


## Article

# An Engine Fault Detection Method Based on the Deep Echo State Network and Improved Multi-Verse Optimizer

Xin Li <sup>1</sup> , Fengrong Bi <sup>1</sup>, Lipeng Zhang <sup>2,\*</sup>, Xiao Yang <sup>1</sup> and Guichang Zhang <sup>3</sup>

<sup>1</sup> State Key Laboratory of Engines, Tianjin University, Tianjin 300350, China; nvh\_lixin@tju.edu.cn (X.L.); fr\_bi@tju.edu.cn (F.B.); yangxiao@tju.edu.cn (X.Y.)

<sup>2</sup> Motorcycle Design Institute, Tianjin Internal Combustion Engine Research Institute, Tianjin 300072, China

<sup>3</sup> College of Aeronautical Engineering, Civil Aviation University of China, Tianjin 300300, China; gczhang@cauc.edu.cn

\* Correspondence: zhanglipeng@tju.edu.cn; Tel.: +86-22-2740-4944

**Abstract:** This paper aims to develop an efficient pattern recognition method for engine fault end-to-end detection based on the echo state network (ESN) and multi-verse optimizer (MVO). Bispectrum is employed to transform the one-dimensional time-dependent vibration signal into a two-dimensional matrix with more impact features. A sparse input weight-generating algorithm is designed for the ESN. Furthermore, a deep ESN model is built by fusing fixed convolution kernels and an autoencoder (AE). A novel traveling distance rate (TDR) and collapse mechanism are studied to optimize the local search of the MVO and speed it up. The improved MVO is employed to optimize the hyper-parameters of the deep ESN for the two-dimensional matrix recognition. The experiment result shows that the proposed method can obtain a recognition rate of 93.10% in complex engine faults. Compared with traditional deep belief networks (DBNs), convolutional neural networks (CNNs), the long short-term memory (LSTM) network, and the gated recurrent unit (GRU), this novel method displays superior performance and could benefit the fault end-to-end detection of rotating machinery.



**Citation:** Li, X.; Bi, F.; Zhang, L.; Yang, X.; Zhang, G. An Engine Fault Detection Method Based on the Deep Echo State Network and Improved Multi-Verse Optimizer. *Energies* **2022**, *15*, 1205. <https://doi.org/10.3390/en15031205>

Academic Editor: Athanasios I. Papadopoulos

Received: 22 December 2021

Accepted: 4 February 2022

Published: 7 February 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** echo state networks (ESNs); multi-verse optimizer (MVO); fault detection; deep learning; engine

## 1. Introduction

As regular power machinery, the diesel engine has superior output torque and fuel economy, which secures its irreplaceable role in the industry, agriculture, and so on. Under worsening energy and environmental crises, many countries are creating stringent legislation for diesel engines [1]. This presents a challenge for researchers; on the other hand, many technologies are developing rapidly as a result of this opportunity [2]. The application of new technologies improves the performance of engines, but it also leads to higher complexity, which results in more frequent failure [3].

Engine fault detection has developed from breakdown maintenance to regular maintenance and is gradually developing into predictive maintenance [4]. Traditional disassembly fault diagnosis technology is evolving to non-disassembly fault diagnosis. The non-disassembly predictive maintenance depends on collecting and analyzing state information [5]. The vibration signal is a common choice because of its rich information, high stability, and low cost. For example, Barszcz et al. [6] used vibration signals to detect the bearing and gear faults of an engine. Benkedjough et al. [7] designed a rotating machinery fault prediction and health management method by vibration signals. Furthermore, vibration analysis does not invade the engine block and can detect multiple kinds of faults, so it is currently considered to be one of the strongest potential methods [8]. The engine has many excitation sources; hence, sensors are usually placed on the block and cylinder head cover to collect vibration signals synthetically. However, the engine vibration signal usually has strong nonlinearity and randomness, and the background noise easily covers the weak

features caused by early faults. Therefore, appropriate algorithms should be researched to recognize faults accurately [9].

The most frequently used method is to begin by processing vibration signals by decomposition, denoising, feature extraction, and feature selection, and then recognizing engine faults by a simple classifier. For example, Wang et al. [10] decomposed vibration signals by variational mode decomposition (VMD) and selected the intrinsic mode function (IMF) with the highest kurtosis as the sensitive component to extract features; finally, the faults were detected by an extreme learning machine (ELM). The characteristic of this step-by-step method is the low requirement for the classifier. In contrast, the algorithm in every step should be controlled manually. Moreover, the recognition rate highly depends on extracted features. Still, the features usually have strong pertinence, leading to their needing to be adjusted according to specific applications with several attempts and having inferior generalization.

End-to-end fault detection is another attractive method because of its high efficiency and generalization capability. On the other hand, the high-performance classifier is required to analyze complex vibration signals, in which deep belief networks (DBNs) [11], convolutional neural networks (CNNs) [12], and recurrent neural networks (RNNs) [13] are frequently used artificial neural networks (ANNs). The DBN is stacked by several restricted Boltzmann machines (RBMs) and can detect the correlation of high-order data in visual layers by hidden layers. Ma et al. [14] used the DBN to detect signals compressed by compressive sensing to realize end-to-end fault recognition. Jiang et al. [15] proposed a DBN model optimized by the locality preserving projection (LPP), which could diagnose bearing faults without manual feature extraction. The CNN could gradually extract local features by convolutional layers, making it more suitable for high-dimensional data. Azamfar et al. [16] proposed a 2-dimensional CNN model based on signals fused from several sensors to detect gearbox faults end-to-end. The CNN could obtain better results in complex conditions with the pretreated vibration signals; for example, Hasan et al. [17] employed the CNN to analyze the vibration signals pretreated by the bispectrum for bearing fault detection in various working conditions. The bispectrum is a higher-order spectrum, the discrete Fourier transform of the higher-order cumulant. It has no definite physical significance but could magnify abnormal impact components for the pattern recognition model. As representative RNNs, the long short-term memory (LSTM) and gated recurrent unit (GRU) have advantages in time series analysis. Alrifayy et al. [18] combined the LSTM with a stacked autoencoder (SAE) to analyze time-dependent vibration signals for electrical gas generator fault detection. Yu et al. [19] used a stacked denoising autoencoder (SDAE) and the GRU to detect planetary gear faults. However, the gradient descent algorithm used in the DBN, the CNN, and the RNN may easily lead to low convergence speed and local minima. They are also prone to over-fitting with small-size training data. Moreover, the DBN and RNN cannot deal with spatial information because they generally require one-dimensional data. The CNN is usually combined with them, such as the DBN stacked by convolutional RBMs [20], convolutional LSTM [21], and multiscale CNN-GRU with attention mechanism (MCNN-AGRU) [22], whereas their efficiency and accuracy still have room for further improvement.

Jaeger et al. [23] proposed echo state networks (ESNs), whose basic principle is to take a randomly generated reservoir, instead of hidden layer neurons, as the basic processing unit to transform computing into linear regression. It shows excellent potential in pattern recognition. For example, Long et al. [24] used the ESN to analyze 3-dimensional printer faults, and Wootton et al. [25] designed a model by optimized ESN for static pattern recognition. However, the ESN lacks the capability for mining deep spatial information from time and frequency domains. Additionally, the hyper-parameters of the ESN are not clear in the mechanism and must be selected according to prior knowledge [26]. Zhang et al. [27] proposed a deep fuzzy ESN by combining fuzzy clustering, and Sun et al. [28] designed a deep belief ESN model based on the DBN to extract deep features. Unfortunately, they have shortcomings in dealing with high-dimensional samples. Although Ma et al. [29]

proposed a convolutional multi-timescale ESN inspired by the CNN, it still needs larger size training data. The particle swarm optimization (PSO) [30] and the binary grey wolf optimizer (GWO) [31] are popular evolutionary metaheuristic algorithms to search for the best result, but they are prone to the local optimum. Inspired by theories of the black hole, the white hole, and the wormhole, Mirjalili et al. [32] proposed a multi-verse optimizer (MVO). It has been widely used to optimize ANN hyper-parameters because of its better global optimization capability and stability. Faris et al. [33] employed the MVO to optimize the multi-layer perceptron (MLP) model and obtained the best result compared with several traditional algorithms. Yang et al. [34] used the MVO for the probability neural network (PNN) optimization to improve the recognition rate of power transformer faults. However, the MVO pours significant computing resources into the global search to avoid the local optimum, which results in insufficient local search and low convergence speed. It is worth improving further for the global and local searches in the MVO to optimize the ESN.

