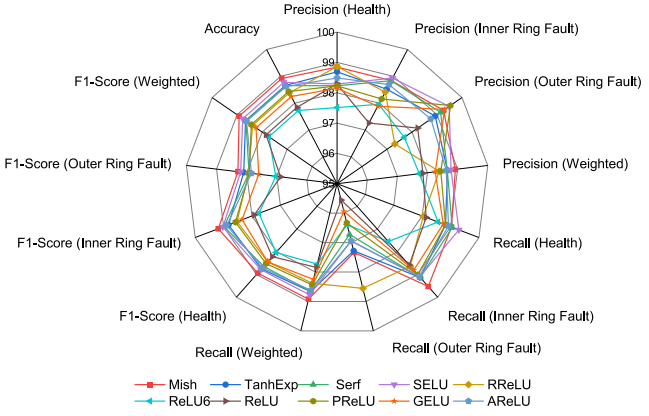


Fig. 8. Results of ISONet with different activation functions.

However, the Mish activation function stands out as the overall leader in the F1-Score metric, demonstrating comprehensive superiority across all three categories of aero-engine system inter-shaft bearing fault diagnosis. This comprehensive advantage underscores the remarkable capabilities of the Mish activation function for the task of aero-engine system inter-shaft bearing fault diagnosis. These findings suggest that the choice of activation function can have a significant impact on the performance of the ISONet for fault diagnosis in aero-engine systems.

4.7. Comparing with state-of-the-art models

To validate the effectiveness of the proposed ISONet for aero-engine system inter-shaft bearing fault diagnosis and compare it with other state-of-the-art (SOTA) 1DCNN networks and new models, we consider the following SOTA models for comparison:

- Channel Attention-based 1DCNN (CA-1DCNN) (Xiang et al., 2021): We add an aggregation perception-recalibration (APR) channel attention module after each convolutional module while keeping the same architecture as ISONet. The APR module has two hyper-parameters: a reduction ratio μ in the dimensionality-reduction layer and the number of FC layers in the representative enhancing layer parameterized by η. We consider six combinations of these hyper-parameters, resulting in six versions of CA-1DCNN: CA-1DCNN-A ($\eta = 1$, $\mu = 8$), CA-1DCNN-B ($\eta = 1$, $\mu = 16$), CA-1DCNN-C ($\eta = 2$, $\mu = 8$), CA-1DCNN-D ($\eta = 2$, $\mu = 16$), CA-1DCNN-E ($\eta = 3$, $\mu = 8$), and CA-1DCNN-F ($\eta = 3$, $\mu = 16$).
- Multi-scale Group-Fusion 1DCNN (MSGF-1DCNN) (Xiang et al., 2022): MSGF-1DCNN uses multi-scale group (MSG) 1D-Conv and PW-Conv instead of standard convolution layers to extract and fuse multi-scale features. MSG 1D-Conv uses multiple group convolution kernels of varying scales to extract features at different levels of detail from the input data, enabling the model to capture more informative spatial information. PW-Conv is utilized to linearly fuse the extracted multi-scale features, enabling the model to effectively combine the local and global features and enhance the discriminative ability of the features. The hyper-parameter of MSGF-1DCNN is the number of different-scale group convolutions Ω in MSG 1D-Conv. We consider four versions of MSGF-1DCNN: MSGF-1DCNN-A ($\Omega = 2$), MSGF-1DCNN-B ($\Omega = 3$), MSGF-1DCNN-C ($\Omega = 4$), and MSGF-1DCNN-D ($\Omega = 5$).

Table 3
Performance of the compared SOTA models in terms of weighted averaged Precision, Recall, and F1-Score. The maximum value in each column is indicated by bolding.

Model	Year	Variants	Precision (%)	Recall (%)	F1-Score (%)
CA-1DCNN	2021	A	98.033	98.031	98.021
		B	97.801	97.792	97.795
		C	97.934	97.937	97.934
		D	97.913	97.906	97.904
		E	97.900	97.875	97.869
		F	97.771	97.751	97.749
MSGF-1DCNN	2022	A	92.123	92.102	92.101
		B	92.579	92.579	92.579
		C	93.532	93.532	93.528
		D	94.751	94.745	94.744
GFAC-1DCNN	2023	–	86.775	86.733	86.739
Transformer	2023	A	66.092	66.034	65.812
		B	67.040	67.413	67.107
		C	70.526	70.782	70.562
		D	61.068	60.759	60.777
		E	66.873	67.143	66.932
1DCNN	–	–	98.419	98.414	98.412
ISONet (the proposed)	2024	–	**98.924**	**98.922**	**98.921**

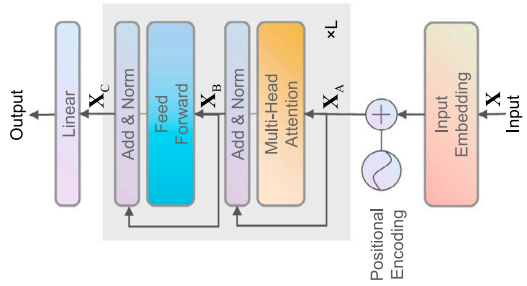


Fig. 9. The Transformer (Xiang et al., 2023b) architecture used for comparing with 1DCNNs.

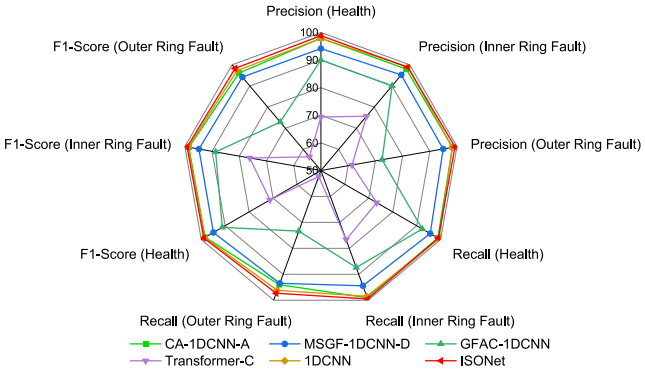


Fig. 10. The radar chart of illustrating the performance of various SOTA models in terms of Precision, Recall, and F1-Score for a specific fault category.

- Group-Fusion 1DCNN with Layer-Wise Auxiliary Classifiers (GFAC-1DCNN) (Xiang et al., 2023a): GFAC-1DCNN reduces model complexity through group convolution and linear fusion layers and combines layer-wise auxiliary classifiers to fuse features from different layers for classification.
- Transformer Model: We use a network structure similar to the Transformer backbone in CsiTransformer (Xiang et al., 2023b), as shown in Fig. 9. The four hyper-parameters in Transformer are patch size (ps), embedding dimension (ed), depth (L), and the number of heads (nh). We consider five models with different parameter quantities: Transformer-A ($ps = 16, ed = 768, L = 1, nh = 12$), Transformer-B ($ps = 16, ed = 768, L = 2, nh = 12$), Transformer-C ($ps = 16, ed = 768, L = 3, nh = 12$), Transformer-D ($ps = 9, ed = 144, L = 2, nh = 12$), and Transformer-E ($ps = 9, ed = 768, L = 2, nh = 12$).

By comparing the performance of ISONet with these models, we aim to demonstrate the effectiveness and superiority of the proposed ISONet for aero-engine system inter-shaft bearing fault diagnosis.

Table 3 presents the performance of the compared SOTA models in terms of weighted averaged Precision, Recall, and F1-Score, which describe the overall classification performance of the models across the three categories. Fig. 10 shows the more detailed performance of the three metrics. From Table 3 and Fig. 10, it can be observed that overall, the best variant for CA-1DCNN is A, for MSGF-1DCNN, the best-performing variant is D, and for Transformer, the best-performing variant is C. CA-1DCNN, MSGF-1DCNN, GFAC-1DCNN, and ISONet are all improvements of 1DCNN, but the overall performance of CA-1DCNN, MSGF-1DCNN, and GFAC-1DCNN is lower than 1DCNN, indicating that the corresponding improvement strategies of these three models are not suitable for the aero-engine system inter-shaft bearing fault diagnosis task. However, from Table 3 and Fig. 10, it can be seen that ISONet outperforms 1DCNN and Transformer in both overall and specific category performance, indicating that the proposed over-parameterization strategy has good adaptability to the aero-engine system inter-shaft bearing fault diagnosis task. Although Transformer has achieved good results in computer vision, time series prediction, and other fields in recent years, its performance is far lower than the 1DCNN-based models for the aero-engine system inter-shaft bearing fault diagnosis task.

