RV reducer Design Using Resnet-based model and integration of Discretized OPtimization

Jiacheng Miao (✉ haomjc@163.com)
Chongqing University https://orcid.org/0000-0001-8868-0095

Chaoyang Li
Chongqing University

Bingkui Chen
Chongqing University

Research Article

Keywords: Sequential Engineering, RV reducer, Multidisciplinary optimization, MP-NSGA-II algorithm, Secondary development

Posted Date: July 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-727232/v1

RV reducer Design Using Resnet-based model and integration of Discretized OPtimization

Jiacheng Miao[1], Chaoyang Li[1], Bingkui Chen[1]

ABSTRACT

A new type of mechanical system structure design model is proposed, which uses a small number of system feature samples to generate a new structure model. In this model, (1) the theory of limited sample recommendation algorithm is used to study the external dimensions recommendation of the reducer, an SG-Resnet network suitable for the generation of reducer structure parameters is established, the main factors affecting the promotion ability and learning rate of the SG-Resnet network structure are analyzed through hyperparameters, and in-depth study of the mechanism of each influencing factor. (2) Establish an optimization design method for the internal dimensions of the reducer, and initially calculate the structural parameters according to the basic performance parameters of the reducer, combine the objective function and constraint conditions to establish the corresponding multi-objective optimization model, and establish the Kriging proxy model. The mixed population NSGA-II algorithm is proposed, the MP-NSGA-II algorithm is used to obtain multiple sets of Pareto optimal solutions, and the multi-objective evaluation method is used to select the optimal solution from the non-dominated solution set.

Experiments were carried out to verify the positive enhancement effect of the structural design model on the stiffness of the reducer. The experiment showed the reliability and generalizability of the model. This research provides a new solution for reducer design and lays a solid foundation for the development of integrated RV reducer forward design software.

Key words: Sequential Engineering; RV reducer; Multidisciplinary optimization; MP-NSGA-II algorithm; Secondary development

1 Introduction

1) **The value of this research.** Under normal circumstances, when designing the structure of the reducer, the dynamic characteristics of the mechanism are not considered too much. The key components are usually fine-tuned based on existing experience, and the effect is not ideal. This design method requires increasing transmission performance of the reducer not effectively. Therefore, in order to achieve good transmission characteristics in the structural design stage, the design process is divided into external parameter recommendation and internal parameter optimization.

2) **The key to system implementation.** It is necessary to introduce a parameter recommendation algorithm in the process of external structure design, and uses a computer to conduct preliminary design of the reducer structure [1]. Regarding the complex mechanical system

[1] State Key Laboratory of Mechanical Transmissions, Chongqing University, Chongqing 400030, PR China

Corresponding author:

Jiacheng Miao, State Key Laboratory of Mechanical Transmissions, Chongqing University, Chongqing 400030, PR China.

Email: haomjc@163.com

scheme design, the parameters obtained from the prototype series are employed to obtain the appropriate structure for the final system. These parameters are provided as background data.

Optimization of internal structure is based on the background data, and its essence is a kind of multi-objective mixed integer nonlinear programming (MOMINLP) problem. Generally, precise algorithms such as cutting plane method and branch and bounding method are used to solve problems with small variable spaces. However, when solving problems with high-dimensional variable spaces, precise algorithms are used. The time complexity is extremely high. As an optimization method with good convergence, high parallelism, and strong robustness, evolutionary algorithm is suitable for solving the multi-objective optimization problem of reducer structure parameter design with many solving variables and large parameter space.

3) **Traditional method/existing method.** The essence of the design of the shape and structure parameters of the reducer is to establish the mapping relationship between performance parameters and structural dimensions. Researchers generally carry out structural design based on experience and existing foreign structures. For situations where performance requirements are different from existing products, more scientific research is needed. Effective design methods, such as Hopfield neural network proposed by Hopfield. The bionic design of biological knowledge modeling provides ideas for the forward design method. SHU et al. [2] bridged biological systems and engineering systems through cross-domain terminology, searching for biological knowledge data and related biological phenomena. NAGEL et al.[3]used functional modeling and biological databases to provide solutions to engineering design problems. Some scholars have carried out in-depth research in the field of small sample learning, and their research work mainly focuses on lightweight network structure design, sample expansion and training strategy adjustment. Since the existing data of the structure parameters and transmission performance of the reducer belongs to the small sample type, the modeling reliability of the small sample mapping model should be considered. Shimizu S et al.[4] For fatigue reliability evaluation, in order to achieve accurate life evaluation of small sample data, artificial neural networks are used to analyze the fatigue life dispersion and obtain a high-precision P-S-N curve. In order to improve the performance and quality of unbalanced data analysis in small sample problems, literature[5] proposed a method of data generation, which provides a data basis for classification. The key to establishing a mapping relationship is to achieve effective data association. Traditional data association methods include joint probabilistic data association (JPDA) algorithm[6] [7] and K nearest neighbor (KNN) algorithm[8], both of which are stable in data error distribution. Based on this algorithm, the method of hypothesis testing is used to distinguish the data association relationship. In order to improve the performance and quality of unbalanced data analysis, literature[9] proposed a method of data generation. The classification algorithm provides the basis for the establishment of the mapping model. Hasnat et al.[10] proposed a fabric defect classification system (PNN) using a probabilistic neural network, and implemented the system using field programmable gate array (FPGA) hardware. Zhang et al.[11] proposed a new method of fabric defect classification based on Gaussian Mixture Model (GMM) improved radial basis function (RBF) network, and verified it on 9 types of defect images;

4) **New method.** The performance of the deep learning model is closely related to the quality and scale of the original data. The deep learning model supported by large-scale samples has excellent performance and good generalization ability, but in many situations where it is difficult to obtain a large amount of labeled data, the prediction performance of existing deep learning models is usually unsatisfactory. Some scholars have conducted in-depth research in the field of small

sample learning. The research work mainly focuses on lightweight network structure design, sample expansion and training strategy adjustment.

The existing deep learning network has a complicated structure, and the effect on tasks with large data volume is remarkable. GoogLeNet, VGGNet, AlexNet and other network structures perform well in the ILSVRC image classification competition, but under small sample conditions, such complex network structures are prone to overfitting. Therefore, in a small sample learning task, a lightweight network structure that matches the number of parameters and the sample size can be designed to appropriately reduce the width and depth of the network.

Sufficient training samples can avoid overfitting. When there are few training samples, sample perturbation or sample synthesis can be used to increase the amount of data. Sample perturbation mainly includes geometric transformation, principal component perturbation, noise perturbation, etc. Sample synthesis can use prior knowledge to generate data, add irrelevant details to existing data, or use the generative confrontation network (GAN) generation model and the discriminant model to play against each other to generate new training samples.

Under the premise that the network structure and training samples are fixed, the training strategy planning has a direct impact on the network performance. Under the condition of big data, the method of randomly initializing parameters can be used, while under the condition of small samples, the training strategy of fine-tuning the pre-trained model can be used. Initializing the network with the weight information trained on a large-scale data set can greatly improve the network prediction performance and accelerate network convergence.

The overall deep learning network model is mainly in the fields of image recognition, data classification and other information processing and pattern recognition, and the design research on the end-to-end physical structure generation model is still in the exploratory stage. Through the continuous development of intelligent design technology, this paper proposes a detailed design model of the serialized product structure generation model.

Considering the mutual restriction of mechanical system performance, evolutionary algorithm can be used to solve the design problem of internal structural parameters of reducer. Swarm intelligence algorithm researchers deduced the evolutionary algorithm[12], and part of the research is based on multi-objective particle swarm optimization (MOPSO) or real number coding Differential Evolution Algorithm (DE), using trigonometric functions, Sigmod functions, etc. to establish the mapping relationship between real numbers and integers [13][14] . The difficulty in structural optimization is how to deal with FEM models that consume a lot of computing resources, and the use of proxy models such as response surface method, support vector machine, Kriging model or artificial neural network can reduce the computational cost of structural optimization. Some scholars use the combination of optimization theory and software design technology to carry out computer-aided design. In order to find the optimal solution for multi-objective optimization problems, Pareto theory is generally used to solve the Pareto non-dominated solution set. Corresponding research has been conducted on the optimization design of the linear reducer. The existing NSGA-II algorithm has a single coding population type, and there are few solutions to the optimal design of the reducer such as the mixed variable optimization problem containing continuous and discrete variables. Therefore, on the basis of studying the traditional optimization method of reducer, improving the coding population and constructing an algorithm suitable for solving multi-objective mixed integer nonlinear programming problems have certain practical value.

5) **Research goals/organizational structure of the paper.** The research in this article is based

on the current domestic reducer as a reference to the positive development of the new model. To achieve serial design, one method is the design of a new structure, which in the structural design optimization process of the existing complex model information utilization is insufficient, and a large number of analysis and test tasks brought by a large number of structural design schemes are difficult to achieve; another way is to improve the design of the reducer, this design method includes the geometric parameter information of a large number of old model models reorganization. In order to make the design more efficient, a simplified parametric model of the reducer, a prototype model that can be repeatedly called in the component library, and a general optimization method for the reducer components are required.