In this paper, a deep ESN model for engine faults end-to-end detection is proposed, and an improved MVO is researched to optimize hyper-parameters of the deep ESN. The main contributions are as follows:

- (1) A sparse input weight matrix is designed for the ESN. Optimized by fixed convolution kernels and the autoencoder (AE), a deep ESN is proposed.
- (2) A novel traveling distance rate (TDR) and universe collapse mechanism are proposed for the MVO to improve the local search and speed it up.
- (3) The bispectrum is employed to transform the one-dimensional time-dependent vibration signal into a two-dimensional matrix with more impact features. An engine fault end-to-end detection model is then built based on the deep ESN, the improved MVO, and the bispectrum.

This paper is organized as follows: the research background and significance are introduced in Section 1. Fundamental theories of the ESN and the MVO are introduced in Section 2. The deep ESN model and the improved MVO are proposed in Section 3. The diesel engine bench test and data collection system are described in Section 4. Section 5 introduces the analytical method of vibration signal and the complete framework of engine fault end-to-end detection model. The proposed method is also verified by experimental data in this section. The conclusion and outlook are presented in Section 6.

## 2. Fundamental Theories

The ESN and the MVO are the two main algorithms researched in this paper. They will be introduced in this section.

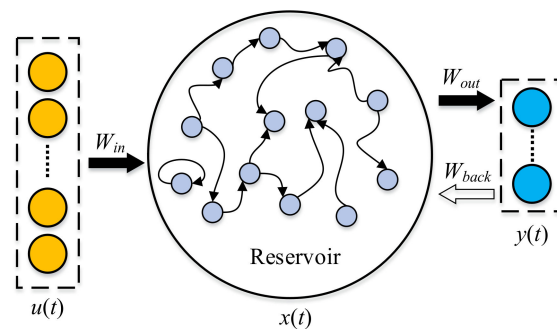
### 2.1. Echo State Networks

The significant characteristic of the ESN is that it takes a randomly generated reservoir as the basic processing unit. The reservoir can be activated into complex internal states, describing features of input signals by a simple linear combination [35]. Moreover, during training of the ESN, the reservoir and input weights are fixed, and only output weights are adjusted by linear regression, which could improve efficiency and avoid local minima, vanishing gradient, and exploding gradient [23]. The basic structure of the ESN is shown in Figure 1.

Suppose  $u = \{u_1, u_2, \dots, u_{n-1}, u_n\}$  is the input signal,  $x = \{x_1, x_2, \dots, x_{N-1}, x_N\}$  is the internal state of reservoir, and  $y = \{y_1, y_2, \dots, y_{m-1}, y_m\}$  is the output signal. The internal state updates with the time step  $t$  as:

$$x(t+1) = f(W_{in}u(t+1) + Wx(t) + W_{back}y(t)) \quad (1)$$

where  $W_{in}$ ,  $W$ , and  $W_{back}$  are randomly generated input, internal, and feedback weights, respectively, and  $f(\bullet)$  is the activation function.



**Figure 1.** Basic structure of the ESN.

The leaky integrate is taken as the neuron when the ESN is used for pattern recognition, so Equation (1) is transformed as:

$$x(t+1) = (1 - \alpha\gamma)x(t) + \gamma f(W_{in}u(t+1) + Wx(t) + W_{back}y(t)) \quad (2)$$

where  $\alpha$  is the leaky rate and  $\gamma$  is the gain. In application,  $\gamma = 1$  and  $W_{back} = 0$  in common [35]:

$$x(t+1) = (1 - \alpha)x(t) + f(W_{in}u(t+1) + Wx(t)) \quad (3)$$

The output of the ESN is:

$$y(t) = g(W_{out}[u(t); x(t)]) \quad (4)$$

where  $W_{out}$  represents the output weight and  $g(\bullet)$  is the activation function.

During the training of the ESN, only the  $W_{out}$  is updated. Its objective function  $L(\cdot)$  is:

$$L(\widehat{W}_{out}) = \left\| g^{-1}(y) - W_{out}[u; x] \right\|_2^2 \quad (5)$$

where  $\|\bullet\|_2$  represents  $L_2$  norm and  $g^{-1}(\bullet)$  is the inverse function of  $g(\bullet)$ .

The estimated output weight  $\widehat{W}_{out}$  is:

$$\widehat{W}_{out} = g^{-1}(y)[u; x]^\dagger = g^{-1}(y)([u; x]^T [u; x])^{-1} [u; x]^T \quad (6)$$

where the superscripts  $\dagger$  and  $T$  represent the pseudo-inverse and the transpose of the matrix, respectively.

## 2.2. Multi-Verse Optimizer

The MVO is an evolutionary metaheuristic algorithm inspired by multi-verse theory. In this theory, several universes are expanding in space with specific inflation rates. The wormhole is a hole that exchanges substances between different universes. The white hole sheds substances into space during universe collision. On the contrary, the black hole absorbs substances from space. Multiple universes could achieve balance through the wormhole and white/black holes [32]. The MVO takes universes as candidate solutions and inflation rate as fitness.

The search process of the MVO is divided into two phases: global search and local deep search. During iteration, the candidate solution is chosen from a better universe selected by the roulette wheel. It is also exchanged by white/black holes and wormholes to reach a solution around the global optimum. Suppose  $d$  represents the number of variables,  $c$  is the number of universes, and the candidate solution is  $U = [z_i^j]$ , where  $i \in [1, c]$ ,  $j \in [1, d]$ .

$z_i^j$  is the  $j$ th parameter in the  $i$ th universe selected by the roulette wheel selection mechanism:

$$z_i^j = \begin{cases} z_k^j & r1 < NI(U_i) \\ z_i^j & r1 \geq NI(U_i) \end{cases} \quad (7)$$

where  $NI(U_i)$  represents the normalized inflation rate of the  $i$ th universe  $U_i$  and  $r1$  is a random number between  $[0, 1]$ .

The lower the inflation rate, the higher the probability that a universe transfers substances by white/black holes. Moreover, the wormhole can transfer substances among different universes without considering the inflation rate to provide local variations for every universe. Supposing that the wormhole exists only between a universe and the current best universe, the mathematical model is:

$$z_i^j = \begin{cases} \begin{cases} Z_j + \text{TDR} \times ((ub_j - lb_j) \times r4 + lb_j) & r3 < 0.5 \\ Z_j - \text{TDR} \times ((ub_j - lb_j) \times r4 + lb_j) & r3 \geq 0.5 \end{cases} & r2 < \text{WEP} \\ z_i^j & r2 \geq \text{WEP} \end{cases} \quad (8)$$

where  $Z_j$  represents the  $j$ th parameter of the current best universe.  $r2$ ,  $r3$ , and  $r4$  are random numbers between  $[0, 1]$ .  $ub$  and  $lb$  represent upper and lower bounds of searched parameters, respectively. WEP and TDR represent wormhole existence probability and traveling distance rate, respectively.

The WEP increases linearly with the number of iterations:

$$\text{WEP} = \text{min} + l \times \left( \frac{\text{max} - \text{min}}{L} \right) \quad (9)$$

where  $l$  is the current iteration and  $L$  is the maximum number of iterations. The min is 0.2 and the max is 1 in common [32].