Fig. 11 illustrates the confusion matrices for the Aero-engine System Inter-shaft Bearing Fault Diagnosis, employing diverse State-of-the-Art (SOTA) DL methodologies. The visualization clearly delineates the misclassification instances for each SOTA model. Notably, the ISONet exhibits a significantly lower quantity of misclassified samples across all categories, particularly for the two fault types. It is posited that the cost of misidentifying a healthy condition as a fault is substantially less than that of misclassifying a fault as a healthy state. Hence, this suggests that the proposed ISONet is more efficacious, capable of circumventing potential safety hazards.

Specifically, for the ISONet, the misclassification details are as follows: within the Health category, a substantial number of Health samples are categorized as Inner Ring Faults. Conversely, for the Inner Ring Fault category, a considerable portion of Inner Ring Fault samples

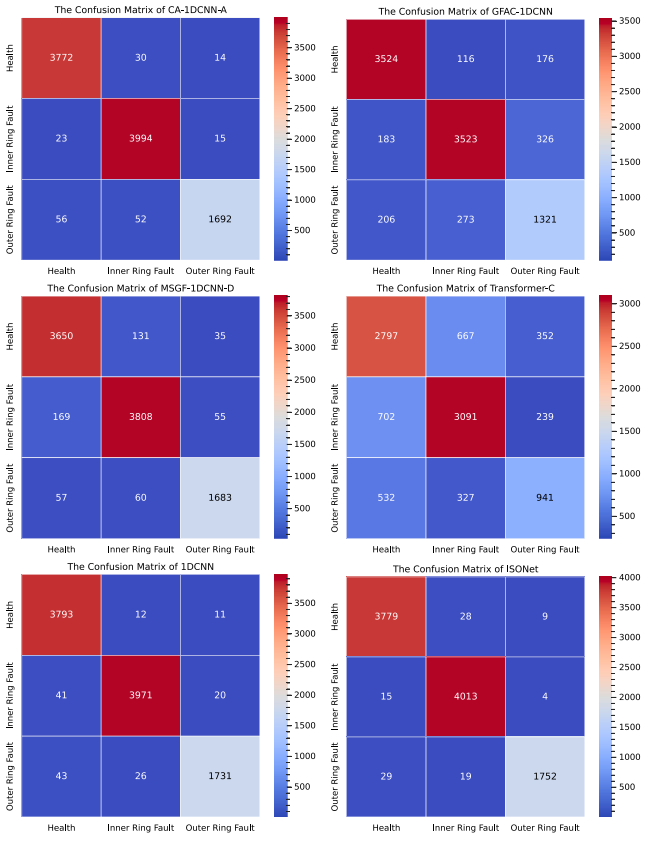


Fig. 11. Confusion matrices for aero-engine system inter-shaft bearing fault diagnosis using various SOTA DL methods. X-axis: predicted labels; Y-axis: ground-truth labels.

are classified as Health. For the Outer Ring Fault category, a notable number of Outer Ring Fault samples are misidentified as Health. Consequently, the challenge in fault identification is primarily concentrated on the differentiation between Inner Ring Fault and Health, which are prone to mutual misclassification, whereas the distinction between the two fault types is relatively more discernible.

To provide a comprehensive assessment of the performance across various models, we constructed the ROC curves and calculated the AUC for each model, as depicted in Fig. 12. From Fig. 12, it is evident that when different categories are considered as the positive class, ISONet consistently demonstrates the highest overall performance. This suggests that the ISONet model possesses superior discriminative ability and generalization, effectively distinguishing between the positive and negative classes across different fault conditions in the aero-engine system inter-shaft bearing diagnosis. The enhanced performance of ISONet underscores its potential as a robust diagnostic tool, with implications for reliable fault detection and system maintenance.

Fig. 13 shows the relationship between the classification accuracy (%) and the number of parameters (M) of the five 1DCNNs, namely CA-1DCNN-A, MSGF-1DCNN-D, GFAC-1DCNN, ISONet, and 1DCNN. From Fig. 13, it can be seen that GFAC-1DCNN has the largest number of parameters among the five 1DCNNs due to the presence of a classifier based on a fully-connected layer after each convolution module, but its classification accuracy is the worst. MSGF-1DCNN has the smallest number of parameters, but its classification accuracy is significantly lower than CA-1DCNN-A, 1DCNN, and ISONet. Since GFAC-1DCNN and MSGF-1DCNN are both based on group convolution architectures, while CA-1DCNN-A, 1DCNN, and ISONet are based on standard convolution, this can to some extent indicate that group convolution is not suitable for the aero-engine system inter-shaft bearing fault diagnosis task.

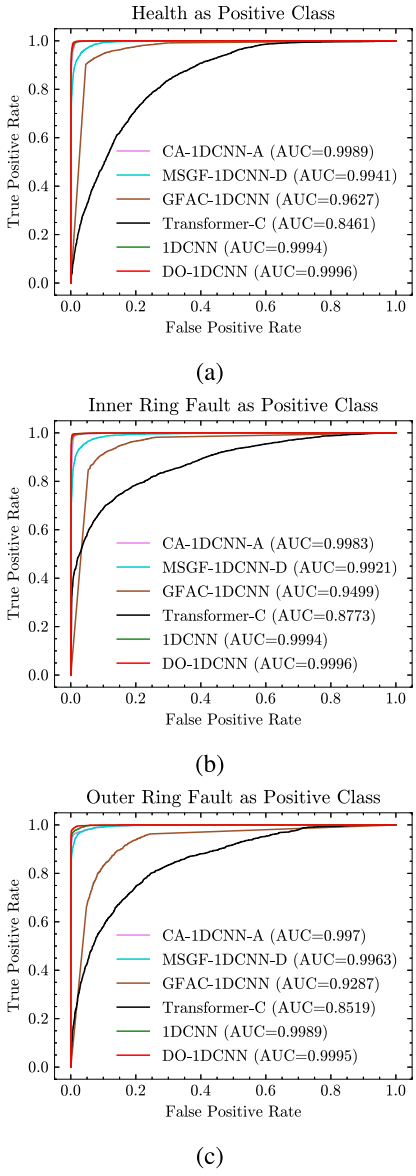


Fig. 12. Comparative analysis of ROC curves and AUC for aero-engine inter-shaft bearing fault diagnosis across multiple fault categories, utilizing distinct SOTA DL models. Each curve represents the performance metrics when a specific fault category is defined as the positive class, thereby providing a comprehensive evaluation of model sensitivity and specificity in detecting various fault conditions within aero-engine systems. (a) health as positive class, (b) inner ring fault as positive class, and (c) outer ring fault as positive class.

CA-1DCNN-A and ISONet both increase the number of parameters compared to 1DCNN, but their effects on 1DCNN are completely different. The classification accuracy of CA-1DCNN-A is lower than 1DCNN, while the classification accuracy of ISONet is higher than 1DCNN, which further demonstrates that over-parameterized convolution has a high promoting effect on the aero-engine system inter-shaft bearing fault diagnosis task.

To further elucidate the performance discrepancies among the compared models, we introduce Fig. 14, which maps the models onto the PQS-FP coordinate system (Xiang et al., 2024). This framework categorizes models into distinct regions based on their behavior as the number of parameters increases: the Underfitting Attenuation Region (UAR) and the Overfitting Exacerbation Region (OER). As shown in Fig.

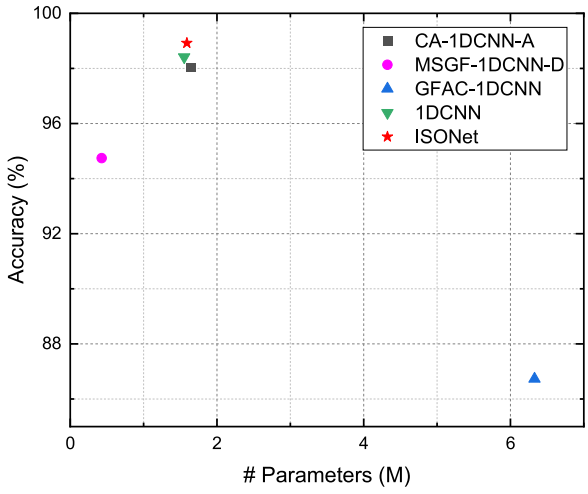


Fig. 13. The relationship between the classification accuracy (%) and the number of parameters (M) of the five SOTA 1DCNNs.