At present, the structure design of reducer mainly adopts two methods of experiment and numerical simulation. There are few reports about the use of parameter recommendation methods for product development. Compared with manual adjustment of parameters, parameter recommendation algorithms can reduce trial and error costs, but they must be available. Structural parameters, training samples and algorithm learning ability are the key point.

The performance of a deep learning model is closely related to the quality and scale of the original data. The deep learning model supported by large-scale samples has excellent performance and good generalization ability. However, in many cases where it is difficult to obtain a large amount of labeled data, the prediction performance of the existing deep learning model is usually not satisfactory. In order to meet engineering requirements and reduce the design workload, research on the establishment of small-sample recommendation algorithms, optimization and implementation effect evaluation is necessary.

Torsional stiffness is one of the key performance parameters of RV reducer. At present, researchers have carried out more in-depth research on the stiffness characteristics of cycloid reducers[15][16], but there are few discussions on the stiffness optimization schemes of RV reducers. The optimization design goals of related documents are mainly based on volume and efficiency. The transmission efficiency, volume and torsional stiffness of the integrated RV reducer restrict each other, and it is necessary to balance each performance index to obtain a more balanced and reasonable optimization result. The current stiffness analysis of reducer usually adopts the method of prototype simulation or numerical analysis. The former has a huge amount of calculation and is difficult to meet the requirements of the optimization algorithm, and the latter is difficult to accurately reflect the nonlinear relationship between the structure parameters and torsional stiffness of the reducer. In order to reduce the time-consuming simulation, it is necessary to establish the Kriging proxy model of the torsional stiffness of the integrated RV reducer[17].

Since the standard NSGA-II cannot handle optimization problems with mixed integer constraints, a modified NSGA-II that can simultaneously process real number populations, integer populations and discrete populations is developed. Compared with the function mapping method, there is no need to design and edit variables separately. The decoding scheme improves the versatility and computational efficiency of the algorithm.

Through the improvement of the coding population, an algorithm suitable for solving multi-objective optimization problems with continuous and discrete variables is constructed and applied to the internal dimension optimization of the reducer, and finally the torsional stiffness of the designed reducer is verified by experiments. This article is organized as follows. In Section 2, the SG-Resnet based network for structure generating is proposed to generate serialized RV reducer structure parameters. In Section 3, the optimal internal parameter is calculated by modified NSGA-

II algorithm to increase torsional stiffness. In Section 4, the prototype of BAJ-25E is manufactured, and the torsional stiffness is measured to validate the proposed method.

2 Methodology

In this section, explanations will be presented for data preprocessing procedure, the proposed model block, and some tricks during the training process.

2.1 Resnet network system structure

In the optimal design of mechanical equipment structure, the BP backpropagation network is widely used, but it has the problem of insufficient expressive ability. Due to this, researchers have developed deep networks such as AlexNet/VGG, with more than 20 layers. However, continuing to increase the number of layers will cause the deep stacking architecture to degrade[18], and it is easy to fall into problems such as over-fitting under small sample conditions. He-Kaiming, Ren-Shaoqing, and Sun-Jian proposed the Resnet network in 2015, which has fewer network structure parameters than VGGNet, which simplifies the network structure and effectively improves the above problems. The difference between the residual network (Resnet) and the traditional deep network is its Residual structure, which allows the jump connection of the network, the so-called shortcut connection

2.2 Batch Normalization

Batch Normalization was proposed by Google's Sergey Ioffe and others. Its purpose is to solve the problem of inconsistent feature distribution between data sets, that is, covariate shift. This inconsistency will lead to gradient dispersion and reduce the training rate of the network. The purpose of regularization is to limit the difference of input eigenvalues to the range where the mean is 0 and the variance is 1. The disappearance of the gradient is avoided, and the convergence rate of the network is greatly improved. The calculation method is as follows:

$$y_i = \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta \tag{1}$$

In the formula, $\mathrm{x_i}$ and $\mathrm{y_i}$ are the input and output of the batch normalization layer, $\mathrm{\mu_B}$ and $\mathrm{\sigma_B^2}$ are the mean and variance of the input parameters, respectively, ϵ is a constant, and γ and β are hyperparameters that can be obtained through training.

2.3 activation function

The activation function is the key to achieving nonlinear mapping. Commonly used activation functions are shown in formulas 2.2~2.6. The original Residual Block uses a modified linear unit (ReLU) function as the activation function, which is closer to the response of the animal brain to the stimulus, considering the input When it is a negative value, the learning rate is lower, and Leaky-ReLU is used instead.

$$\mathrm{ReLU} \quad f(x) = max(0, x) \tag{2}$$

$$\mathrm{Leaky-ReLU} \quad f(x) = max(\alpha x, x) \tag{3}$$

2.4 Dropout algorithm

In the process of model training, due to too few training samples of the reducer, if the number of layers of the network is too large, over-fitting is prone to occur, that is, the generalization ability of the model is insufficient. Hinton proposed the Dropout method in 2012. Its essence is to set the hidden layer node value of some neurons to 0 according to a certain probability p, which can reduce the symbiosis between feature detectors. The essence is to generate A large number of sub-networks constructed randomly. The algorithm for applying Dropout on the *i*-th neuron is:

$$o_i = X_i a\left(\sum_{k=1}^{d_i} w_k x_k + b\right) = \begin{cases} a\left(\sum_{k=1}^{d_i} w_k x_k + b\right) & if\ X_i = 1 \\ 0 & if\ X_i = 0 \end{cases} \tag{4}$$

Among them, P (X_i=0) = p, w_k and b are related parameters of linear projection.

In addition to considering the risk of overfitting, the structural parameters of a single training sample are also considered to be incomplete. If the established Resnet model is overly dependent on the interaction of features, the prediction performance will be greatly affected. Therefore, the experiment uses the Dropout strategy to improve the Resnet pair Robustness with missing features.

2.5 Kriging proxy model

Using Latin hypercube experimental design sampling in the design variable space, the original sample points of the Kriging proxy model are obtained, and the minimum value of the following formula is the condition that the sample points meet:

$$\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{1}{\|x_i - x_j\|^2} \tag{5}$$

In the formula, the number of sample points is N; the distance between two sample points is $\|x_i - x_j\|$.

The relationship between the predicted value y and the design variable x is defined by the Kriging proxy model, and the expression is as follows[17]:

$$\mathrm{y}(x) = \mathrm{F}(\beta, x) + \mathrm{z}(x) \tag{6}$$

Among them, the global model of the design variable space is F (β, x), and the local deviation of random distribution according to N (0, σ^2) is z(x). The statistical characteristics of z(x) are as follows:

$$\mathrm{E}[z(x)] = 0 \tag{7}$$

$$\mathrm{Var}[z(x)] = \sigma_z^2 \tag{8}$$

$$\mathrm{cov}\left(z(x_i), z(x_j)\right) = \sigma_z^2 R(x_i, x_j) \tag{9}$$

In the formula, R (x_i, x_j) is the correlation model about sample points x_i and x_j, which is used to describe the degree of correlation between the sample points. The Gaussian correlation model is usually used:

$$\mathrm{R}(x_i, x_j) = \exp\left(-\sum_{k=1}^{n_v}\theta_k |x_i^k - x_j^k|^2\right) \tag{10}$$

In the formula, the relevant parameter vector to be determined is θ; the dimension of the design variable is n_v; θ_k and x_i^k, x_j^k are the kth component of θ, x_i and x_j, respectively.

The prediction variance and response value of the Kriging model at the prediction point x are obtained by the linear weighted interpolation method:

$$\hat{\mathrm{y}}(x) = \mathrm{F}(\beta, x) + r^T(x)R^{-1}(g - \hat{\beta}F) \tag{11}$$

$$\hat{e}^2(x) = \sigma^2\left[1 - r^T R r + \left[\frac{(1 - q^T R^{-1} r)^2}{q^T R^{-1} q}\right]\right] \tag{12}$$

In the formula, the correlation model vector between the sample point and the x point is r(x); the correlation model matrix is R; the unit column vector with the number of elements n_v and all

1 is q; the vector of the sample point response is g.

In order to effectively improve the accuracy of the Kriging proxy model, sample points can be added in the optimization process, usually adding points where the prediction variance is large or expected to improve (EI) as the Kriging adding point criterion. The Pareto optimal set is introduced into the point-adding criterion to effectively use the information of evolutionary algorithm in the optimization process. The number of points $\mathrm{S_E}$ where the prediction variance is large and the number of points $\mathrm{S_P}$ where the Pareto optimal solution is added in a single evolution are:

$$S_E = \left\lfloor \frac{g - g_c}{g} \cdot C_{E_1} + C_{T_1} \right\rfloor \tag{13}$$

$$S_P = \left\lfloor \frac{g_c}{g} \cdot C_{E_2} + C_{T_2} \right\rfloor \tag{14}$$

In the formula, the current number of iterations is $\mathrm{g_c}$; the total number of iterations is g; the adjustment coefficient and the minimum number of points added are $\mathrm{C_E}$ and $\mathrm{C_T}$, respectively. During the evolution process, the strategy of updating the Kriging model is to select scattered sample points according to the crowded distance of the Pareto optimal set. The dynamic number of points not only improves the efficiency of solving the optimal set, but also satisfies the accuracy of the model.