The TDR can describe the distance rate of substances transferred by the white hole around the current best universe:

$$\text{TDR} = 1 - \frac{l^{1/p}}{L^{1/p}} \quad (10)$$

where  $p$  is the control parameters of the global and local searches. Generally,  $p = 6$  [32].

### 3. Deep ESN and Improved MVO

#### 3.1. Deep ESN

##### 3.1.1. Fixed Convolution Kernel

The CNN is a frequently used model for mining deep information from high dimensional data, but it should be trained by large data. Advantages of the CNN include its convolutional and pooling layers. The convolutional layer is mainly used for gradually extracting local features, and the pooling layer is down sampling. However, complex convolutional layers require large training data. Considering that reservoir and input weights of the ESN are fixed during training, fixed convolution kernels are employed to extract features.

The engine fault will lead to an abnormal impact, characterized by energy concentration in a certain region. Therefore, the primary aim of the fixed convolution kernel is to detect the energy concentration region, and this could be realized by the edge detector. The Prewitt filter and the Sobel filter are classical edge detectors, which have simple structures and adjustable sizes and are especially suitable for analyzing the single-channel matrix. The Prewitt filter and the Sobel filter include horizontal and vertical operators, respectively (recorded as Ph, Pv, Sh, and Sv). The four types of filters of  $3 \times 3$  are shown

as follows:  $Ph = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$ ,  $Pv = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$ ,  $Sh = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$ , and  $Sv = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$ .

After edge detecting, characteristics should be integrated further. The Gaussian lowpass filter could reduce the Gaussian noise and improve smoothness to integrate slight features. Suppose  $G = [k_{ij}]_{d \times d}$  is a Gaussian lowpass filter:

$$k_{ij} = \exp\left(-\frac{(2i - d - 1)^2 + (2j - d - 1)^2}{8\sigma^2}\right) \tag{11}$$

where  $\sigma$  represents the standard deviation.

The Gaussian lowpass filter should be set behind the other two filters in fixed convolutional layers for accurate feature extraction.

### 3.1.2. Autoencoder (AE)

The fixed convolutional layers do not need large training data, but this results in low generalization capability. An autoencoder (AE) is introduced behind the fixed convolutional layers to provide generalization capability and further compress the features. The AE includes two phases: encoder and decoder, as shown in Figure 2.

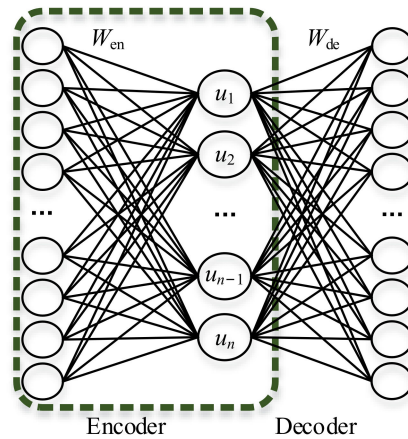


Figure 2. Basic structure of the AE.

Supposing the data analyzed by the fixed convolutional layer is  $C$ , the encoding phase is:

$$u = f(CW_{en}) \tag{12}$$

The decoding phase is:

$$\widehat{C} = g(uW_{de}) \tag{13}$$

The  $W_{en}$  and  $W_{de}$  are encoder and decoder matrices, respectively, which are trained by the gradient descent algorithm. The AE is incomplete when the dimension of the  $u$  is lower than  $C$ 's, and the  $W_{en}$  could be used for dimensionality reduction and feature extraction. In particular, the AE only has one hidden layer and is trained with one epoch in this study for a low computational burden.



### 3.1.3. Sparse Input Matrix of the ESN

As shown in Equation (3), the input matrix  $W_{in}$  and reservoir matrix  $W$  have significant influences on the internal state. When designing the  $W$ , Jaeger et al. [23] built a sparse square matrix of  $N \times N$  with the spectral radius:

$$W = \rho W_{original} / \lambda_{ei} \quad (14)$$

where  $N$  is the reservoir size,  $W_{original}$  is the randomly generated matrix,  $\lambda_{ei}$  is the maximum eigenvalue of  $W_{original}$ , and  $\rho$  is the spectral radius.

The  $W_{in}$  is a randomly generated dense matrix because it is a non-square matrix in common that has no eigenvalues. The size of the  $W_{in}$  relates to the dimension of the input data, which is much higher than the reservoir size generally. This may result in over-fitting and low generalization capability. A novel designing method of  $W_{in}$  similar to the  $W$  is proposed. Generating a sparse matrix:

$$W_{original}^{in} = [w_{ij}]_{n \times N} \quad (15)$$

where  $n$  is dimension of the input signal.

Next, computing the maximum singular value of  $W_{original}^{in}$ : suppose unitary matrices  $U$  and  $V$  and the diagonal matrix  $\Sigma = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$  could make  $W_{original}^{in} = U\Sigma V^T$ . The maximum singular value of  $W_{original}^{in}$  is  $\lambda_{\max} = \max(\lambda_1, \lambda_2, \dots, \lambda_r)$ , where  $r$  is the rank of  $W_{original}^{in}$ .

Finally, the input matrix is:

$$W_{in} = W_{original}^{in} / \lambda_{\max} = [w_{ij} / \lambda_{\max}]_{n \times N} \quad (16)$$

### 3.1.4. Deep ESN Model

As a type of RNN, the ESN needs signals with more than one time-step to activate internal states. After the AE, the processed signal is copied twice to obtain three time-steps, that is,  $u(1) = u(2) = u(3) = u$ . Equation (3) could be unfolded as:

$$\begin{cases} x(1) = f(W_{in}u) \\ x(2) = (1 - \alpha)x(1) + f(W_{in}u + Wx(1)) \\ x(3) = (1 - \alpha)x(2) + f(W_{in}u + Wx(2)) \end{cases} \quad (17)$$

Based on that, a deep ESN model is built as Figure 3, and the details are as follows:

- ① Design fixed convolutional layers based on the Prewitt filter, the Sobel filter, and the Gaussian lowpass filter.
- ② Process the input data by the designed convolutional and pooling layers.
- ③ Train the AE by processed data to obtain the encoder matrix  $W_{en}$ .
- ④ Compress features of the processed data further with Equation (12).
- ⑤ Build an ESN model with the sparse input matrix based on Equation (16).
- ⑥ Copy the data twice to obtain three time-steps to activate the internal state of the ESN based on Equation (17).
- ⑦ Train the ESN based on Equation (6).
- ⑧ Predict the output labels based on ②, ④, ⑥, and the trained model.

### 3.2. Optimization of MVO

The reservoir size  $N$ , spectral radius  $\rho$ , and leaky rate  $\alpha$  greatly influence the ESN, but there is no definitive method to select them [35]. Moreover, Verstraeten et al. [36] showed that a spectral radius higher than one might achieve the best result, contrary to [35]. The MVO is improved in this section to optimize the hyper-parameters.

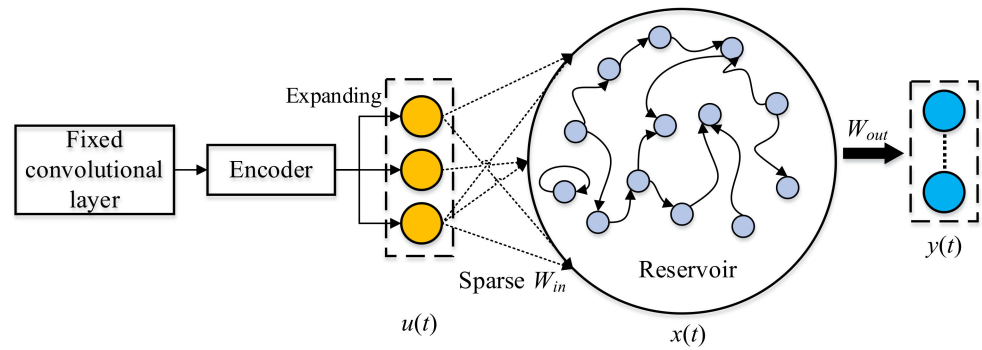


Figure 3. Model of deep ESN.