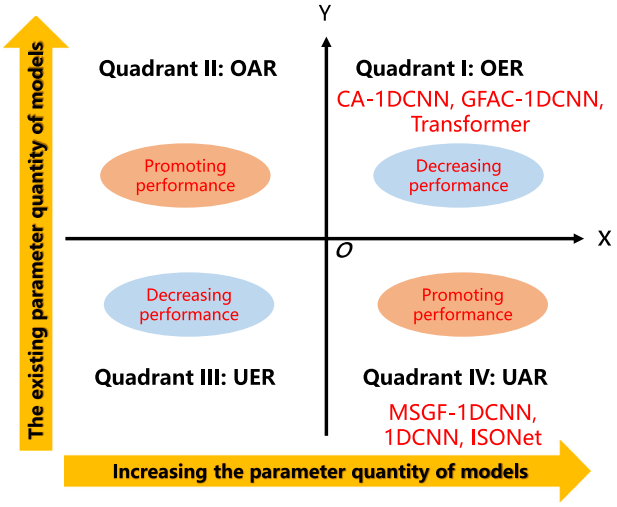


Fig. 14. The positioning of various models in the PQS-FP coordinate system (Xiang et al., 2024).

14, MSGF-1DCNN, standard 1DCNN, and ISONet reside in the UAR, where increasing the number of parameters alleviates underfitting, leading to improved classification accuracy. This aligns with the trends observed in Fig. 13, where ISONet achieves the highest accuracy by effectively leveraging its increased parameter count to capture complex patterns in the data without overfitting.

In contrast, CA-1DCNN, GFAC-1DCNN, and Transformer-based models fall into the OER, where additional parameters exacerbate overfitting, particularly in scenarios with limited training data or high feature complexity. This explains their inferior performance despite their architectural enhancements, as seen in Fig. 13. The overparameterization in these models leads to a degradation in generalization, as they tend to overfit the training data, resulting in lower accuracy on the test set.

The distinction between UAR and OER underscores the importance of balancing model complexity and generalization. ISONet's design, which incorporates input spatial over-parameterization, strikes this balance effectively, enabling it to outperform other methods.

It is crucial to highlight that, although the ISONet incorporates additional parameters compared to the standard 1DCNN, these extra

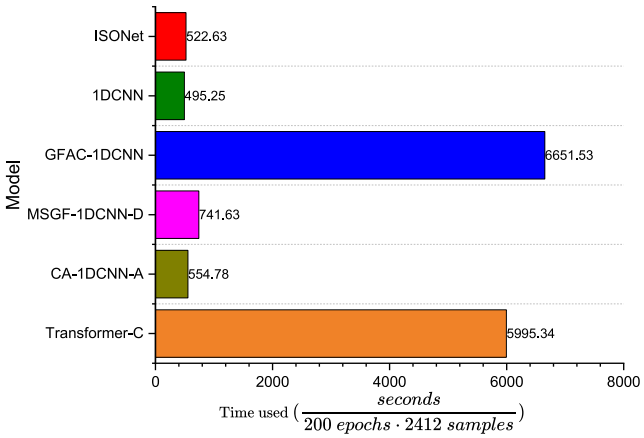


Fig. 15. Time used for training different models.

parameters are exclusive to the training process. As demonstrated by Eq. (10), once the model has been trained, we can utilize the folding operation to collapse the $F^{(l)}$ and $G^{(l)}$ tensors. This folding operation ensures that, in practical inference scenarios, the ISONet essentially maintains the same parameter quantity as the 1DCNN. Consequently, the ISONet offers the advantage of increased learning capacity during training, without compromising computational efficiency during deployment.

Fig. 15 illustrates the total training time required for various models when trained on 2412 samples over 200 epochs. It can be seen that ISONet, with a training time of 522.63 s, demonstrates a modest increase in training duration compared to the 1DCNN, which recorded a time of 495.25 s. This slight augmentation of approximately 57.38 s indicates that the ISONet incurs a minimal penalty in training time while potentially offering enhanced capabilities. Furthermore, when juxtaposed with models such as Transformer-C, CA-1DCNN-A, and GFAC-1DCNN, ISONet exhibits a significant advantage in training speed. This comparative analysis underscores the efficiency of ISONet in the context of high-dimensional sequence modeling, suggesting that it provides a commendable trade-off between training efficiency and model performance.

4.8. Comparison across datasets with limited training data

In real-world scenarios, it is conceivable that there may be a dearth of fault training data. Therefore, it is imperative to verify the performance of models under varying quantities of training samples. To this end, we randomly selected 20% of all samples to be fixed as a validation dataset. Subsequently, we selected a certain proportion of the remaining samples to train the model, with the proportion ranging from {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. We trained all models for 30 epochs and recorded the optimal results during the training process, repeating the experiment five times to take the average. In addition to recording the metrics of Accuracy, Precision, Recall, and F1-Score, we also documented the value of the loss function $\mathcal{J}$ and AUC. The results are presented in Table 4.

An examination of Table 4 reveals that as the sampling rate of the samples increases, the values of the six evaluated metrics for the compared models correspondingly improve. This trend aligns with our intuition that an expanded dataset can significantly enhance the accuracy of fault detection based on deep neural networks. The Transformer-C model consistently underperforms, indicating that models predicated on attention mechanisms may not be well-suited for the task of aero-engine system inter-shaft bearing fault diagnosis. The CA-1DCNN-A

Table 4
The performance of various models under different conditions of training sample sizes. The values of all metrics are the averaged outcomes following the conduct of five experimental trials.

Model	Metric	The Sampling Ratio									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Transformer-C	Accuracy (%)	57.7114	64.5937	63.8474	76.6998	72.7197	81.5920	78.3167	74.6683	84.1211	84.1625
	Precision (%)	57.8393	64.9841	63.5921	76.6100	72.8370	81.5877	78.2856	74.7890	84.1421	84.1266
	Recall (%)	57.7114	64.5937	63.8474	76.6998	72.7197	81.5920	78.3167	74.6683	84.1211	84.1625
	F1-Score (%)	55.7997	64.7151	63.5927	76.6468	72.7598	81.5063	78.2602	74.7174	84.1259	84.1207
	AUC	0.7026	0.7939	0.7878	0.9009	0.8743	0.9390	0.9217	0.8992	0.9496	0.9486
	$\mathcal{J}$	0.9853	1.5992	1.8306	0.8480	1.2815	0.5234	0.7111	0.7349	0.6293	0.9880
CA-1DCNN-A	Accuracy (%)	95.2322	97.8027	98.9635	99.0879	99.5025	99.4196	99.6683	99.8342	99.6683	99.8871
	Precision (%)	95.2492	97.8050	98.9633	99.0889	99.5023	99.4200	99.6684	99.8344	99.6688	99.8872
	Recall (%)	95.2322	97.8027	98.9635	99.0879	99.5025	99.4196	99.6683	99.8342	99.6683	99.8871
	F1-Score (%)	95.2376	97.8036	98.9630	99.0879	99.5020	99.4196	99.6680	99.8342	99.6684	99.8871
	AUC	0.9925	0.9981	0.9990	0.9992	0.9995	0.9997	0.9998	1.0000	0.9999	1.0000
	$\mathcal{J}$	0.1870	0.0971	0.0661	0.0552	0.0372	0.0396	0.0230	0.0117	0.0236	0.0101
MSGF-1DCNN-D	Accuracy (%)	88.8889	94.1542	96.9735	98.2172	98.4660	98.5489	99.0464	99.1294	99.4610	99.6269
	Precision (%)	88.9371	94.1966	97.0328	98.2198	98.4658	98.5518	99.0465	99.1298	99.4611	99.6276
	Recall (%)	88.8889	94.1542	96.9735	98.2172	98.4660	98.5489	99.0464	99.1294	99.4610	99.6269
	F1-Score (%)	88.8979	94.1635	96.9794	98.2180	98.4658	98.5490	99.0463	99.1291	99.4608	99.6269
	AUC	0.9727	0.9906	0.9978	0.9988	0.9990	0.9993	0.9995	0.9996	0.9999	1.0000
	$\mathcal{J}$	0.3483	0.2192	0.1138	0.0735	0.0604	0.0520	0.0363	0.0378	0.0187	0.0128
GFAC-1DCNN	Accuracy (%)	70.7297	87.3964	88.3914	89.3864	89.2620	90.2985	90.0083	91.6667	91.6667	91.3764
	Precision (%)	70.9663	87.1091	88.3554	89.4211	89.1762	90.3569	90.1542	91.5440	91.8408	91.4419
	Recall (%)	70.7297	87.3964	88.3914	89.3864	89.2620	90.2985	90.0083	91.6667	91.6667	91.3764
	F1-Score (%)	70.6787	87.1943	88.3726	89.3870	89.1898	90.3198	90.0714	91.5862	91.7284	91.3997
	AUC	0.8216	0.9367	0.9404	0.9510	0.9691	0.9542	0.9657	0.9629	0.9750	0.9605
	$\mathcal{J}$	30.8900	12.7131	13.0591	11.1353	2.8343	10.9095	5.5059	7.2835	3.7390	8.2459
1DCNN	Accuracy (%)	94.0298	98.4489	98.8806	99.3781	99.6683	99.5025	99.5025	99.7098	99.7171	99.8342
	Precision (%)	94.0360	98.4506	98.8797	99.3780	99.6687	99.5050	99.5033	99.7100	99.7172	99.8345
	Recall (%)	94.0298	98.4489	98.8806	99.3781	99.6683	99.5025	99.5025	99.7098	99.7171	99.8342
	F1-Score (%)	94.0299	98.4489	98.8792	99.3776	99.6682	99.5026	99.5021	99.7096	99.7171	99.8342
	AUC	0.9898	0.9990	0.9994	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	1.0000
	$\mathcal{J}$	0.1764	0.0549	0.0389	0.0242	0.0156	0.0194	0.0176	0.0139	0.0060	0.0079
ISONet (the proposed)	Accuracy (%)	97.3466	98.5075	99.3366	99.4196	99.6683	99.5439	99.8342	99.8442	99.8756	99.9585
	Precision (%)	97.3383	98.5091	99.3372	99.4208	99.6693	99.5449	99.8349	99.8449	99.8760	99.9586
	Recall (%)	97.3466	98.5075	99.3366	99.4196	99.6683	99.5439	99.8342	99.8442	99.8756	99.9585
	F1-Score (%)	97.3354	98.5070	99.3354	99.4190	99.6683	99.5436	99.8342	99.8441	99.8757	99.9585
	AUC	0.9974	0.9989	0.9997	0.9998	0.9999	0.9999	1.0000	1.0000	1.0000	1.0000
	$\mathcal{J}$	0.0952	0.0533	0.0273	0.0220	0.0141	0.0162	0.0076	0.0096	0.0067	0.0048