2.6 Pareto selection based on entropy method

The entropy method is an objective weighting method that uses the entropy of the decision index to calculate the entropy weight. In the evaluation problems of i evaluation indexes such as $\mathrm{X_1, X_2, \dots, X_i}$, and j Pareto optimal solutions, the mth One evaluation index $X_m = \{x_1, x_2, \cdots, x_j\}$, define the entropy H_m of X_m as:

$$H_m = -k \sum_{n=1}^{j} f_{mn} ln f_{mn} \ (m = 1,2, \cdots i) \tag{15}$$

Where: $f_{mn} = \frac{r_{mn}}{\sum_{n=1}^{j} r_{mn}}$; $k = \frac{1}{lnj}$

Among them, $\mathrm{r_{mn}}$ is the normalized index, $r_{mn} = \frac{X_{mn} - min(X_m)}{max(X_m) - min(X_m)}$; the greater the entropy of the index, the smaller the entropy weight, the mth The entropy weight $\mathrm{w_m}$ of each performance index is defined as:

$$w_m = \frac{1 - H_m}{i - \sum_{m=1}^{i} H_m} \tag{16}$$

Using the entropy weight to rank the scores of each scheme, a more objective Pareto optimal solution can be obtained.

3 Novel SG-Resnet based network for structure generating

3.1 Structure of the proposed model

Resnet includes 18 layers, 34 layers, 50 layers, 101 layers and 152 layers. Although the depth of the network is an important guarantee for the effect of deep learning, in practice, its depth is limited by many aspects: the more the number of network layers, the higher the GPU performance requirements, the more memory requirements; the lower the iteration efficiency, the slower the convergence speed; the network depth is too deep, and overfitting may occur. Therefore, considering the complexity of the problem and the running performance, the construction of SG-Resnet is realized on the basis of Resnet18.

3.1.1 Modified Resnet-18

The solution in this paper is based on the deep residual network Resnet proposed by He et al. [19], but its core structural parameters and training strategies are optimized and adjusted to obtain a model framework suitable for generating structural parameters of RV reducers. Considering that the convolutional layer is more suitable for extracting image features, and the structural parameters of the RV reducer are abstract data, the convolutional layers in the long and short connections in the Residual Block are replaced with linear transformation layers (nn.Linear). The system structure of SG-Resnet is shown in Figure 1.

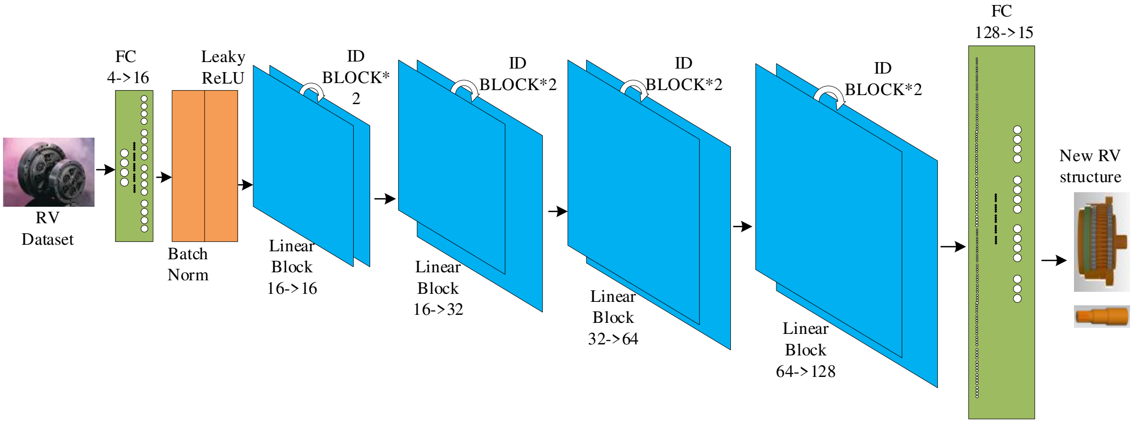


Fig. 1 SG-Resnet network structure diagram

The structure of Resnet18 is shown in Figure 2. The core of Resnet18 network consists of 2 conv2_x (2 convolutional layers), 2 conv3_x (2 convolutional layers), 2 conv4_x (2 convolutional layers), 2 conv5_x (2 convolutional layers), its purpose is to extract the features of the data. The first layer is a 7×7 convolutional connection pooling layer, and the last layer is a 512-dimensional fully connected layer, which implements a linear transformation of a specific number of feature spaces.

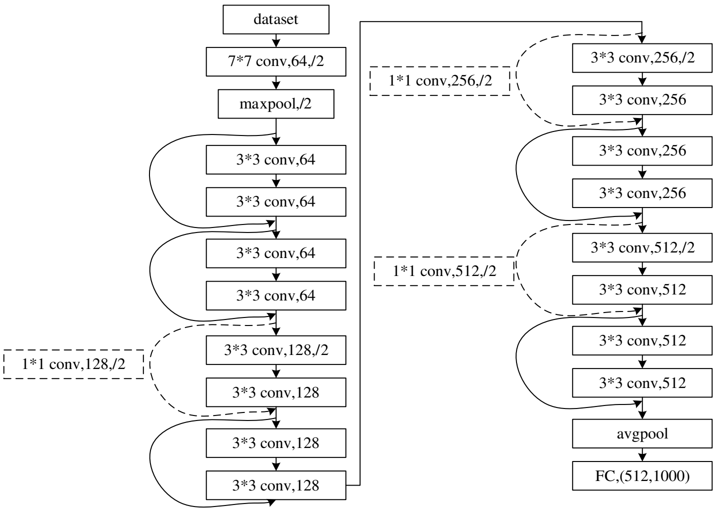
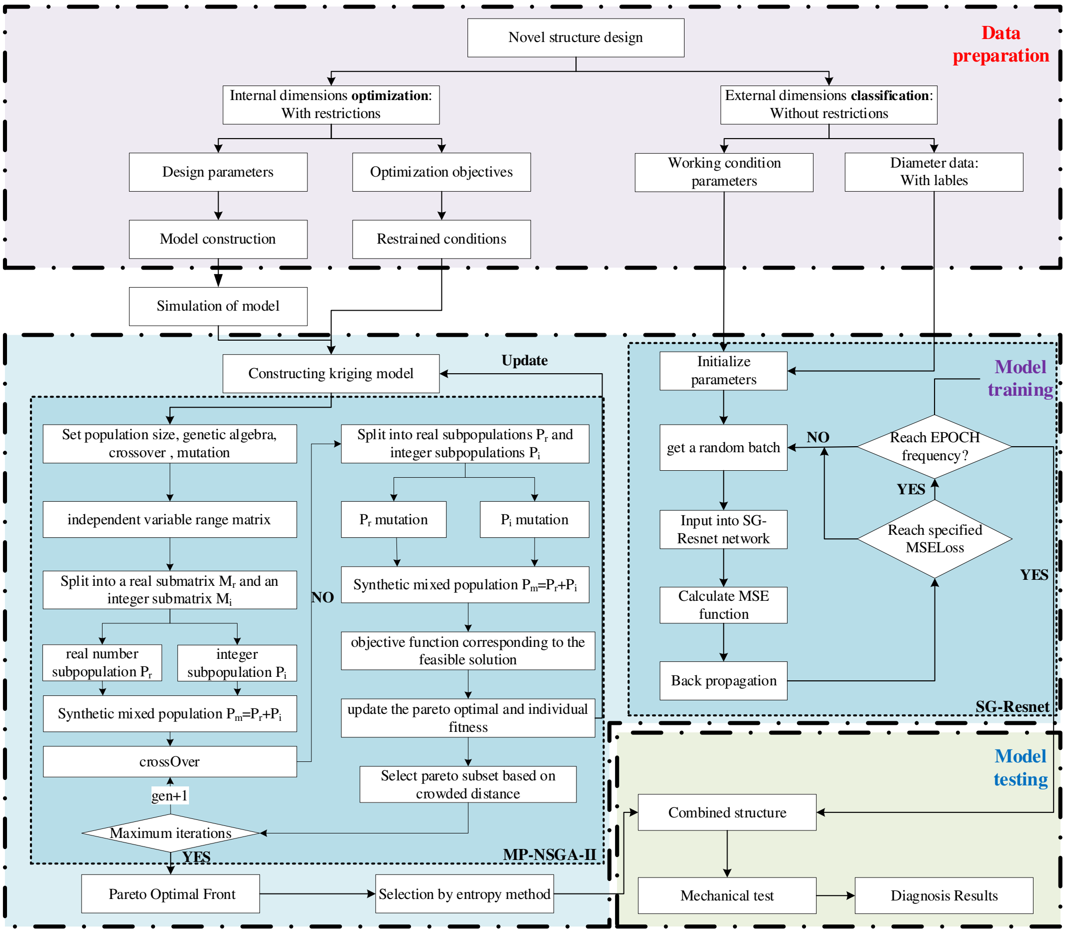


Fig. **Error! No text of specified style in document.** Typical structure of Resnet18

Multi-objective genetic algorithm (Multi-objective genetic algorithm) has the ability to search for optimal solutions globally, and is suitable for dealing with optimization problems of discrete variables. The current practice of multi-objective genetic algorithms mainly focuses on multi-objective optimization algorithms using the concept of Pareto optimal solutions. Among them, the second-generation non-dominated sorting genetic algorithm (Non-dominated sorting genetic algorithm-II) has the advantages of good solution set distribution, low time complexity, and fast convergence speed.