The TDR, as Equation (10), is an important parameter to control global and local searches. Its value changing with the current number of iterations  $l$  is shown as the black dotted line in Figure 4, where the maximum number of iterations is  $L = 500$ . A big TDR value is beneficial for the global search, whereas a small one is beneficial for the local search. Considering that the wormhole could also provide diversity for candidate solutions, the original TDR, which changes slowly from a big value, consumes too many computational resources in the global search, leading to disadvantages in the accuracy and efficiency of the local search. Therefore, a novel TDR is employed:

$$TDR = \omega_1 / e^{\omega_2(l/L)^2} \tag{18}$$

where  $\omega_1$  and  $\omega_2$  are exponential adjustment factors.

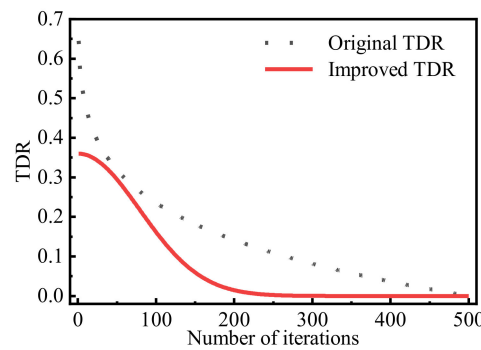


Figure 4. Curves of the TDRs.

Supposing  $\omega_1 = 0.36$  and  $\omega_2 = 20$ , the TDR value changing with the  $l$  is shown as the solid red line in Figure 4. The curve approaches 0 quickly, which means the MVO would centralize computational resources in the local search after a short global search period. Meanwhile, the WEP rises linearly, as Equation (9), to maintain the diversity of candidate solutions and avoid local optimum.

Inspired by the collapse of the universe in physics, an accelerated search mechanism is also proposed. In this mechanism, the number of universes  $c$  is reducing while the  $l$  is increasing:

$$c_l = \begin{cases} c_{\max} - l & c > c_{\max}/10 \\ c_{\max}/10 & c \leq c_{\max}/10 \end{cases} \tag{19}$$

where  $c_{\max}$  is the maximum number of universes.

For comparison, the improved MVO and original MVO are tested by several unimodal benchmark functions ( $f_1$ – $f_7$ ) and multi-modal benchmark functions ( $f_8$ – $f_{13}$ ) from [32], and the results are listed in Table 1. The maximum number of iterations and the maximum number of universes are both set at 500. The average value of 10 searches is taken as the



final result for reliability. The  $f_{\min}$  is the searched minimum value of the function. The smaller the  $f_{\min}$ , the better the algorithm.

**Table 1.** Comparison of improved MOV and original MVO.

Function	Dim	Bounds	$f_{\min}$	
			Original MVO	Improved MVO
$f_1 = \sum_{i=1}^n x_i^2$	10	[-100,100]	$2.78 \times 10^{-3}$	$9.66 \times 10^{-15}$
$f_2 = \sum_{i=1}^n  x_i  + \prod_{i=1}^n  x_i $	10	[-10,10]	$2.78 \times 10^{-2}$	$2.42 \times 10^{-8}$
$f_3 = \sum_{i=1}^n \left( \sum_{j=1}^i x_j \right)^2$	10	[-100,100]	$7.26 \times 10^{-3}$	$2.09 \times 10^{-14}$
$f_4 = \max( x_1 ,  x_2 , \dots,  x_n )$	10	[-100,100]	$2.88 \times 10^{-2}$	$5.48 \times 10^{-8}$
$f_5 = \sum_{i=1}^{n-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$	10	[-30,30]	$1.25 \times 10^2$	$1.06 \times 10^2$
$f_6 = \sum_{i=1}^n ( x_i + 0.5 )^2$	10	[-100,100]	$1.86 \times 10^{-3}$	$9.93 \times 10^{-15}$
$f_7 = \sum_{i=1}^n (ix_i^4) + \text{random}(0, 1)$	10	[-1.28,1.28]	$7.58 \times 10^{-4}$	$1.72 \times 10^{-3}$
$f_8 = \sum_{i=1}^n [-x_i \sin(\sqrt{ x_i })]$	10	[-500,500]	$-3.11 \times 10^3$	$-3.36 \times 10^3$
$f_9 = \sum_{i=1}^n [x_i^2 - \cos(2\pi x_i) + 10]$	10	[-5.12,5.12]	13.04	9.95
$f_{10} = -20 \exp\left(-0.2 \sqrt{\sum_{i=1}^n x_i^2/n}\right) - \exp\left(\sum_{i=1}^n \cos(2\pi x_i)/n\right) + 22.72$	10	[-32,32]	0.18	$3.92 \times 10^{-8}$
$f_{11} = \sum_{i=1}^n x_i^2/4000 - \prod_{i=1}^n \cos(x_i/\sqrt{i}) + 1$	10	[-600,600]	0.32	$9.31 \times 10^{-2}$
$f_{12} = \pi/n \{10 \sin(\pi y_i) + \sum_{i=1}^{n-1} (y_i - 1)^2 \times [1 + 10 \sin^2(\pi y_{i+1})] + (y_n - 1)^2\} + \sum_{i=1}^n u(x_i, 10, 100, 4)$	10	[-50,50]	$3.12 \times 10^{-2}$	$5.37 \times 10^{-5}$
$f_{13} = 0.1 \left\{ \sin^2(3\pi x_1) + \sum_{i=1}^{n-1} (x_i - 1)^2 \times [1 + (\sin^2(3\pi x_{i+1}))] + (x_n - 1)^2 \times [1 + \sin^2(2\pi x_n)] \right\} + \sum_{i=1}^n u(x_i, 5, 100, 4)$	10	[-50,50]	$2.50 \times 10^{-4}$	$1.27 \times 10^{-16}$

$$\text{where } y = (x + 5)/4 \text{ and } u(x, a, k, m) = \begin{cases} k(x - a)^m & x > a \\ 0 & -a < x \leq a \\ k(-x - a)^m & x \leq -a \end{cases} .$$

Except for  $f_7$ , results of the improved MVO are all smaller than the original one's, and the differences are by several orders of magnitude (the result of  $f_7$  may be caused by the stochastic term). Moreover, the number of total iterations in the improved MVO is  $L \times \sum_{i=1}^L c_i = 124805$ , and the number in the original MVO is  $L \times c_{\max} = 250000$ . The computational load of the improved MVO is almost half of the original one because of the acceleration search mechanism.

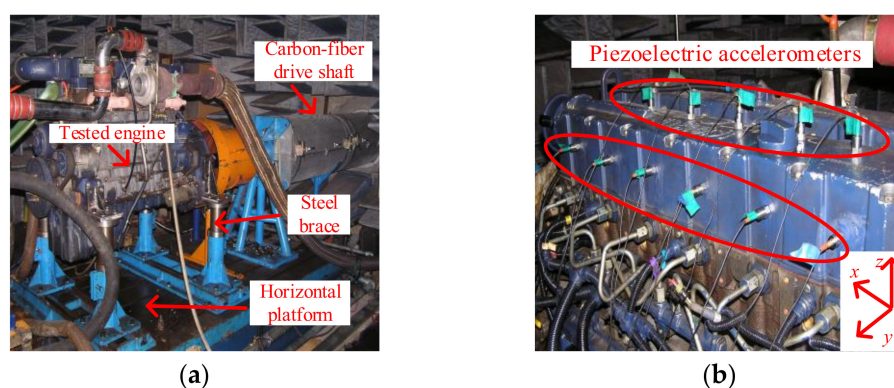
#### 4. Experiment

A turbocharged inline 6-cylinder diesel engine designed for heavy-duty vehicles is tested for a high reference value. The main technical parameters of the tested engine are listed in Table 2.