model demonstrates superior performance relative to the 1DCNN, suggesting that the incorporation of channel attention mechanisms is effective within this experimental framework. Conversely, the MSGF-1DCNN-D and GFAC-1DCNN models exhibit diminished performance compared to the 1DCNN, implying that the introduction of grouped convolutions and multiple auxiliary classifiers may have a detrimental impact on the task at hand. Notably, the ISONet outperforms other methodologies across all metrics, achieving an identification accuracy of over 97% even at a sample adoption rate as low as 0.1.

To delve deeper into the classification boundary dynamics of various DL models under different sampling rates, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten & Hinton, 2008) with different distance metrics to dimensionally reduce and graphically represent the outputs of several deep neural network models. Fig. 16, Fig. 17, Fig. 18, and Fig. 19 lucidly delineate this scenario. A clear observation from Fig. 16, Fig. 17, Fig. 18, and Fig. 19 is that as the volume of training samples increases, the intra-class compactness and inter-class separability of the models incrementally improve. Interestingly, our investigation reveals that the ISONet consistently achieves discernible classification boundaries across various distance metrics, such as Chebyshev, Cosine, Euclidean, and Manhattan. Given that these metrics capture different aspects of similarity, this finding suggests that the classification boundaries produced by ISONet possess a robust and consistent level of discriminability across different fault categories. This consistency in performance across different distance metrics is a testament to the model's ability to generalize and effectively delineate the decision boundaries for fault classification, regardless of the specific measure of similarity employed. Such a property is highly desirable in the context of fault diagnosis, where the model must reliably identify and distinguish between various fault types under potentially varying conditions and data representations.

However, the clarity of classification boundaries for Transformer-C and GFAC-1DCNN does not exhibit a marked enhancement. In contrast, MSGF-1DCNN-D, CA-1DCNN-A, 1DCNN, and ISONet demonstrate a pronounced improvement in boundary clarity at sampling rates less than or equal to 0.6. Beyond this threshold, however, the distinctness of classification boundaries remains largely invariant. This indicates that when the volume of samples reaches a certain threshold, the effectiveness of fault diagnosis encounters a performance plateau, beyond which advanced CNNs such as MSGF-1DCNN-D, CA-1DCNN-A, 1DCNN, and ISONet tend to perform at similar levels. As can also be observed from Table 4, the performance improvement of ISONet over MSGF-1DCNN-D, CA-1DCNN-A, and 1DCNN is not significant when the sampling rate exceeds 0.6. Thus, the ISONet's primary advantage lies in its enhanced intra-class compactness at lower sampling rates, thereby conferring greater reliability to fault diagnosis under conditions of sample scarcity. This trait fortifies the ISONet's performance, ensuring accurate and dependable diagnostic outcomes even when faced with limited data.

Additionally, from Fig. 16, Fig. 17, Fig. 18, and Fig. 19, it is observable that ISONet, akin to the majority of SOTA 1DCNNs, still encounters a challenge when the sample sampling rate is relatively low (for instance, at 0.1 and 0.2). Although ISONet exhibits commendable intra-class compactness, there remains a necessity to enhance the inter-class distinctiveness. This observation suggests that while the model demonstrates robust clustering within classes, the differentiation between distinct classes could be further optimized to bolster the