Novel structure design
Data preparation
Internal dimensions optimization: With restrictions
External dimensions classification: Without restrictions
Design parameters
Optimization objectives
Working condition parameters
Diameter data: With lables
Model construction
Restrained conditions
Simulation of model
Constructing kriging model
Update
Initialize parameters
Model training
Set population size, genetic algebra, crossover , mutation
Split into real subpopulations P_r and integer subpopulations P_i
get a random batch
NO
Reach EPOCH frequency?
independent variable range matrix
P_r mutation
P_i mutation
YES
Input into SG-Resnet network
Reach specified MSELoss
Split into a real submatrix M_r and an integer submatrix M_i
Synthetic mixed population P_m=P_r+P_i
NO
YES
Calculate MSE function
objective function corresponding to the feasible solution
real number subpopulation P_r
integer subpopulation P_i
Back propagation
Synthetic mixed population P_m=P_r+P_i
update the pareto optimal and individual fitness
SG-Resnet
Model testing
crossOver
Select pareto subset based on crowded distance
gen+1
Combined structure
Maximum iterations
YES
MP-NSGA-II
Pareto Optimal Front
Selection by entropy method
Mechanical test
Diagnosis Results

Fig.3 Modified mixed population NSGA-2 for internal structure optimization flow chart

The initial step to solve the optimization problem is to randomly generate the coding population according to the design variable range matrix. In the design variables $\{z_1, z_2, z_g, Z, m\}^T$ are discrete variables, $\{b, D_z, d_z, B, K_1, D_m, D_r, L\}^T$ is a continuous variable. Improved on the basis of the evolutionary algorithm library Geatpy. Aiming at the problem that the NSGA-II algorithm can only deal with a single type of population, the design variable range matrix is divided into a real number independent variable range sub-matrix and an integer independent variable range sub-matrix. The mutation stage is processed separately and synthesized into the same population in other stages. At the same time, non-dominated sorting considering the crowding distance is introduced to enhance the distribution of the population. The flow of the mixed population NSGA-II algorithm is shown in Figure 3.

The specific processing steps for mixed populations are:

① Initial population generation: The independent variable range matrix is generated from the value range of the design variable, and it is divided into two sub-matrices M_i and M_r using the sequence of continuous variables and discrete variables in the matrix. Use the rand() function to generate a real-valued initial population P_r corresponding to the sub-matrix M_r, and use the rand() function and rounding method to generate a decimal integer initial population P_i corresponding to the sub-matrix M_i.

② Cross-mixed population: The populations P_i and P_r are both the number of columns as the design variable number, the number of rows is the matrix of population size, and the two matrices are merged horizontally to form the mixed population P_m. The reorganization of chromosomes between individuals is realized by single-point crossover.

③ Subpopulation mutation: divide the mixed population into subpopulations P_i and P_r. The integer value mutation operator is realized by the mutation of the individual in the matrix P_i, and the mutation operation on P_r is completed by the real-valued Gaussian mutation operator. The mixed population P_m consists of two matrices. synthesis.

Since the truncation method [12] only applies to the processing of continuous integer variables, in order to solve the problem of processing discrete variables in the MIP problem, this chapter proposes a general coding scheme for discrete variables based on the array index as shown in Figure 4. The array Z_{arr} represents n_d possible values of the discrete variable Z_d, and the integer subpopulation Z_{num} with the value (0,1, ..., n_d-1) is the encoding form of the array index. In the stage of evaluating the constraints and function values, the discrete subpopulation Z_{dis} is decoded by the array index population Z_{num} and the array Z_{arr}.

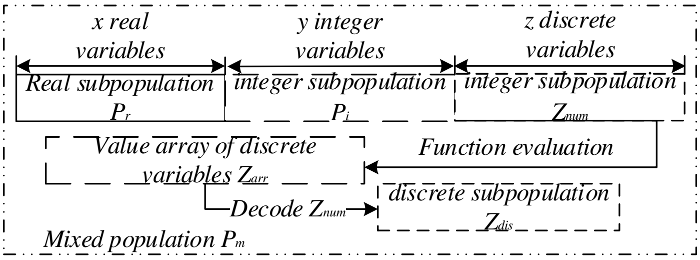


Fig.4 A universal coding scheme for discrete variables

Introduce non-dominated sorting considering crowded distance to reduce the computational cost and enhance the distribution of the population. Since the calculation efficiency of the Euclidean distance is low, the crowded distance is represented by the percentage of the offset calculation target value. The crowding distance between two adjacent individuals x_i, x_j can be defined according to the objective function value of the mixed population individuals sorted from small to large:

$$d_c = \frac{f(x_j) - f(x_i)}{f(x)_{max} - f(x)_{min}} \tag{17}$$

In the formula, the objective function value of individual x is f(x), and the smallest and largest objective function values in the mixed population are $f(x)_{min}$ and $f(x)_{max}$. In order to select the next generation of individuals, the individual fitness V_{fit} is updated based on the crowding distance.

Table 1. benchmark simulation of MINLP

No.	SOTA	MP-NSGA-II	Best gen
P1	85.500000	85.50002	3417
P2	7.666566	7.667180	1
P3	4.579582	4.579589	2034
P4	2.000000	2.000001	2699
P5	2.124470	2.124468	6444
P6	1.076555	1.076625	9994

P7	99.239635		
P8	3.557464	3.558061	89
P9	-32217.427	-35270.129	8
P10	-0.793323	-0.793323	50
P11	-0.974565	-0.974565	1
P12	-0.999949	-0.999954	153
P13	5850.38376	5850.50079	1625
P14	-75.134168	-75.133805	3864

The 14 MINLP problems in the literature[14] are simulated to evaluate the feasibility of the MP-NSGA-II algorithm. The existing best comes from ridPSO, ridDE, MDE, MDELS and MDEIHS.

Table 1 shows 14 minimization target problems, and the mixed population NSGA-II is a population size of 1000. The evolution algebra is set to 10000 to show the convergence characteristics of the algorithm. The gap between the known optimal and the single-run mixed population NSGA-II optimal solution is less than 0.1%, and the 3 known optimal solutions are improved: P5(x=1.3748231,y=1), P9(x=[27, 27,27], y=[88,44]), P12(x=[0.90219,0.88775,0.94918, 0.84872], y=[5,5,4,6]). When the independent variable x_2=0, the existing optimal objective function denominator in P7 is 0, this group has not been compared.

3.3 General procedure of the proposed system

Resnet includes 18 layers, 34 layers, 50 layers, 101 layers and 152 layers. Although the depth of the network is an important guarantee for the effect of deep learning, in practice, its depth is limited by many aspects: the more the number of network layers, the The higher the GPU performance requirements, the more memory requirements; the lower the iteration efficiency, the slower the convergence speed; the network depth is too deep, and overfitting may occur. Therefore, considering the complexity of the problem and the running performance, the construction of SG-Resnet is realized on the basis of Resnet18, and the MP-NSGA-II is a discrete version of NSGA-II. The overall process of new type RV reducer design is illustrated in figure 3. (Data preparation->Model training ->Model testing).

4 Case studies and results

4.1 Experiment preparation

The experimental platform for comparing the Resnet model and other methods proposed in this chapter is Windows10 64bit system, Intel core i9-9900 2.6GHz, 16GB RAM, 1TB SSD, NVIDIA GTX 1070 8GB. The algorithm is implemented using Pytorch 1.3+CUDA10.1 framework and Python 3.6.

For the RV reducer with integrated structure, the arm bearing adopts integrated bearing. The three-dimensional model is shown in Figure 5. Considering the size of the standard parts and the parts to be matched, the external dimensions to be determined are $S = \{y_{v1}, y_{v2}, y_{v3}, y_{v4}, y_{v5}, y_{v6}, y_{v7}, y_{v8}, y_{h1}, y_{h2}, y_{h3}, y_{h4}, y_{h5}, y_{h6}, y_{h7}\}$, a total of 15 size parameters, reducer output speed n_2 (rpm), output torque T_2 (N·mm), the input power P (w), and the number of planet wheels n_p are the original input samples. Corresponding working condition parameters of each model of RV reducer C=$\{n_2,T_2,P,n_p\}$ There are a large difference within 9 categories, a total of 9 types; the difference between the categories is small, a total of 8 types of reducers, the simulation of small sample conditions can be fully investigated The generalization ability of the model, so as to fully consider the performance of the structure generation algorithm in the case of small sample fine-grained model generation.