**Table 2.** Main technical parameters of testing diesel engine.

Items	Parameters
Number of Cylinders	6
Arrangement	Inline
Displacement	7.14 L
Air inlet model	Turbocharged and intercooled
Firing order	1-5-3-6-2-4
Rated power	220 kW@2300 r/min
Maximum torque	1250 N•m@1200–1600 r/min

The diesel engine bench test is shown in Figure 5a. The support system is composed of a horizontal platform and four air springs, whose natural frequency is below 2 Hz. The engine is connected rigidly with the platform by steel braces. An electrical dynamometer connects with the engine by a carbon-fiber driveshaft outside of the test bench room. Sensors are ICP 621B40 piezoelectric accelerometers produced by PCB Piezotronics, and the data acquisition system is SCM05 LMS Testlab produced by Siemens. An SPSR-115 photoelectric rotating-speed sensor produced by Monarch Electric Co. is placed near the connecting shaft of the dynamometer to synchronize engine speeds and vibration signals.

**Figure 5.** Diesel engine test bench: (a) Testing bench; (b) Sensor arrangement.

According to [3], fuel injection and valve systems are the most frequent failure parts (accounting for about 40%). The injection system fault is divided into three types: abnormal fuel injection timing is selected for simulating the control system failure, abnormal injection quantity for the injector failure, and abnormal rail pressure for the high-pressure common rail failure. The valve clearance faults are selected to simulate wear and carbon deposits in the valve system. In application, the faults should be detected at an early stage. Several early faults are designed, whose details are listed in Table 3, where “+” and “−” represent increasing and decreasing parameters from the normal condition, respectively, and CA is the crankshaft angle. Testing speeds include 700 r/min, 1300 r/min, 1600 r/min, 2000 r/min, and 2300 r/min, and testing loads include 100%, 75%, and 50%. Data in 15 s are collected under various working conditions, respectively.

Considering that cylinder pressure acts on block and cylinder head directly, and that, moreover, intake and exhaust valves are near cylinder head, six sensors are placed on the Y-direction (the horizontal direction perpendicular to the crankshaft) of the cylinder head cover near the 1st–6th cylinders. Five additional sensors are also placed on the Z-direction (vertical direction) as reference and comparison. The eleven sensor placements are shown in Figure 5b. The sampling frequency is set as 25.6 kHz based on the fault feature frequency of the testing engine and the Nyquist theorem.

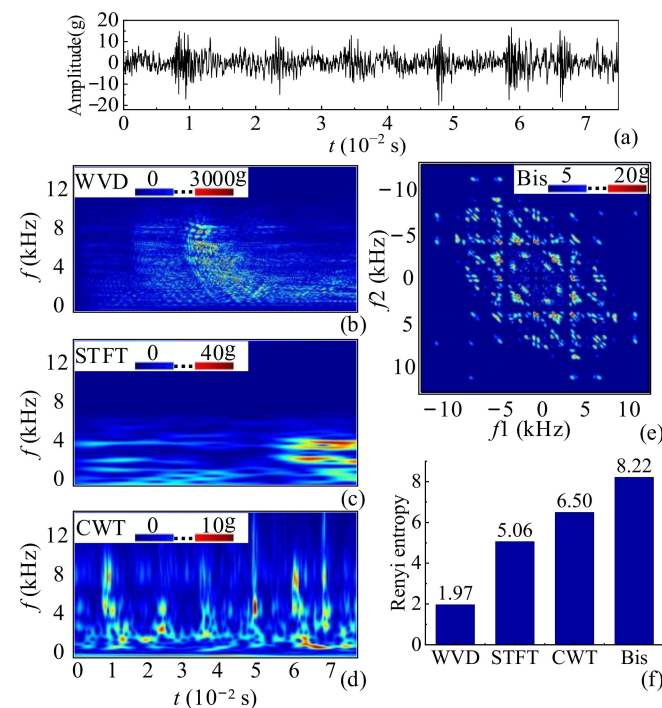
**Table 3.** Main technical parameters of testing diesel engine.

Fault Type	Adjusting Parameters		
Abnormal injection quantity	75%		
Advance injection timing	−2 °CA		
Delayed injection timing	+2 °CA		
Low rail pressure	−200 bar		
High rail pressure	+200 bar		
Small Valve clearance	Intake	−0.05 mm	
	Outtake	−0.05 mm	
Big Valve clearance	Intake	+0.05 mm	
	Outtake	+0.05 mm	

## 5. Results and Analysis

### 5.1. Data Pre-Processing Method

A signal during one engine cycle of normal working condition collected from the Y-direction near the third cylinder in 1600 r/min and 100% load is selected to analyze the testing engine vibration feature, as shown in Figure 6a, where  $g = 9.8 \text{ m/s}^2$ .



**Figure 6.** Diesel engine test bench: (a) Original vibration signal; (b) WVD result of the signal; (c) STFT result of the signal; (d) CWT result of the signal; (e) Bispectrum result of the signal; (f) Renyi entropies of the four results.

The original vibration data is the time-dependent signal, whereas the frequency information is essential for fault detection, especially the rotating machinery with periodic excitation sources. The signal should be analyzed simply to provide more frequency information for the neural network model. Four kinds of common methods are analyzed. The first one is Cohen's class distribution. The Wigner–Ville distribution (WVD) is a representative algorithm that describes energy distribution in the time-frequency domain [37]. The second method is the short-time Fourier transform (STFT), which carries on the Fourier transform in a sliding window of the time domain to obtain the time-frequency signal. The third method is continuous wavelet transform (CWT), which uses a wavelet basis function to obtain time-frequency information. The analyzing results of three classical time-frequency representation methods are shown in Figure 6b–d. The STFT adopts Hamming

window to reduce frequency leakage, and the window is set at 1/6 data length because the testing engine has six cylinders. The CWT employs the complex Morlet wavelet basis because its small window area is beneficial for the time-frequency domain analysis. The results of the WVD and the STFT are completely distorted, and the time domain information cannot reflect the cyclic impact components caused by the in-cylinder combustion. The CWT obtains a better result, but there are still distortion and illusion components in the low-frequency band. The WVD is bedeviled with quadratic cross-terms and fake harmonic trajectories. Although quadratic cross-terms have certain relationships to the faults, they still interfere with the analysis of the multicomponent vibration signal. The window function in the STFT cannot thoroughly avoid frequency leakage and restricts the performance in local and nonstationary signals. Due to Heisenberg's uncertainty principle, the CWT fails to maintain high resolutions both in low and high frequencies. The fourth method is the higher-order spectrum, in which the bispectrum (abbreviated as Bis in Figure 6) is frequently used for engine vibration analysis. The bispectrum can describe the non-Gaussian components with asymmetry and nonlinearity. In the engine vibration signal, the non-Gaussian components are usually the impact components caused by the excitation sources, such as the in-cylinder combustion and valve seating. The impact features in normal and abnormal working conditions are different, and the bispectrum is especially suitable for magnifying the vibration differences brought by engine faults [38]. Figure 6e shows the bispectrum analysis results of the vibration signal.

Renyi entropies are computed to compare them quantitatively [39], as shown in Figure 6f. The entropy can describe the information content, and the higher the entropy, the larger the information content. The bispectrum has the highest Renyi entropy of 8.22, which shows it is advantageous in information representation. Moreover, the two dimensions of the bispectrum result both describe frequency information, which makes the size of the matrix independent from the time domain. The duration of one engine cycle changes with its speed, so the result sizes of the other three methods also vary. Although the size could be unified by interpolation, it has a negative impact on the recognition, especially in a wide speed range.