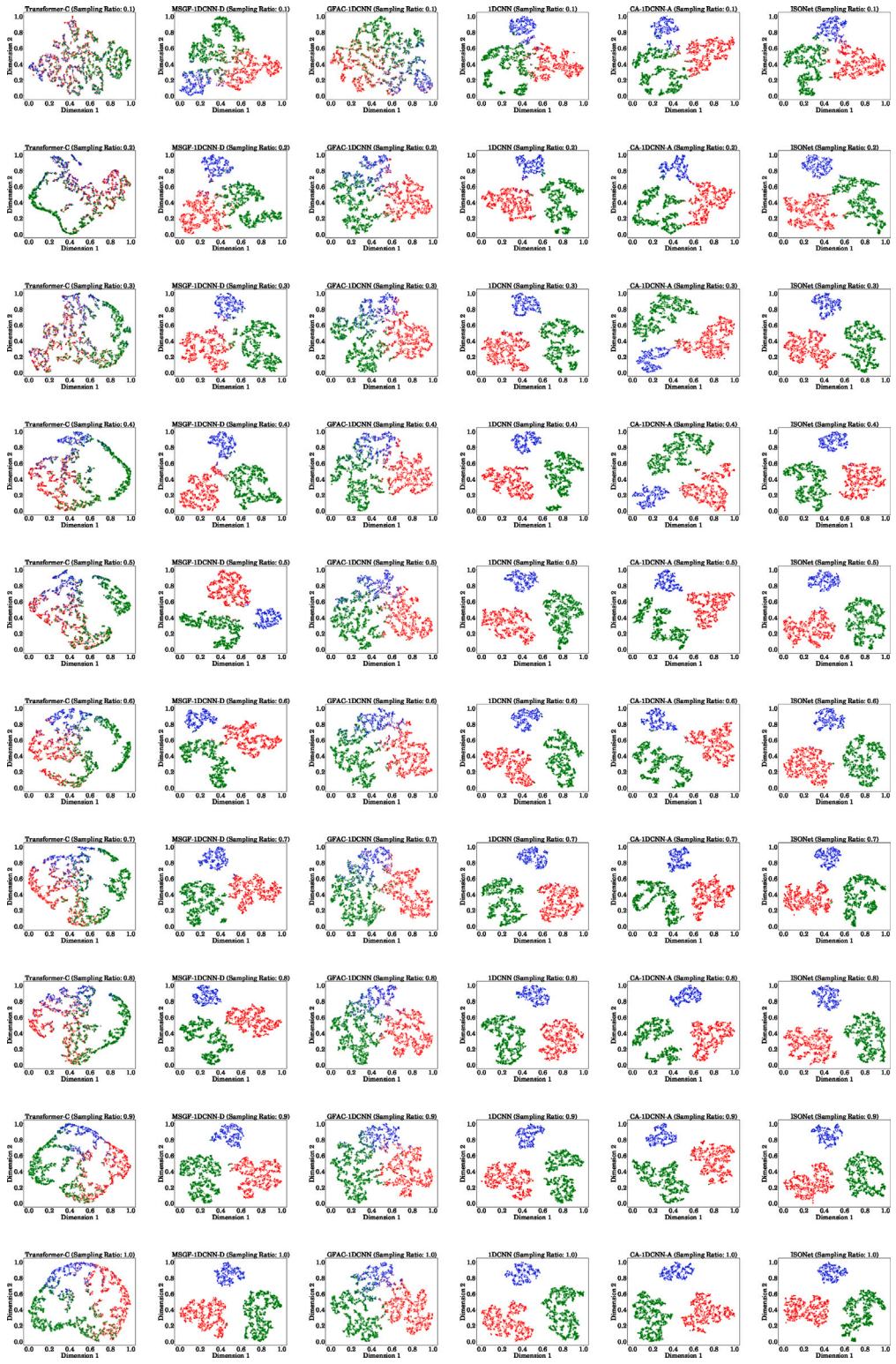


Fig. 16. Visualization of deep neural network model outputs via t-SNE (Maaten & Hinton, 2008) reduction to two dimensions with Chebyshev distance. Red: Health, Green: Inner Ring Fault and Blue: Outer Ring Fault.

diagnostic capabilities of the network in scenarios characterized by limited sampling.

In conclusion, this comparative analysis within the scope of our research underscores the unique advantages of incorporating spatial over-parameterization in the context of fault diagnosis for aero-engine systems, highlighting its potential as an indispensable strategy for improving diagnostic accuracy under constrained sampling conditions.

4.9. Comparison with the archived results on HIT dataset

Given that the HIT dataset has only been examined in Berghout et al. (2023) and Hou et al. (2023), which lack sufficient detail to replicate the models from the papers, we have redrawn the HIT dataset according to the 70% training and 30% testing split method described in Berghout et al. (2023) and Hou et al. (2023), retrained it using the

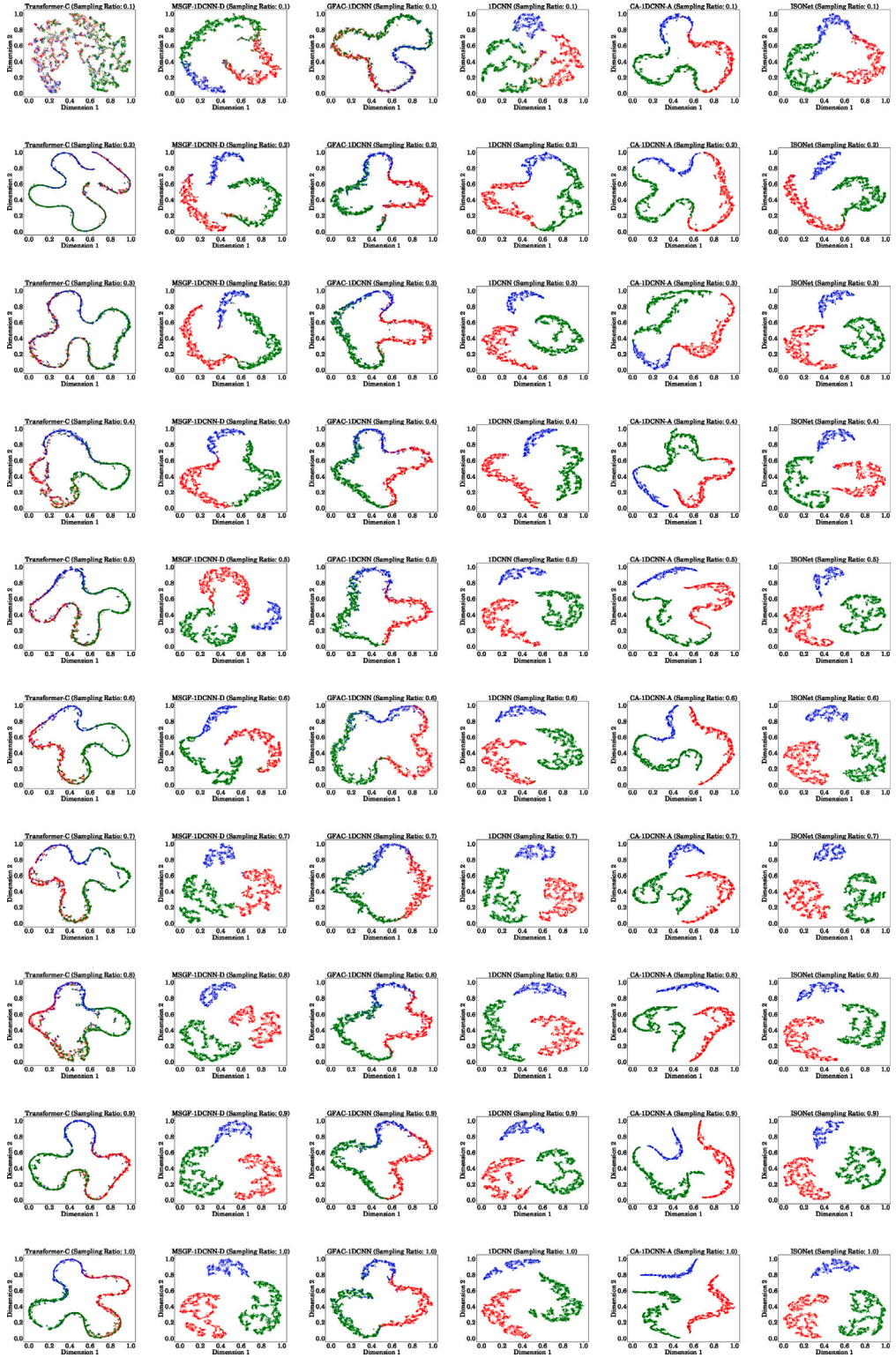


Fig. 17. Visualization of deep neural network model outputs via t-SNE (Maaten & Hinton, 2008) reduction to two dimensions with Cosine distance. Red: Health, Green: Inner Ring Fault and Blue: Outer Ring Fault.

methodologies presented in this paper, and compared the results with the archived outcomes as shown in Table 5. As can be discerned from Table 5, ISONet still maintains a significant advantage. This strongly demonstrates the effectiveness of ISONet presented in this paper for aero-engine system inter-shaft bearing fault diagnosis.

It is important to note that in the paper by Hou et al. (2023), the CNN accuracy was only 83.13%, while our ISONet achieved 99.39%. While the standard 1DCNN in our experiment also performed close to ISONet, the significant improvement in performance should not be solely attributed to the strategy of Input Spatial Over-parameterization. Several factors may contribute to this result:

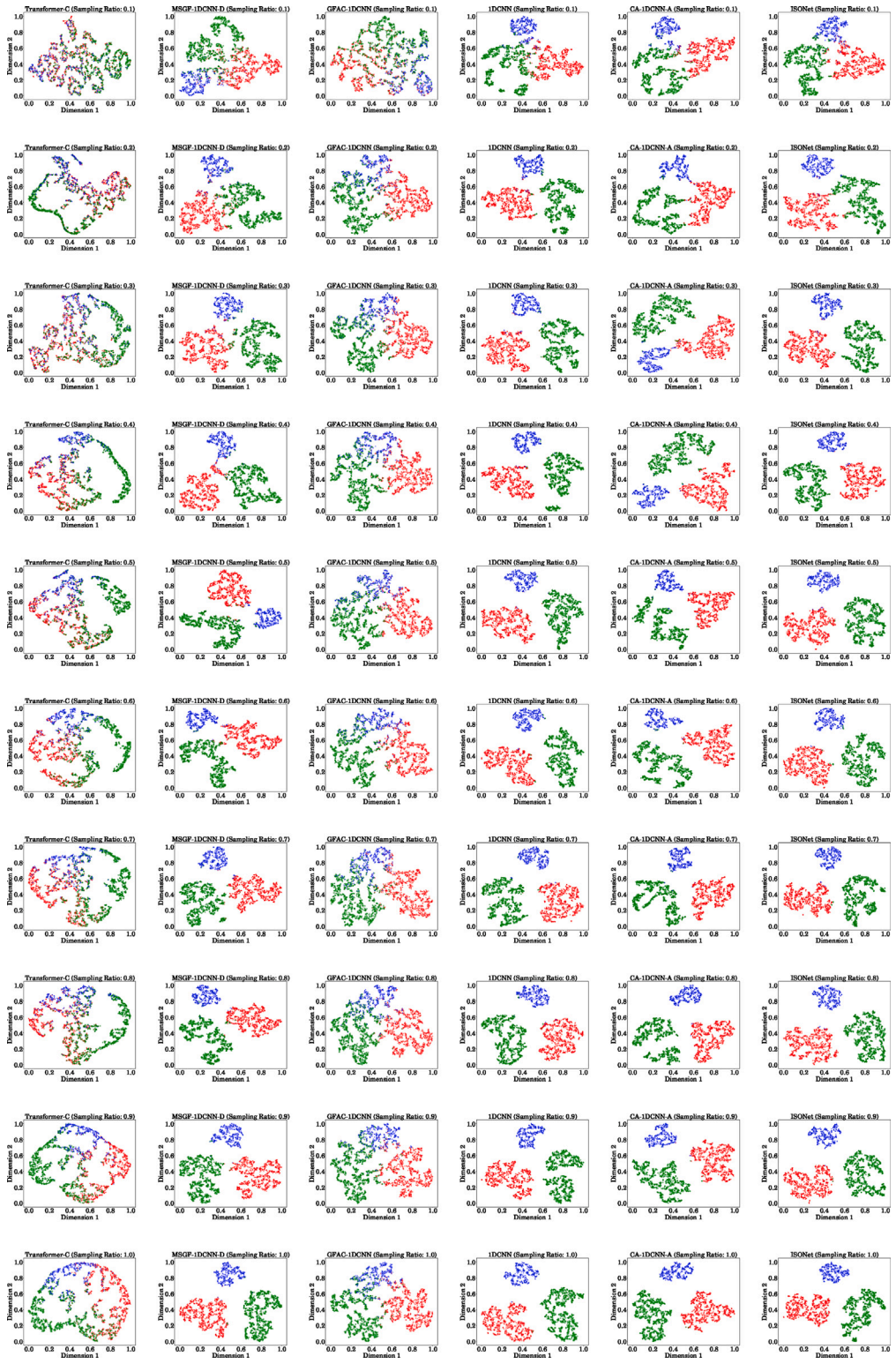


Fig. 18. Visualization of deep neural network model outputs via t-SNE (Maaten & Hinton, 2008) reduction to two dimensions with Euclidean distance. Red: Health, Green: Inner Ring Fault and Blue: Outer Ring Fault.

- Model Architecture and Hyperparameters: Our ISONet and the standard 1DCNN were carefully designed with optimized hyperparameters, including the number of layers, kernel sizes, and activation functions, which may differ from the CNN architecture used by Hou et al. (2023). These architectural choices can significantly impact the model's performance.
- Training Strategy and Evaluation Protocol: We employed the Adan optimizer with a specific learning rate and batch size, which may differ from the training strategy used by Hou et al. (2023). The optimization algorithm and training regimen can greatly influence the convergence and generalization of the model. While we ensured the consistency of the training data quantity with that of Hou et al. (2023), other aspects of the evaluation protocol, such

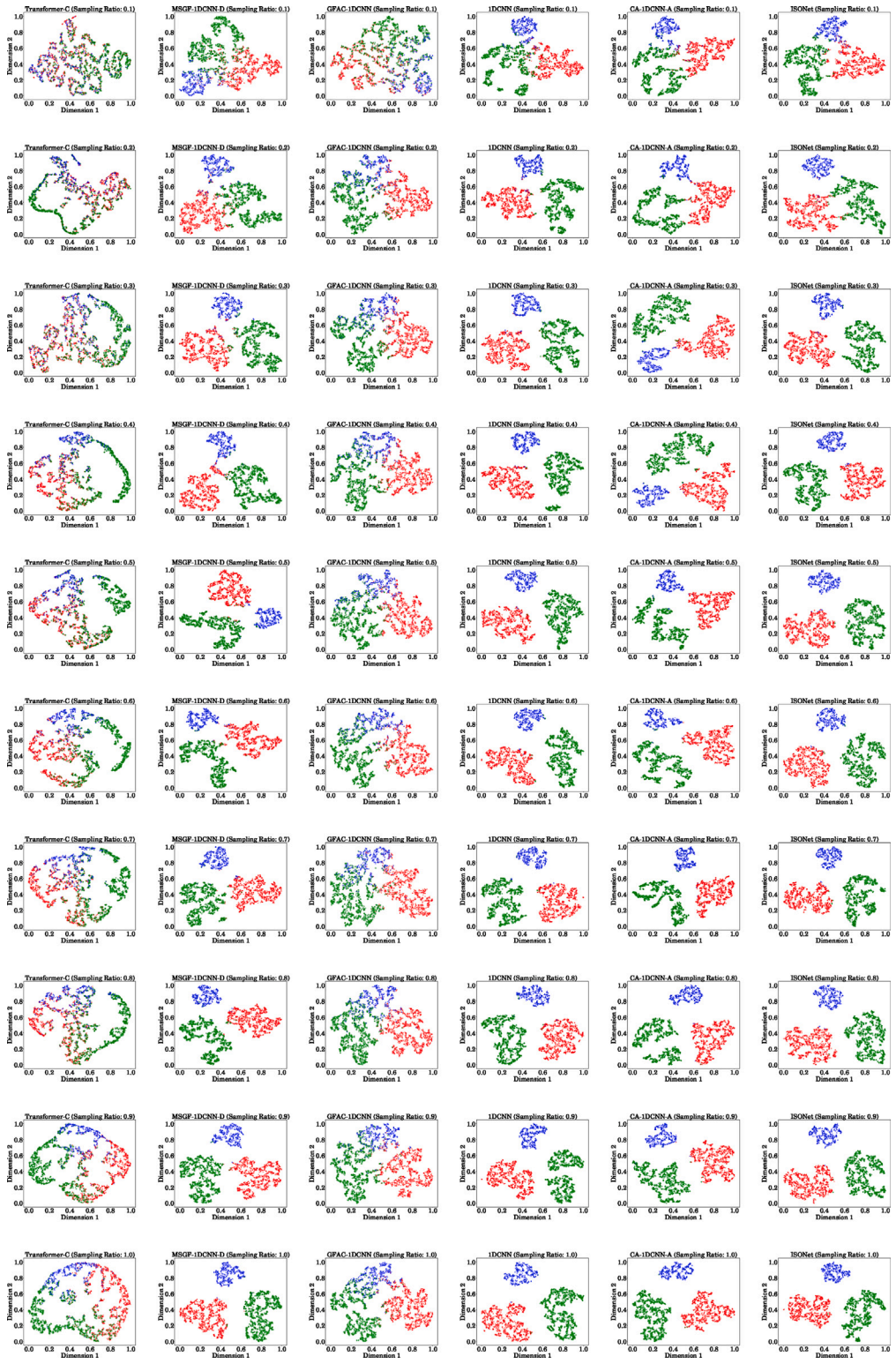


Fig. 19. Visualization of deep neural network model outputs via t-SNE (Maaten & Hinton, 2008) reduction to two dimensions with Manhattan distance. Red: Health, Green: Inner Ring Fault and Blue: Outer Ring Fault.

as the random seed for data splitting and the number of training epochs, may differ. These factors can introduce variability in the results.

- Data Preprocessing: While both studies used raw vibration signals, our preprocessing included batch normalization and adaptive pooling, which were not explicitly detailed in Hou et al. (2023). These steps could impact feature scaling and model stability.

While multiple factors may contribute to the superior performance of ISONet over the CNN in Hou et al. (2023), the comparative experiments in the preceding sections demonstrate that, under fair and identical experimental conditions, the contribution of Input Spatial Over-parameterization to performance improvement is notably significant and generalizable. This underscores the effectiveness of the

Table 5
Comparative results of ISONet against the recorded results under the 70% training and 30% testing split scenario.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
SLFN (Berghout et al., 2023)	34.38	34.38	34.38	34.38	–
RWI-LSTM (Berghout et al., 2023)	92.00	92.90	93.80	92.00	–
LSTM (Hou et al., 2023)	85.41	–	–	–	–
CNN (Hou et al., 2023)	83.13	–	–	–	–
TST (Hou et al., 2023)	71.07	–	–	–	–
ISONet (the proposed)	99.39	99.39	99.39	99.39	0.9998

proposed approach in enhancing diagnostic accuracy across various scenarios.