4.2 Experimental setup

SG-Resnet uses the SGD optimizer by default, the optimizer hyperparameter learning rate= 1×10^{-6} , Momentum=0.5, the loss function selects the mean square loss function (MSELoss), according to the optimization process of Resnet network, the training process of SG-Resnet is shown in Figure 3.

In order to prevent overfitting of the function, the dropout method is adopted, and the hyperparameter p=0.5. On the one hand, during each training, inactive nodes will appear randomly in the network. When different training samples are input to the network, they correspond to different network structures, preventing the assimilation of the structure. Drawing on the method of image augmentation (data augmentation), transforming the initial data to augment the training set, and adding Gaussian noise to the original data can effectively improve the learning ability of the system.

Step1. External dimensions classification (without restrictions)

Data description

The integrated structure RV reducer is a new type of high-precision RV reducer proposed by this research team. Its structure is shown in Figure 5. The new high-precision RV reducer input mechanism intermediate arm bearing adopts an integrated design integrated in the double eccentric shaft and cycloid gear coordinate hole full cylindrical roller bearing structure, the output mechanism support bearing adopts an integrated design integrated in the planet carrier and Staggered roller bearing structure of pin gear shell.

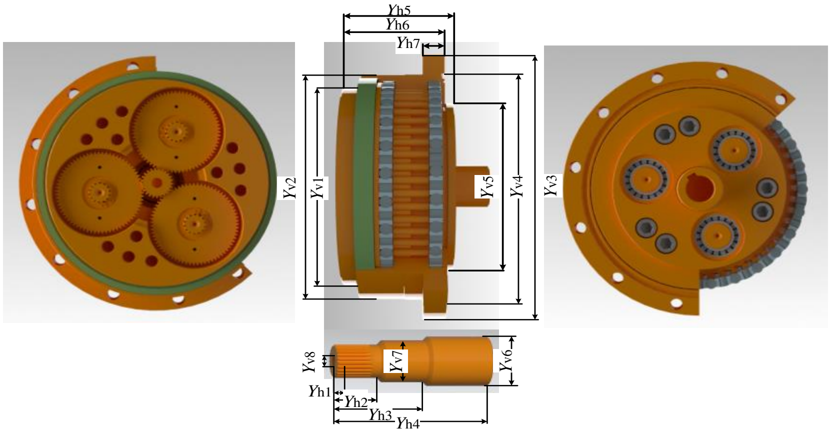


Fig.5 RV reducer model with integrated structure

Jointly carried out industrial development with a cooperative reducer manufacturer, and formed an integrated RV reducer structure parameter data set based on the industrial development process. Choose RV-6E, RV-20E, RV-40E, RV-80E, RV-110E, RV-160E, RV-320E, RV-450E, a total of 8 models, each model corresponds to 9 working condition, a total of 72 pairs of data sample.

Parameter selection of the proposed method

The improved Resnet model evaluates the accuracy and generalization performance of the model through two indicators during the training process: train loss and test loss. Since the mechanism to prevent overfitting in the model is fully considered, no verification set is introduced (Val).

In the model training stage, the basic structure of the model can be kept unchanged, and the network performance can be optimized by adjusting the batch size when the network weights are updated. Set the batch size to 8, 16, 24, and 32, respectively. The 18-layer SG-Resnet iteration loss function changes in 7,500 epochs as shown in Figure 2.8. In the figure, epoch is the number of traversal training samples, and loss is the value of the loss function.

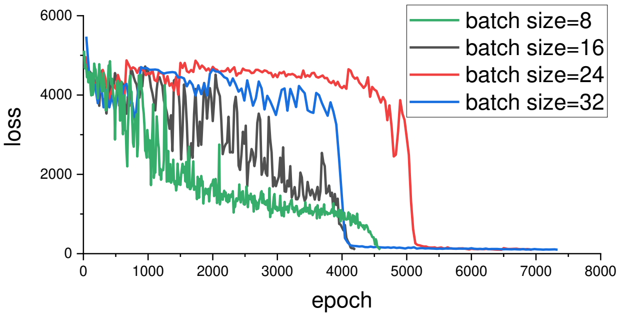


Fig.6 SG-Resnet iterative loss trend

It can be seen from Figure 6 that during the Mini-Batch Gradient Descent (MBGD) process, the general trend is that the larger the batch size, the smaller the loss function oscillation and the better the directionality of the gradient descent. The small batch size can bring more random weight updates and it is easier to jump out of the local optimal. When the batch size is 16, the system has the optimal convergence ability. When the batch size is 24 and 32, respectively, the loss function has a cliff-like decline phenomenon. Because it is easy to fall into the local optimal solution, the convergence of the loss function value is below 150 The speed is slow. When the batch size is 8, the convergence speed is the highest before 4000 epochs, and the convergence ability decreases due to excessive fluctuations in the later period. In the study of Keskar et al.[20], this phenomenon was theoretically discussed. When the batch size reaches a larger scale, the model tends to converge to the steep minimum value of the loss function, and a small difference in the parameter space will lead to the validation set, the larger the prediction offset, the smaller the batch size, the smaller the relative error on the new data set, and the easier it is to converge to a smooth minimum.

Model comparison and results

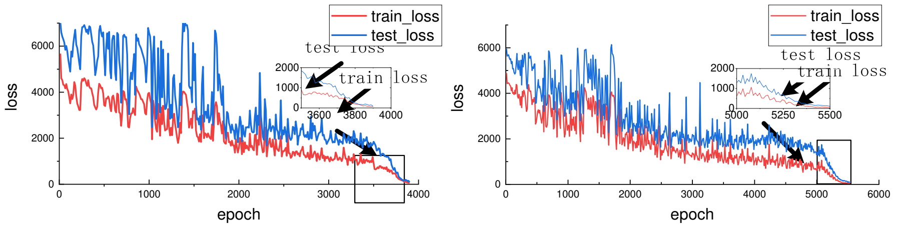
To validate the performance of the SG-Resnet model, the training sample set, the test sample set, and the ratio of the training sample and the test sample all have a greater impact on the accuracy and generalization ability of the model. The experiment can be carried out in the following three ways, according to these three methods are as follows:

1) Use different ratio samples for training and testing;

2) Use the same sample for cross-training and testing;

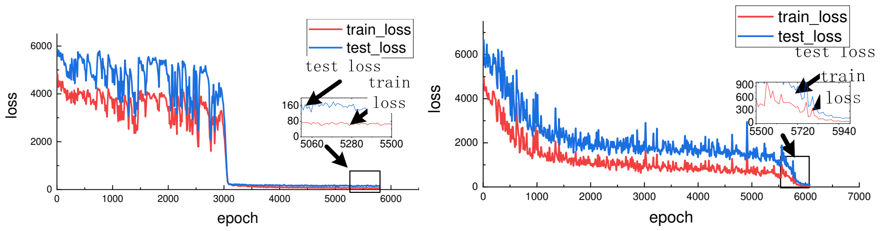
3) Use a fixed ratio of samples for training and testing, the results of the experiment;

① When using samples of different proportions for training and testing, the 18-layer SG-Resnet network is uniformly selected, RV-80E is used as the test set, and the remaining part is used as the training set, and the network performance is observed experimentally.



(a) Test set：RV-80E

(b) Test set：RV-40E RV-160E



(c) Test set：RV-20E RV-80E RV-160E (d) Test set：RV-20E RV-80E RV-160E RV-450E

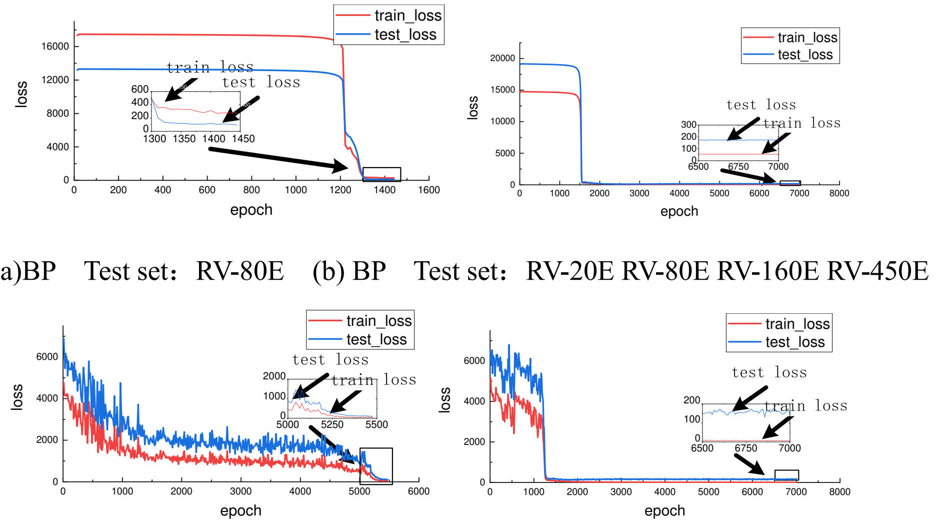
Fig.7 Loss trend of training set and test set with different proportions of samples

As shown in Figure 7, this SG-Resnet network can converge on the test set when isolating the data sets of different types of RV reducers, indicating that the network can correctly predict the size parameters of the reducer under unknown conditions, and has strong learning and reasoning ability. When the training set: test set=7:1, the loss of the test set in the pre-training period is always higher than that of the test set, and the convergence speed increases after 3500 epochs; when the training set: test set=6:2, the training samples are reduced, Convergence starts to accelerate after 5000 epochs; when training set: test set=5:3, although the loss function drops to about 200 after only 3000 epochs, as the iteration proceeds, the test set loss function value does not decrease but increases. The phenomenon of over-fitting indicates that for the problem of too few samples, the training sample is not enough to cover all the effective size features in the test sample, and the complexity of this model is higher than the complexity of the problem itself. When training set: test set=4:4, the model converges the slowest, and over-fitting also occurs.