Based on this, the complete framework of the fault detection model is shown in Figure 7. First, the bispectrum is employed to transform the one-dimensional time-dependent vibration signal into a two-dimensional matrix with more impact features. Secondly, the deep ESN model is built. Thirdly, the improved MVO is used to optimize the hyper-parameters of the deep ESN. Finally, the trained model can detect engine faults end-to-end.

## 5.2. Dataset

Detecting multiple engine faults in a single working condition requires advanced research. In this paper, multiple engine faults in several working conditions, including different speeds and loads, are analyzed to verify the advantages of the proposed method. The dataset built based on the experiment in Section 4 is listed in Table 4, including 6480 sets of samples at speeds of 1300 r/min, 1600 r/min, and 2000 r/min, and loads of 100% and 50%, respectively. Every sample contains the vibration signal during one engine cycle. The duration of one engine cycle is 120/speed seconds, and the sample is intercepted from the corresponding data without overlap.

## 5.3. Results and Comparisons

Two-thirds of the samples are selected randomly as the training data and the rest as the testing data. The one-dimensional signal is first processed by the bispectrum, and a simplified matrix of  $256 \times 256$  is obtained. Based on this, three convolutional layers are designed. The first layer contains three horizontal and vertical Prewitt filters of  $5 \times 5$ , respectively. The second layer contains three horizontal and vertical Sobel filters of  $7 \times 7$ , respectively. The third layer contains three Gaussian lowpass filters of  $3 \times 3$ . An average pooling layer of  $2 \times 2$  is added after every convolutional layer. An AE with one hidden layer of 1500 nodes is trained to process the 2523-dimensional data ( $29 \times 29 \times 3$ ). The

reservoir size  $N$ , spectral radius  $\rho$ , and leaky rate  $\alpha$  of the ESN are optimized by the improved MVO, where the search ranges are  $N \in \{2, 3, \dots, 15\}$ ,  $\rho \in (0, 1.5)$ , and  $\alpha \in (0, 1)$ , respectively. The original ESN, the DBN, the LSTM, the GRU, and the CNN are analyzed for comparisons. The results are listed in Table 5. Confusion matrices of these models are shown in Figure 8 to further analyze the results, where the labels of the working conditions in Table 4 are recorded as 0–7 from top to bottom.

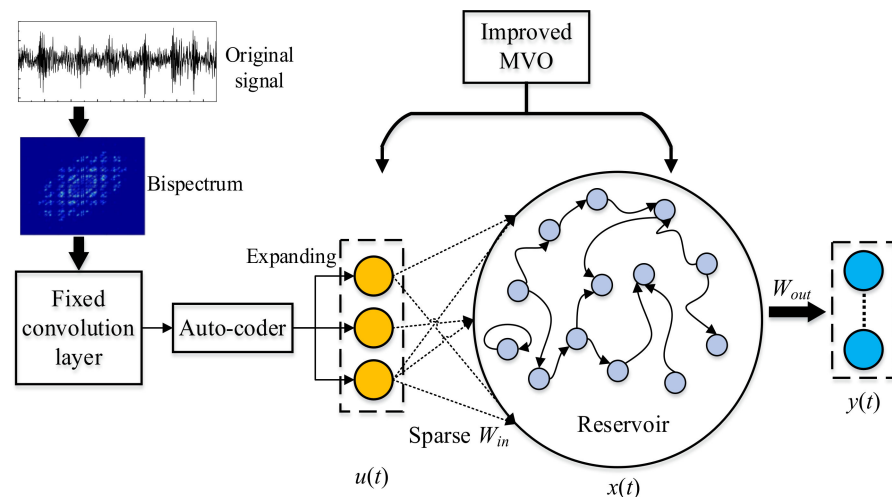


Figure 7. Framework of engine faults end-to-end detection model.

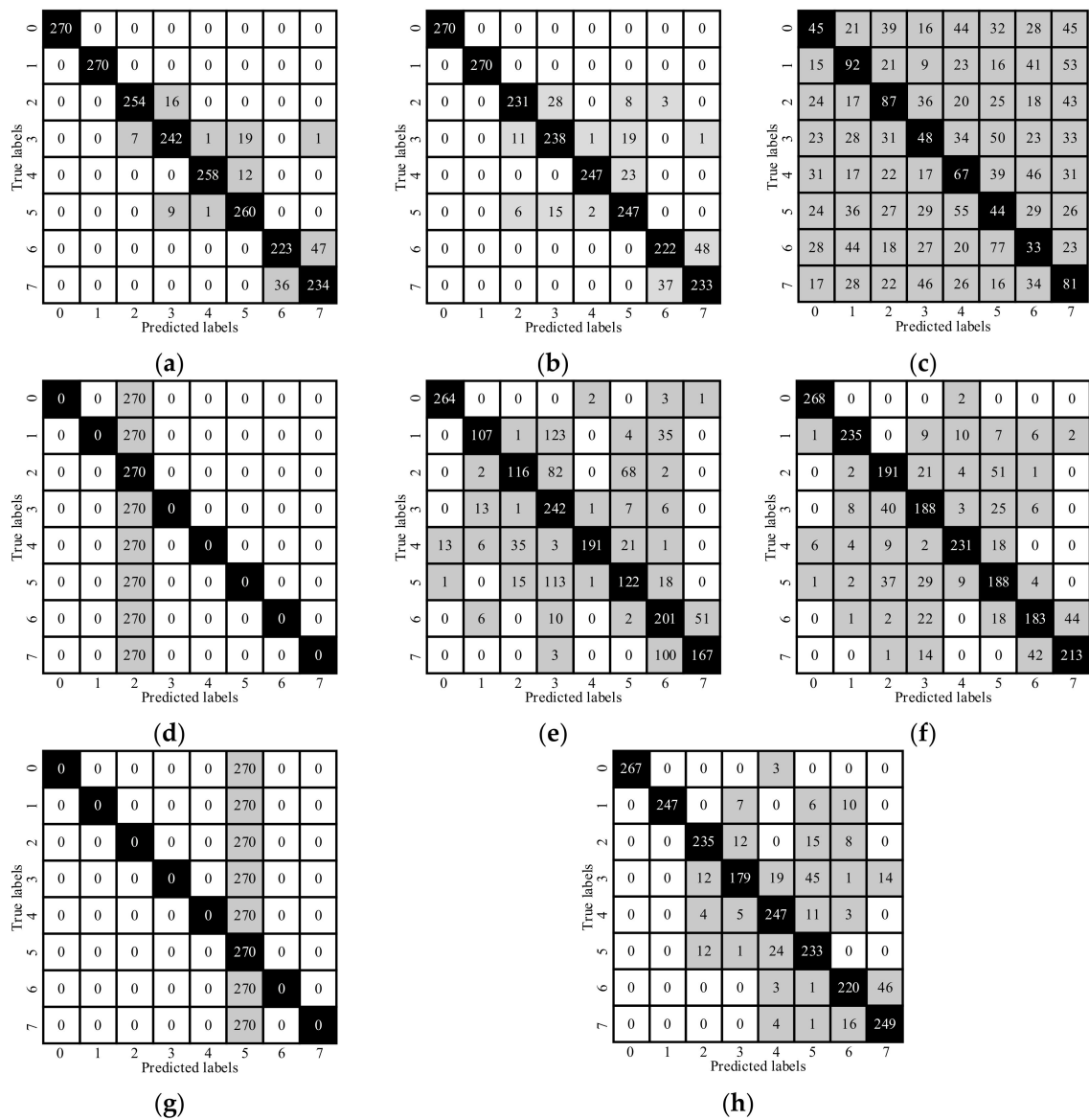
Table 4. Dataset of engine faults.