5. Conclusions and future works

This paper introduces a novel approach to enhance the performance of 1DCNNs for aero-engine inter-shaft bearing fault diagnosis. The proposed ISONet architecture, which incorporates input spatial over-parameterization, has demonstrated significant improvements in diagnostic accuracy, particularly under limited-sample conditions. This advancement is attributed to the innovative application of an 1D-ISOConv, which introduces additional learnable parameters to the model. The theoretical underpinnings of input spatial over-parameterization are elucidated through an analysis of matrix and vector operations, providing a solid foundation for the observed performance gains. The transformation of tensor operations into matrix/vector operations within the 1D-ISOConv layer allows for a more tractable theoretical framework, enabling the development of a specific preconditioning scheme that combines the benefits of momentum and adaptive learning rates. Empirical validation using real-world vibration data from aero-engine test rigs confirms the superiority of ISONet over existing deep learning models. The comparative analyses highlight the robustness of ISONet in fault diagnosis, even when faced with the challenges of data scarcity, thus addressing a critical need in the field of aero-engine maintenance and reliability. The findings of this research not only contribute to the theoretical understanding of neural network optimization but also offer a practical solution for enhancing the reliability and performance of aero-engine diagnostic systems. The transformative potential of ISONet marks a substantial stride forward in the domain of predictive maintenance, with implications for the broader field of engineering diagnostics.

However, at lower sampling ratio such as 0.1 and 0.2, ISONet, despite its robust intra-class compactness, exhibits a need for heightened inter-class discriminability. Therefore, to achieve inter-class discriminability, future research endeavors could concentrate on the exploration of adversarial examples as a means to fortify the model's robustness and its capacity to discern subtle distinctions within classes. Besides, exploring the potential of two-dimensional (2D) CNN for this task is a promising direction. However, it is essential to ensure the interpretability and scientific nature of the data transformation process when converting vibration data into a 2D format.

CRediT authorship contribution statement

Qian Xiang: Conceptualization, Data curation, Methodology, Software, Investigation, Resources, Validation, Visualization, Writing – original draft. **Xiaodan Wang:** Funding acquisition, Project administration, Supervision, Formal analysis. **Yafei Song:** Funding acquisition, Project administration, Writing – review and editing. **Lei Lei:** Formal analysis, Funding acquisition, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, grant numbers 62402521, 62203461, 62203365 and 62227814; the Young Talent Fund of University Association for Science and Technology in Shaanxi, China, grant number 20220106; the Young Talent Promotion Program of Shaanxi Association for Science and Technology, grant numbers 20220121, 20230125; and National Key R&D Program of China, grant number 2024YFB3311204.

Data availability

The data is publicly available via https://github.com/HouLeiHIT/HIT-dataset.

References

Arora, S., Cohen, N., & Hazan, E. (2018). On the optimization of deep networks: Implicit acceleration by overparameterization. In J. Dy, & A. Krause (Eds.), *Proceedings of machine learning research: Vol. 80, Proceedings of the 35th international conference on machine learning* (pp. 244–253). PMLR, URL: https://proceedings.mlr.press/v80/arora18a.html.

Berghout, T., Bentrcia, T., Lim, W. H., & Benbouzid, M. (2023). A neural network weights initialization approach for diagnosing real aircraft engine inter-shaft bearing faults. *Machines*, *11*(12), http://dx.doi.org/10.3390/machines11121089.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier, & G. Saporta (Eds.), *Proceedings of cOMPSTAT'2010* (pp. 177–186). Heidelberg: Physica-Verlag HD.

Che, C., Zhang, Y., Wang, H., & Xiong, M. (2024). Interpretable multi-domain meta-transfer learning for few-shot fault diagnosis of rolling bearing under variable working conditions. *Measurement Science and Technology*, *35*(7), Article 076103. http://dx.doi.org/10.1088/1361-6501/ad36d9.

Chen, D., Li, J., & Xu, K. (2020). AReLU: Attention-based rectified linear unit. http://dx.doi.org/10.48550/arXiv.2006.13858, ArXiv E-Prints, arXiv:2006.13858.

Dong, Z., Zhao, D., & Cui, L. (2024). Rotating machinery fault classification based on one-dimensional residual network with attention mechanism and bidirectional gated recurrent unit. *Measurement Science and Technology*, *35*(8), Article 086001. http://dx.doi.org/10.1088/1361-6501/ad41fb.

Fei, C. w., Han, Y. j., Wen, J. r., Li, C., Han, L., & Choy, Y. s. (2024). Deep learning-based modeling method for probabilistic LCF life prediction of turbine blisk. *Propulsion and Power Research*, *13*(1), 12–25. http://dx.doi.org/10.1016/j.jppr.2023.08.005.

Fu, L., Yan, K., Zhang, Y., Chen, R., Ma, Z., Xu, F., & Zhu, T. (2023). EdgeCog: A real-time bearing fault diagnosis system based on lightweight edge computing. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–11. http://dx.doi.org/10.1109/TIM.2023.3298403.

Gao, T., Yuan, S. m., Liu, Y. q., Cao, S. q., & Sun, J. q. (2024). Prediction and analysis of paroxysmal impulse vibration in aero-engine inter-shaft bearings induced by localized faults in the outer ring. *Nonlinear Dynamics*, http://dx.doi.org/10.1007/s11071-024-09729-y.

Guo, J., Yang, Y., Li, H., Dai, L., & Huang, B. (2024). A parallel deep neural network for intelligent fault diagnosis of drilling pumps. *Engineering Applications of Artificial Intelligence*, *133*(A), Article 108071. http://dx.doi.org/10.1016/j.engappai.2024.108071.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE international conference on computer vision* (pp. 1026–1034). http://dx.doi.org/10.1109/ICCV.2015.123.

Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). http://dx.doi.org/10.48550/arXiv.1606.08415, ArXiv E-Prints, arXiv:1606.08415.

Hou, Z. g., Wang, H. w., Lv, S. l., Xiong, M. l., & Peng, K. (2022b). Siamese multiscale residual feature fusion network for aero-engine bearing fault diagnosis under small-sample condition. *Measurement Science and Technology*, *34*(3), Article 035109. http://dx.doi.org/10.1088/1361-6501/aca044.

Hou, L., Yi, H., Jin, Y., Gui, M., Sui, L., Zhang, J., & Chen, Y. (2023). Inter-shaft bearing fault diagnosis based on aero-engine system: A benchmarking dataset study. *Journal of Dynamics, Monitoring and Diagnostics*, *2*(4), 228–242. http://dx.doi.org/10.37965/jdmd.2023.314.

Hou, L., Zhao, J., Dun, S., Cai, Y., Yang, Y., Xu, J., & Sun, C. (2022a). Feature extraction of weak-bearing faults based on Laplace wavelet and orthogonal matching pursuit. *Shock and Vibration*, *2022*(8154492), http://dx.doi.org/10.1155/2022/8154492.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. http://dx.doi.org/10.48550/arXiv.1704.04861, ArXiv E-Prints, arXiv:1704.04861.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach, & D. Blei (Eds.), *Proceedings of machine learning research*: *Vol. 37*, *Proceedings of the 32nd international conference on machine learning* (pp. 448–456). Lille, France: PMLR.

Kang, Y., Cao, S., Hou, Y., & Chen, N. (2022). Analysis of backward whirling characteristics of a dual-rotor system caused by unbalance. *Measurement*, *203*, Article 111982. http://dx.doi.org/10.1016/j.measurement.2022.111982.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, ICLR 2015*.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 972–981). Red Hook, NY, USA: Curran Associates Inc..

Liu, X., Chen, G., Cheng, Z., Wei, X., & Wang, H. (2022). Convolution neural network based particle filtering for remaining useful life prediction of rolling bearing. *Advances in Mechanical Engineering*, *14*(6), Article 16878132221100631. http://dx.doi.org/10.1177/16878132221100631.

Liu, X., Chen, G., Hao, T., & Pan, W. (2023). A combined deep learning model for damage size estimation of rolling bearing. *International Journal of Engine Research*, *24*(4), 1362–1373. http://dx.doi.org/10.1177/14680874221086601.