Table **Error! No text of specified style in document.**. Structure of 18-layer and 34-layer model

layer name	18-layer	34-layer
FC1	16-d fc	
Linear2_x	$\begin{bmatrix}16\\16\end{bmatrix}\times 2$	$\begin{bmatrix}16\\16\end{bmatrix}\times 3$
Linear3_x	$\begin{bmatrix}32\\32\end{bmatrix}\times 2$	$\begin{bmatrix}32\\32\end{bmatrix}\times 4$
Linear4_x	$\begin{bmatrix}64\\64\end{bmatrix}\times 2$	$\begin{bmatrix}64\\64\end{bmatrix}\times 6$
Linear5_x	$\begin{bmatrix}128\\128\end{bmatrix}\times 2$	$\begin{bmatrix}128\\128\end{bmatrix}\times 3$
FC6	15-d fc	

② When using the same sample for cross-training and testing, select 18-layer BP neural network, 18-layer and 34-layer SG-Resnet models for comparative experiments. The BP network is based on the multilayer feedforward network proposed by McCelland and Rumelhart in 1986. Error back propagation algorithm, the batch size is uniformly set to 8. Compared with the 18-layer SG-Resnet, the BP network has the same parameters except for the lack of the Residual Block structure. Table 2 shows the structure of the 18-layer and 34-layer SG-Resnet model.



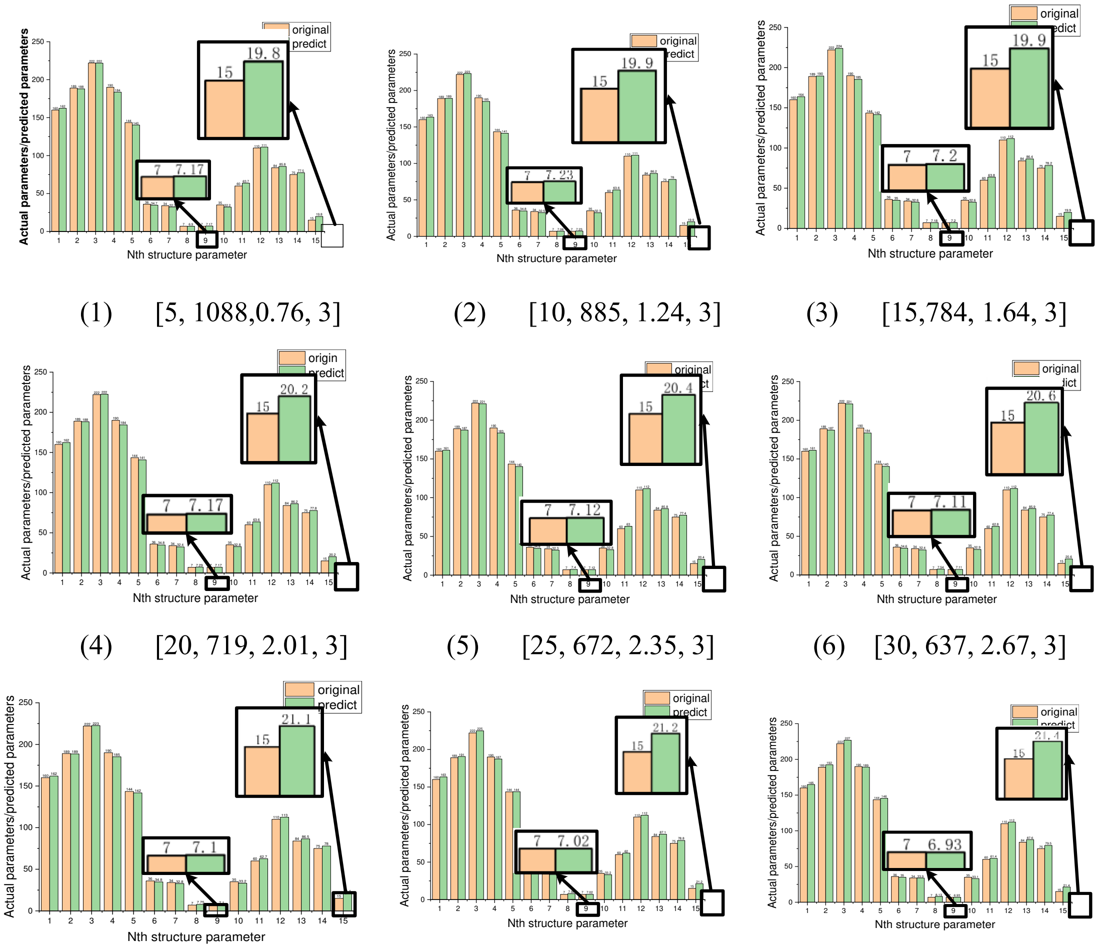
(a)BP Test set：RV-80E (b) BP Test set：RV-20E RV-80E RV-160E RV-450E

(c)SG-Resnet18 Test set：RV-80E (d) SG-Resnet34 Test set：RV-80E

Fig.8 Loss trend of different network types

The results in Figure 8 show that under the same experimental environment (GPU floating-point computing ability, video memory capacity), the ordinary 18-layer BP network has no residual structure, lacks the jump connection of the network, and the early loss function value is higher when the training samples are sufficient. And the loss function is very small. When the training sample: test sample=7:1, the convergence starts to accelerate at 1300 epochs, but the training sample: test sample=4:4, although the convergence is extremely fast, the test set loss When the function value reaches about 150, it does not decrease but rises, indicating that the model has low flexibility, insufficient generalization ability, and is easy to overfit and lose its prediction credibility; although the convergence rate of SG-Resnet18 is low, there is no overfitting. The phenomenon indicates that the model accuracy is very high; when the 34-layer SG-Resnet is used, serious over-fitting occurs again, indicating that the two deep networks leads to the high complexity of the model. In summary, it is necessary to adopt an improved network structure for small sample problems, and SG-Resnet with an improved structure still needs to adjust the network depth appropriately according to the characteristics of the sample, and reasonably select the dropout probability p and other hyperparameters.

③Through the foregoing experimental conclusions, an 18-layer SG-Resnet network is selected, RV-80E is the test set, the batch size is 8, the dropout probability p = 0.7, and the network size is appropriately reduced to test the performance of this system. When epoch=72455, the loss function values of the training set and the test set are 4.297 and 8.748, respectively. This model obtains the analysis results of the consistency between the real structure parameters and the predicted structure parameters.



(1) [5, 1088,0.76, 3] (2) [10, 885, 1.24, 3] (3) [15,784, 1.64, 3]

(4) [20, 719, 2.01, 3] (5) [25, 672, 2.35, 3] (6) [30, 637, 2.67, 3]

(7) [40, 584, 3.26, 3] (8) [50, 546, 3.81, 3] (9) [60, 517, 4.33, 3]

Fig.9 Experimental results of RV-80E structural parameter prediction

The results in Figure 9 show that the size parameters predicted by the 9 pairs of working conditions of RV-80E differ very little from the actual size parameters, and only the 15th structural parameter y_{h7} has a large difference. The RV-80E is used as a test set. The structure of the RV-80E is unknown to this system, indicating that the network system has a certain "structural design capability". Combining the structural size parameters generated by SG-Resnet18 and the designers' actual design parameters in the database, calculate the correlation parameter values between subjective and objective design parameters: Pearson's r, R^2 (COD) and adjusted R^2. The results are shown in Table 3. The correlation parameter values of the three evaluation methods are all higher than 0.998, indicating that the proposed SG-Resnet model has high accuracy.

Table 3. Correlation parameter values of 9 samples of RV-80E

Eval	Spl 1	Spl 2	Spl 3	Spl 4	Spl 5	Spl 6	Spl 7	Spl 8	Spl 9
Pearson's r	0.99928	0.99932	0.99933	0.99932	0.99933	0.99935	0.99939	0.99946	0.99948
R^2 (COD)	0.99856	0.99863	0.99866	0.99863	0.99866	0.99870	0.99878	0.99892	0.99897
adjust R^2	0.99845	0.99853	0.99855	0.99853	0.99856	0.99860	0.99868	0.99883	0.99889

Step2.