Fault Types	Number of Samples							Total
	Speed	1300 r/min		1600 r/min		2000 r/min		
		Load	100%	50%	100%	50%	100%	
Normal working condition		90	150	150	150	120	150	810
Abnormal fuel delivery		90	150	150	150	120	150	810
High rail pressure		90	150	150	150	120	150	810
Low rail pressure		90	150	150	150	120	150	810
Big valve clearance		90	150	150	150	120	150	810
Small valve clearance		90	150	150	150	120	150	810
Delayed injection timing		90	150	150	150	120	150	810
Advanced injection timing		90	150	150	150	120	150	810
Total		720	1200	1200	1200	960	1200	6480

Table 5. Recognition rates of models.

Model	Improved MVO						Deep ESN	
	Deep ESN	Original ESN	DBN	LSTM	GRU	CNN		CNN-BN
Recognition rate	93.10%	23.01%	12.50%	65.28%	78.56%	12.50%	86.90%	90.65%

Except for the ESNs, the max-epochs of the models are set at 500, and learning rates are set as  $lr = lr_0 \times 0.9^{ep/20}$  to obtain a better convergence, where  $lr_0$  is the initial learning rate and  $ep$  the current epoch. The initial learning rate and the batch size are optimized by the improved MVO. Additionally, the momentum of the DBN, and the reservoir size, spectral radius, and leaky rate of the original ESN are optimized.



**Figure 8.** Confusion matrices of different models. (a) Deep ESN-improved MVO, (b) Deep ESN, (c) Original ESN, (d) DBN, (e) LSTM, (f) GRU, (g) CNN, (h) CNN-BN.

The original one-dimensional signals and two-dimensional matrices are tested by the DBN and the original ESN, but unsatisfying results are obtained. Results of one-dimensional signals are listed in Table 5. The DBN contains three layers of 1000, 500, and 100 nodes, respectively. The results show that the DBN and the original ESN cannot mine deep information from limited one-dimensional signals. Primarily, the DBN identifies all samples as the high rail pressure and obtains a recognition rate of 12.50% (1/8), which means that the small-size training data could not provide enough information for the DBN to distinguish vibration signals in different conditions.

Considering the advantages of the RNN in time-dependent signals, the LSTM and the GRU are employed to analyze the original one-dimensional data. The LSTM and the GRU both have a three-layer structure with 32 nodes in the hidden layer. The recognition rates of 65.28% and 78.56%, respectively, show that the two classical RNN models could not detect engine faults accurately. Besides the performances of the classifiers, the deep information hidden in the time-dependent signal is another obstacle for fault detection, which proves the necessity of the two-dimensional matrix. The convolutional layer structure of the CNN is designed the same as the fixed one in the deep ESN for a clear comparison. The CNN



model can be considered a variant of the LeNet. The CNN identifies all samples as the small valve clearance and obtains a recognition rate of 12.50% as well. Dropout is employed to improve the DBN and the CNN; unfortunately, the results remain unchanged. Meanwhile, batch normalization (BN) [40] is used to optimize the CNN, and a high recognition rate of 86.90% is obtained. The deep ESN without the improved MVO is also analyzed, whose reservoir size, spectral radius, and leaky rate are set as 4, 0.2, and 0.2, respectively, based on [35]. It obtains a higher recognition rate of 90.65%, which shows the proposed deep ESN has advantages in mining deep information in complex data. The deep ESN optimized by the improved MVO obtains the highest recognition rate of 93.10%, in which the reservoir size, spectral radius, and leaky rate selected by the improved MVO are 7, 1.05, and 0.59, respectively. The ideal result shows that the improved MVO could select the best hyper-parameters for the deep ESN.

As shown in Figure 8, most of the errors in the deep ESN occur between the same fault in different degrees, such as the delayed and advanced injection timing conditions and high and low rail pressure conditions. The small valve clearance is the most deceptive condition. However, besides lower recognition rates, the other models all have a dangerous problem of recognizing fault conditions as normal ones. In particular, the CNN and traditional RNN models should be trained by GPU because they would spend several days in the CPU. Although the DBN could be trained by CPU, it is still seriously slower than the deep ESN. The hyper-parameter optimization aside, the DBN, LSTM, GRU, and CNNs are trained with 500 epochs, whose three layers of weights at least need to be adjusted based on the backpropagation (BP) in every epoch. As for the deep ESN, the ESN is trained only once by linear regression, and the AE has two weight matrices that are trained with one epoch. The results show that the proposed deep ESN optimized by the improved MVO is advantageous and practical.

Moreover, the deep ESN optimized by the improved MVO is tested by another nine cases, in which the training and testing datasets are re-selected randomly for cross-validation. Results are listed in Table 6, and Case 1 is drawn from Table 5. The recognition rates fluctuate around 93%, showing that the proposed method does not benefit from unique data and has stable performance. The results also verify that the experiment in this paper is reasonable and obtains reliable data.

**Table 6.** Recognition rates of the deep ESN-improved MVO under other datasets.

<b>Case</b>	1	2	3	4	5	6
<b>Recognition Rate</b>	93.10%	93.70%	92.92%	93.89%	93.24%	92.78%
<b>Case</b>	7	8	9	10	Average	
<b>Recognition Rate</b>	92.87%	93.06%	93.10%	93.38%	93.20%	

## 6. Conclusions and Outlook

An engine early fault end-to-end detection model is proposed based on a novel deep ESN and an improved MVO. In the deep ESN, a sparse input matrix is proposed based on the maximum singular value to improve generalization capability, and fixed convolution kernels and the AE are designed to mine deep spatial information. An improved MVO with a novel TDR and universe collapse mechanism is studied to search for the best hyper-parameters of the deep ESN. Compared with the original ESN, the DBN, the LSTM, the GRU, and the CNN, the proposed model obtains the highest recognition rate of 93.10% in multiple engine faults of different speeds and loads.

Much work needs to be done in the future. The mechanism of the reservoir should be analyzed further to improve the accuracy and efficiency of the ESN. Wormholes and white holes are potential fields in the MVO, and they will be researched in the following work.