Liu, X., & Di, X. (2021). TanhExp: A smooth activation function with high convergence speed for lightweight neural networks. *IET Computer Vision*, *15*(2), 136–150. http://dx.doi.org/10.1049/cvi2.12020.

Luo, Y., Ren, X., Zheng, Z., Jiang, Z., Jiang, X., & You, Y. (2023). CAME: Confidence-guided adaptive memory efficient optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 4442–4453). Association for Computational Linguistics.

Luo, L., Xiong, Y., Liu, Y., & Sun, X. (2019). Adaptive gradient methods with dynamic bound of learning rate. In *Proceedings of the 7th international conference on learning representations, New Orleans, Louisiana*.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605, URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Misra, D. (2020). Mish: A self regularized non-monotonic activation function. In *The 31st British machine vision conference* (pp. 1–14).

Nag, S., Bhattacharyya, M., Mukherjee, A., & Kundu, R. (2023). Serf: Towards better training of deep neural networks using log-Softplus ERror activation Function. In *2023 IEEE/CVF winter conference on applications of computer vision* (pp. 5313–5322). http://dx.doi.org/10.1109/WACV56688.2023.00529.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on international conference on machine learning* (pp. 807–814). Madison, WI, USA: Omni Press.

Nemirovskii, A., & Nesterov, Y. (1985). Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, *25*(2), 21–30. http://dx.doi.org/10.1016/0041-5553(85)90100-4.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., Devito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ..., Bai, J. et al. (2019). PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd international conference on neural information processing systems* (pp. 1–12). Red Hook, NY, USA: Curran Associates Inc..

Su, W., Boyd, S., & Candès, E. J. (2014). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In *Proceedings of the 27th international conference on neural information processing systems - vol. 2* (pp. 2510–2518). Cambridge, MA, USA: MIT Press.

Sun, L., Zhu, X., Xiao, J., Cai, W., Ma, Q., & Zhang, R. (2024). A hybrid fault diagnosis method for rolling bearings based on GGRU-1DCNN with AdaBN algorithm under multiple load conditions. *Measurement Science and Technology*, *35*(7), Article 076201. http://dx.doi.org/10.1088/1361-6501/ad3669.

Tian, J., Liu, L., Zhang, F., Ai, Y., Wang, R., & Fei, C. (2020). Multi-domain entropy-random forest method for the fusion diagnosis of inter-shaft bearing faults with acoustic emission signals. *Entropy*, *22*(1), http://dx.doi.org/10.3390/e22010057.

Tian, J., Zhang, Y., Zhang, F., Ai, X., & Wang, Z. (2023). A novel intelligent method for inter-shaft bearing-fault diagnosis based on hierarchical permutation entropy and LLE-RF. *Journal of Vibration and Control*, *29*(23–24), 5357–5372. http://dx.doi.org/10.1177/10775463221134166.

Wang, J., Li, T., Sun, C., Yan, R., & Chen, X. (2022b). Improved spiking neural network for intershaft bearing fault diagnosis. *Journal of Manufacturing Systems*, *65*, 208–219. http://dx.doi.org/10.1016/j.jmsy.2022.09.003, URL: https://www.sciencedirect.com/science/article/pii/S0278612522001443.

Wang, C., Tian, J., Zhang, F. l., Ai, Y. t., Wang, Z., & Chen, R. z. (2024). Simulation method and spectrum modulation characteristic analysis of typical fault signals of inter-shaft bearing. *Mechanical Systems and Signal Processing*, *209*, Article 111145. http://dx.doi.org/10.1016/j.ymssp.2024.111145.

Wang, H., Xu, J., Sun, C., Yan, R., & Chen, X. (2022a). Intelligent fault diagnosis for planetary gearbox using time-frequency representation and deep reinforcement learning. *IEEE/ASME Transactions on Mechatronics*, *27*(2), 985–998. http://dx.doi.org/10.1109/TMECH.2021.3076775.

Wibisono, A., Wilson, A. C., & Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, *113*(47), E7351–E7358. http://dx.doi.org/10.1073/pnas.1614734113, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1614734113, URL: https://www.pnas.org/doi/abs/10.1073/pnas.1614734113.

Xiang, Q., Wang, X., Lai, J., Lei, L., Song, Y., He, J., & Li, R. (2024). Quadruplet depth-wise separable fusion convolution neural network for ballistic target recognition with limited samples. *Expert Systems with Applications*, *235*, Article 121182. http://dx.doi.org/10.1016/j.eswa.2023.121182.

Xiang, Q., Wang, X. D., Lai, J., Song, Y. F., Li, R., & Lei, L. (2022). Multi-scale group-fusion convolutional neural network for high-resolution range profile target recognition. *IET Radar Sonar and Navigation*, *16*(12), 1997–2016. http://dx.doi.org/10.1049/rsn2.12312.

Xiang, Q., Wang, X., Lai, J., Song, Y., Li, R., & Lei, L. (2023a). Group-fusion one-dimensional convolutional neural network for ballistic target high-resolution range profile recognition with layer-wise auxiliary classifiers. *International Journal of Computational Intelligence Systems*, *16*(1), 190. http://dx.doi.org/10.1007/s44196-023-00372-w.

Xiang, Q., Wang, X., Song, Y., Lei, L., Li, R., & Lai, J. (2021). One-dimensional convolutional neural networks for high-resolution range profile recognition via adaptively feature recalibrating and automatically channel pruning. *International Journal of Intelligent Systems*, *36*(1), 332–361. http://dx.doi.org/10.1002/int.22302.

Xiang, Q., Wang, X., Wu, X., Lai, J., He, J., & Song, Y. (2023b). CsiTransformer: A limited-sample 6G channel state information feedback model. In *2023 IEEE 6th international conference on pattern recognition and artificial intelligence* (pp. 1160–1166). http://dx.doi.org/10.1109/PRAI59366.2023.10331944.

Xiao, B., Zhao, Y., Zhou, C., Ou, J., & Huang, G. (2024). A noise-robust CNN architecture with global attention and gated convolutional kernels for bearing fault detection. *Measurement Science and Technology*, *35*(8), Article 086142. http://dx.doi.org/10.1088/1361-6501/ad4d16.

Xie, X., Zhou, P., Li, H., Lin, Z., & Yan, S. (2022). Adan: Adaptive Nesterov momentum algorithm for faster optimizing deep models. arXiv preprint arXiv:2208.06677.

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. http://dx.doi.org/10.48550/arXiv.1505.00853, ArXiv E-Prints, arXiv:1505.00853.

Yang, R., Zhang, Z., & Chen, Y. (2022). Analysis of vibration signals for a ball bearing-rotor system with raceway local defects and rotor eccentricity. *Mechanism and Machine Theory*, *169*, Article 104594. http://dx.doi.org/10.1016/j.mechmachtheory.2021.104594, URL: https://www.sciencedirect.com/science/article/pii/S0094114X21003359.

Yu, M., Fang, M., Chen, W., & Cong, H. (2023). Compound faults feature extraction of inter-shaft bearing based on vibration signal of whole aero-engine. *Journal of Vibration and Control*, *29*(1–2), 51–64. http://dx.doi.org/10.1177/10775463211041871.

Yu, M., Pan, X., Meng, G., & Chen, W. (2021). A study on the diagnosis of compound faults in rolling bearings based on ITD-SVD. *Journal of Vibroengineering*, *23*(3), 587–602. http://dx.doi.org/10.21595/jve.2020.21590.

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. ArXiv E-Prints, arXiv:1212.5701.

Zhang, W. t., Ji, X. f., Huang, J., & Lou, S. t. (2021). Compound fault diagnosis of aero-engine rolling element bearing based on CCA blind extraction. *IEEE Access*, *9*, 159873–159881. http://dx.doi.org/10.1109/ACCESS.2021.3130637.

Zhuang, J. T., Tang, T., Ding, Y. F., Tatikonda, S., Dvornek, N., Papademetris, X., & Duncan, J. S. (2020). AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients. In *Advances in neural information processing systems*: *Vol. 33*, *34th conference on neural information processing systems*. LA JOLLA: Neural Information Processing Systems (Nips).

Zuo, L., Zhang, L., Zhang, Z.-h., Luo, X. l., & Liu, Y. (2021). A spiking neural network-based approach to bearing fault diagnosis. *Journal of Manufacturing Systems*, *61*, 714–724. http://dx.doi.org/10.1016/j.jmsy.2020.07.003.