Internal dimensions optimization (with restrictions):

The optimization objectives of this paper are transmission efficiency η, volume V and torsional stiffness K′, and the constraint function is geometric constraint and stress constraint, then the optimization model is:

$$\begin{cases} min[-\eta(X,C),V(X,C),-K'(X,C)] \\ s.t.\ g(X,C) \le 0 \\ h(X,C)=0 \end{cases} \tag{18}$$

In the formula, g (X, C) is an inequality constraint; h (X, C) is an equality constraint.

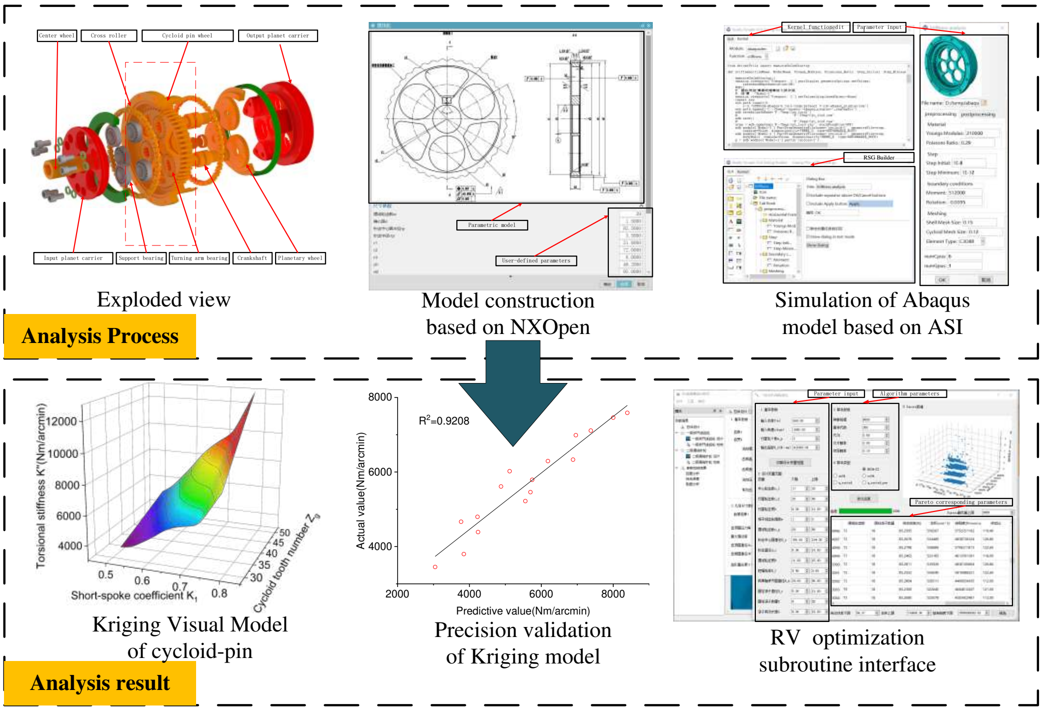


Fig.10 the flow chart of building kriging model

The parameter settings of the mixed population NSGA-II algorithm are shown in figure 10. Run this algorithm many times at the same time to verify that the stability of the Pareto optimal solution is extremely high. Run the optimization program to get the Pareto frontier as shown in Figure 11. It can be seen that the three optimization goals are mutually restricted, and the ideal solution needs to be selected from the Pareto frontier solution set.

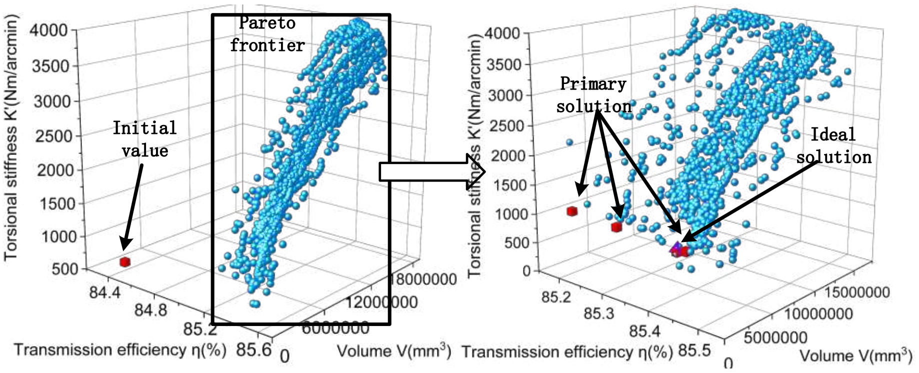
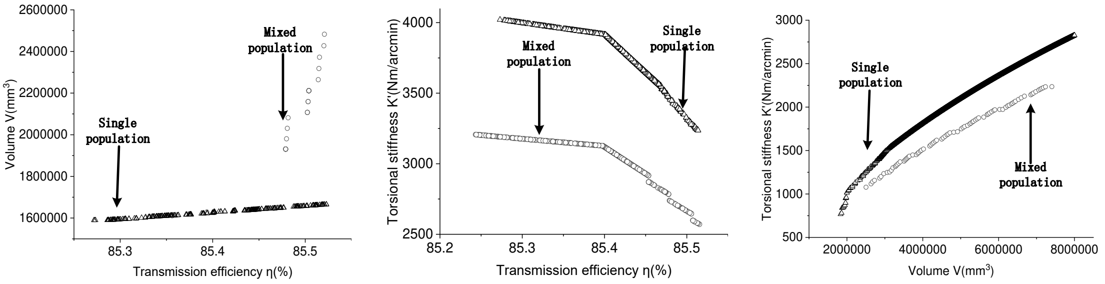


Fig.11 Pareto frontier

Since the objective functions compete with each other through the structure and size parameters, the solution set of the Pareto front end can be compared when the mixed population and the single population are used. At the same time, the Pareto front end of the two objectives is used to study the coupling relationship between the objective functions of the RV reducer, as the follow-up Pareto The basis for selection is that the single population removes the constraints of discrete variables and integer variables. The relationship curve is shown in figure 12. Because part of the independent variables in the mixed population are discrete variables, the Pareto front end appears discontinuous.



(a) (b) (c)

Fig.12 Competitive relationship between objective functions

Figure 12(a), figure 12(b), and figure 12(c) show the competitive relationship between transmission efficiency and volume, transmission efficiency and torsional stiffness, volume and torsional stiffness, respectively. When volume and torsional stiffness change, it is shown in figure 12(a) As shown in figure 12(b), the transmission efficiency of 85.2%~85.6% is basically maintained, so the coupling level of the remaining objective functions and transmission efficiency is not obvious. When the optimization goal tends to reduce the volume, as shown in figure 12(c), it will result in a substantial decrease in torsional stiffness; if the torsional stiffness of the reducer is increased, the volume will increase. Therefore, based on the design requirements of torsional stiffness and volume, the optimization scheme is initially screened as shown in figure 10.

4.3 Experiment analysis

1) Torsional stiffness of RV reducer and its testing principle

Fix one end of the input and output ends of the RV reducer, and apply torque at the other end. Due to the internal clearance of the reducer and the elastic deformation of the material, the loading end will produce a certain degree of torsion. The rotation applied in this process The ratio of the moment value to the generated torsion angle is called the torsional stiffness of the RV reducer. In the actual test, the movement of the input end of the tested reducer is usually restricted, and the output end is loaded to the rated torque of the reducer, and the corresponding torque and rotation angle of a section near the rated torque of the hysteresis curve are calculated. As shown in Figure 13, b is the torque value near the rated torque, and a is the corner value corresponding to b. The formula for calculating the torsional stiffness of the RV reducer is:

$$k = \frac{b}{a} \tag{19}$$

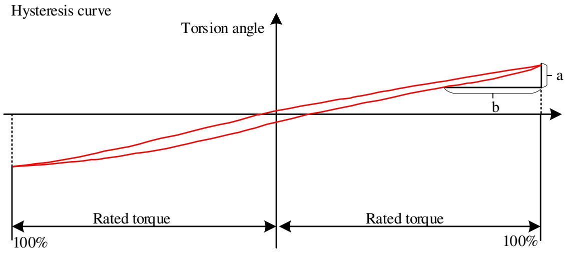


Fig.13 Schematic diagram of torsional stiffness calculation

There are many types of RV reducers, and their rated torque is as small as tens of Nm (for example, the rated torque of RV-6E is 58Nm), and as large as several thousand Nm (for example, the rated torque of RV-450E is 4410Nm). When using the existing calculation method to design the torsional stiffness test instrument, the load torque of several thousand Nm poses a great challenge to the load capacity of the entire system and the torsional stiffness of the shafting. Especially in the batch test of RV reducer, a large range of reciprocating loading will reduce the detection efficiency of the product, and also shortens the service life of the detection device.

2) Torsional measurement of BAJ-25E

According to the parameters in Figure 10, an RV reducer is designed (as shown in Figure 14), and the torsional stiffness of the RV reducer in different torque ranges are tested using a special RV reducer torsional stiffness test device.