**Author Contributions:** Conceptualization, X.L.; methodology, X.L.; software, X.L.; validation, X.L.; investigation, X.L., F.B. and X.Y.; resources, F.B., L.Z. and G.Z.; data curation, X.L. and X.Y.; writing—original draft preparation, X.L.; writing—review and editing, F.B., L.Z. and G.Z.; visualization, X.L.; supervision, F.B.; funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China Joint Funding Project, grant number U1833108.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Hirose, K.; Matsumura, T. A comparison between emission intensity and emission cap regulations. *Energy Policy* **2020**, *137*, 111115. [[CrossRef](#)]
2. Resitoglu, I.A.; Altinisik, K.; Keskin, A. The pollutant emissions from diesel-engine vehicles and exhaust aftertreatment systems. *Clean Technol. Environ. Policy* **2015**, *17*, 15–27. [[CrossRef](#)]
3. Nahim, H.M.; Younes, R.; Shraim, H.; Quladsine, M. Oriented review to potential simulator for faults modeling in diesel engine. *J. Mar. Sci. Technol.* **2016**, *21*, 533–551. [[CrossRef](#)]
4. Tahan, M.; Tsoutsanis, E.; Muhammad, M.; Karim, K.Z.A. Performance-based health monitoring, diagnostics and prognostics for condition-based maintenance of gas turbines: A review. *Appl. Energy* **2017**, *198*, 122–144. [[CrossRef](#)]
5. Vijay, G.S.; Pai, S.P.; Sriam, N.S.; Rao, R.B.K.N. Radial basis function neural network based comparison of dimensionality reduction techniques for effective bearing diagnostics. *Proc. Inst. Mech. Eng. Part J J. Eng. Tribol.* **2013**, *227*, 640–653.
6. Barszcz, T.; Jablonski, A. A novel method for the optimal band selection for vibration signal demodulation and comparison with the Kurtogram. *Mech. Syst. Signal Process* **2011**, *25*, 431–451. [[CrossRef](#)]
7. Benkedjouh, T.; Medjaher, K.; Zerhouni, N.; Rechak, S. Remaining useful life estimation based on nonlinear feature reduction and support vector regression. *Eng. Appl. Artif. Intell.* **2013**, *26*, 1751–1760. [[CrossRef](#)]
8. Jablonski, A.; Dworakowski, Z.; Dziedzic, K.; Chaari, F. Vibration-based diagnostics of epicyclic gearboxes-From classical to soft-computing methods. *Meas. J. Int. Meas. Confed.* **2019**, *147*. [[CrossRef](#)]
9. Zeng, R.L.; Zhang, L.L.; Mei, J.M.; Shen, H.; Zhao, H.M. Fault detection in an engine by fusing information from multivibration sensors. *Int. J. Distrib. Sens. Netw.* **2017**, *13*, 1–9. [[CrossRef](#)]
10. Wang, H.D.; Deng, S.E.; Yang, J.X.; Liao, H.; Li, W.B. Parameter-adaptive VMD method based on BAS optimization algorithm for incipient bearing fault diagnosis. *Math. Probl. Eng.* **2020**, *2020*, 5659618. [[CrossRef](#)]
11. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [[CrossRef](#)] [[PubMed](#)]
12. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, MA, USA, 1998.
13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comp.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
14. Ma, Y.F.; Jia, X.S.; Bai, H.J.; Wang, G.L.; Liu, G.Z.; Guo, C.M. A new fault diagnosis method using deep belief network and compressive sensing. *J. Vibroeng.* **2020**, *22*, 83–97.
15. Jiang, H.K.; Shao, H.D.; Chen, X.X.; Huang, J.Y. A feature fusion deep belief network method for intelligent fault diagnosis of rotating machinery. *J. Intell. Fuzzy Syst.* **2018**, *34*, 3513–3521. [[CrossRef](#)]
16. Azamfar, M.; Singh, J.; Bravo-Imaz, I.; Lee, J. Multisensor data fusion for gearbox fault diagnosis using 2-D convolutional neural network and motor current signature analysis. *Mech. Syst. Signal Process* **2020**, *144*, 106861. [[CrossRef](#)]
17. Hasan, M.J.; Sohaib, M.; Kim, J.M. A multitask-aided transfer learning-based diagnostic framework for bearings under inconsistent working conditions. *Sensors* **2020**, *20*, 7205. [[CrossRef](#)]
18. Alrifay, M.; Lim, W.H.; Ang, C.K. A novel deep learning framework based rnn-sae for fault detection of electrical gas generator. *IEEE Access* **2020**, *9*, 21433–21442. [[CrossRef](#)]
19. Yu, J.; Gao, L.L.; Yu, G.B.; Liu, K.; Guo, Z.Y. Fault identification of planetary gears based on the SDAE and GRUNN. *J. Vib. Shock* **2021**, *40*, 156–163.
20. Xie, J.Q.; You, W.; Shen, C.Q.; Zhu, Z.K. Bearing fault diagnosis based on improved convolution deep belief network. *J. Electron. Meas. Instrum.* **2020**, *34*, 36–43.
21. Liang, K.W.; Qin, N.; Huang, D.Q.; Fu, Y.Z. Convolutional recurrent neural network for fault diagnosis of high-speed train bogie. *Complexity* **2018**, *5*, 4501952. [[CrossRef](#)]

22. Zhang, X.C.; Cong, Y.W.; Yuan, Z.; Zhang, T.; Bai, X.T. Early fault detection method of rolling bearing based on MCNN and GRU network with an attention mechanism. *Shock Vib.* **2021**, *3*, 6660243. [[CrossRef](#)]
23. Jaeger, H.; Haas, H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science* **2004**, *304*, 78–80. [[CrossRef](#)] [[PubMed](#)]
24. Long, J.Y.; Sun, Z.Z.; Li, C.; Hong, Y.; Bai, Y.; Zhang, S.H. A novel sparse echo autoencoder network for data-driven fault diagnosis of delta 3-D printers. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 683–692. [[CrossRef](#)]
25. Wootton, A.J.; Taylor, S.L.; Day, C.R.; Haycock, P.W. Optimizing echo state networks for static pattern recognition. *Cognitive Comput.* **2017**, *9*, 391–399. [[CrossRef](#)]
26. Ozturk, M.C.; Xu, D.M.; Principe, J.C. Analysis and design of echo state networks. *Neural Comp.* **2007**, *19*, 111–138. [[CrossRef](#)] [[PubMed](#)]
27. Zhang, S.H.; Sun, Z.Z.; Wang, M.; Long, J.Y.; Bai, Y.; Li, C. Deep fuzzy echo state networks for machinery fault diagnosis. *IEEE Trans. Fuzzy Syst.* **2020**, *28*, 1205–1218. [[CrossRef](#)]
28. Sun, X.C.; Li, T.; Li, Q.; Huang, Y.; Li, Y.Q. Deep belief echo-state network and its application to time series prediction. *Knowl. Based Syst.* **2017**, *130*, 17–29. [[CrossRef](#)]
29. Ma, Q.L.; Chen, E.H.; Lin, Z.X.; Yan, J.Y.; Yu, Z.W.; Wing, W.Y.N. Convolutional multitime-scale echo state network. *IEEE Trans. Cybern.* **2021**, *51*, 1613–1625. [[CrossRef](#)] [[PubMed](#)]
30. Chouikhi, N.; Ammar, B.; Rokbani, N.; Alimi, A.M. PSO-based analysis of echo state network parameters for time series forecasting. *Appl. Soft Comput. J.* **2017**, *55*, 211–225. [[CrossRef](#)]
31. Liu, J.X.; Sun, T.N.; Luo, Y.L.; Yang, S.; Cao, Y.; Zhai, J. Echo state network optimization using binary grey wolf algorithm. *Neurocomputing* **2020**, *385*, 310–318. [[CrossRef](#)]
32. Mirjalili, S.; Mirjalili, S.M.; Hatamlou, A. Multi-verse optimizer: A nature-inspired algorithm for global optimization. *Neural Comput. Appl.* **2016**, *27*, 495–513. [[CrossRef](#)]
33. Faris, H.; Aljarah, I.; Mirjalili, S. Training feedforward neural networks using multi-verse optimizer for binary classification problems. *Appl. Intell.* **2016**, *45*, 322–332. [[CrossRef](#)]
34. Yang, X.H.; Chen, W.K.; Li, A.Y.; Yang, C.S. A hybrid machine-learning method for oil-immersed power transformer fault diagnosis. *IEEE Trans. Electr. Electron. Eng.* **2020**, *15*, 501–507. [[CrossRef](#)]
35. Jaeger, H.; Lukosevicius, M.; Popovici, D.; Siewert, U. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Netw.* **2007**, *20*, 335–352. [[CrossRef](#)]
36. Verstraeten, D.; Schrauwen, B.; D’Haene, M.; Stroobandt, D. An experimental unification of reservoir computing methods. *Neural Netw.* **2007**, *20*, 391–403. [[CrossRef](#)]
37. Cohen, L. Generalized phase-space distribution functions. *J. Math. Phys.* **1966**, *7*, 781–786. [[CrossRef](#)]
38. Bi, X.Y.; Cao, S.Q.; Zhang, D.M. Diesel engine valve clearance fault diagnosis based on improved variational mode decomposition and bispectrum. *Energies* **2019**, *12*, 661. [[CrossRef](#)]
39. Baraniuk, R.G.; Flandrin, P.; Janssen, A.J.E.; Michel, J.J. Measuring time-frequency information content using the Renyi entropies. *IEEE Trans. Inf. Theory* **2001**, *47*, 1391–1409. [[CrossRef](#)]
40. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 37, pp. 449–456.