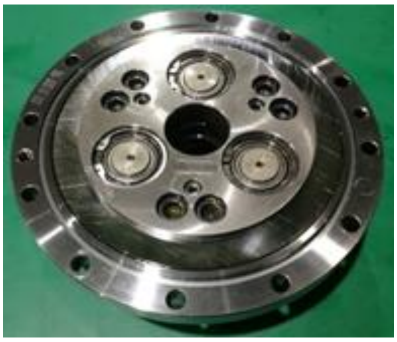


Fig.14 RV reducer prototype

The torsional stiffness test device of RV reducer is shown in Figure 15. During the test, the input end of the RV reducer is fixed, the loading device loads its output end, and the torque sensor and the rotation angle measuring device obtain the torque and rotation angle values of the output end respectively. The upper computer displays the hysteresis curve of the tested reducer in real time. The specific loading process of the output terminal is as follows: firstly load to the maximum torque value in the positive direction, then the motor reverses, unload the torque to 0 and continue to load to the negative maximum value, and finally unload the load to 0 again and load in the forward direction until Load to coincide with the curve of the first forward load. In this test, the torsional stiffness of the RV reducer are tested four times. The maximum torque values in the four tests were 1100Nm, 900Nm, 700Nm and 500Nm, respectively. The test results are shown in Figure 16~19.

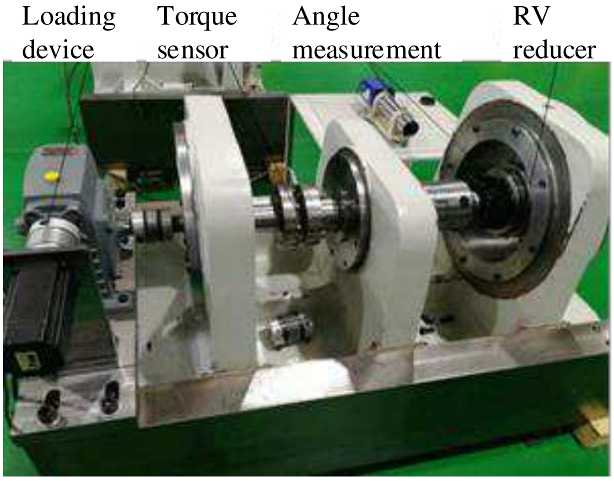


Fig.15 RV reducer stiffness test

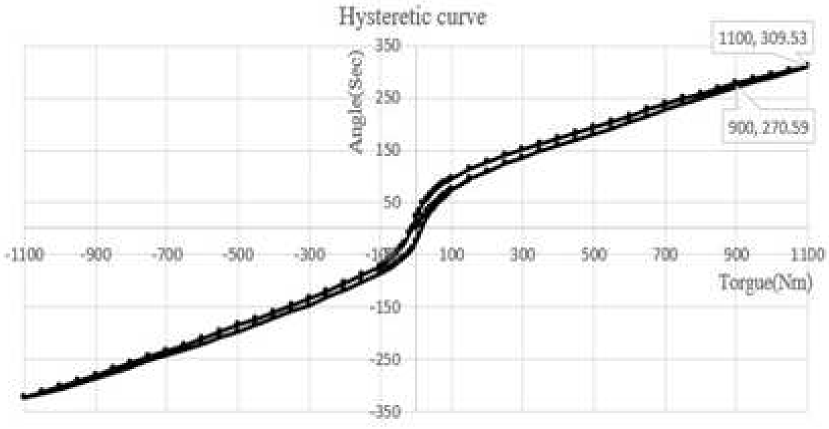


Fig.16 Hysteresis curve of 1100Nm

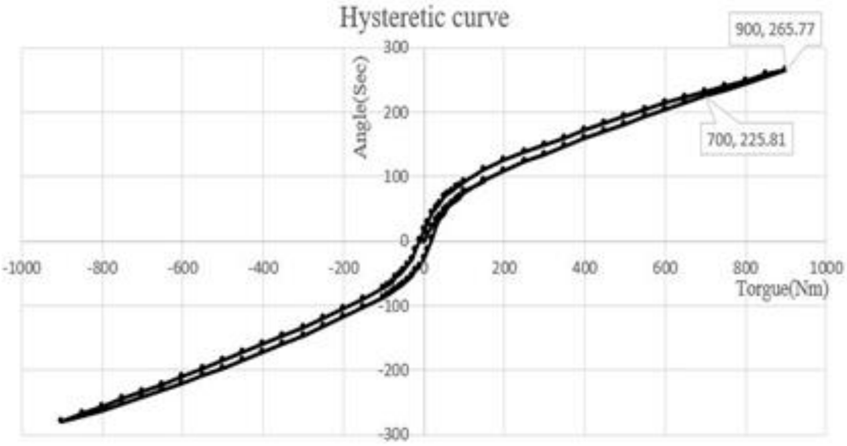


Fig.17 Hysteresis curve of 900Nm

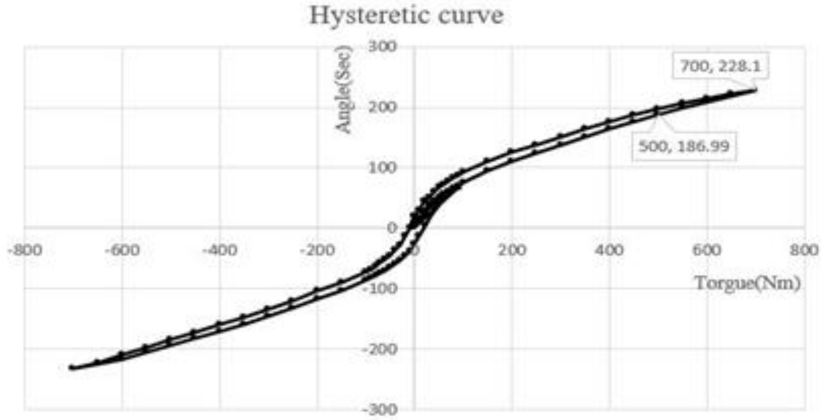


Fig.18 Hysteresis curve of 700Nm

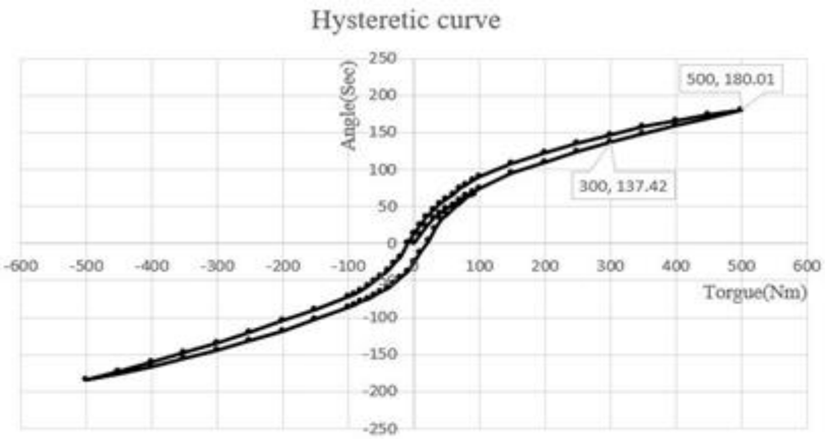


Fig.19 Hysteresis curve of 500Nm

It can be seen from the hysteresis curve of the RV reducer that, except near the origin of the coordinate, the load torque and torsion angle of the RV reducer in other ranges have a good linear relationship, which indicates the torsional stiffness within this range. The amount of change is small. It can be seen from the results of the test data analysis that the torsional stiffness of the RV reducer increases with the increase of the load torque, and the torsional stiffness changes in different torque ranges are roughly the same as the results of the finite element analysis.

5 Conclusions

Based on the research of the small-sample structure parameter recommendation algorithm, a novel NSGA-II and ResNet based model has been proposed according to the structural characteristics of the integrated structure RV reducer. The generalization performance and accuracy of the proposed model are significantly higher than the Backpropagation model. Compared with the actual design parameters of the engineering designers, the correlation of the structural size parameters generated by this method can reach 0.998, which verifies the rationality of the design of serialized reducers using the parameter recommendation method.

Through the Eric6 development environment and a series of function interfaces of PyQt5, the forward design software of the integrated structure RV reducer is realized, the integration of the RV reducer component design and verification function is completed, and the MP-NSGA-II optimization module is integrated in the software by utilizing the Geatpy framework.

In order to realize the practical application of the intelligent structure design method of RV reducer, in the future research, the application of multidisciplinary design optimization technology in the design of RV reducer will be explored, and the auxiliary design program of RV reducer under multiple platforms will be integrated.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported by the National Key Research and Development Program of China (Grant

No. 2019YFB2006402).

Nomenclature

References

[1] Lin L, Guo F, Xie X. Novel informative feature samples extraction model using cell nuclear pore optimization[J]. Engineering Applications of Artificial Intelligence, 2015, 39: 168-180.

[2] Shu L H, Cheong H. A natural language approach to biomimetic design[M]//Biologically Inspired Design. Springer, London, 2014: 29-61.

[3] Nagel J K S, Nagel R L, Stone R B, et al. Function-based, biologically inspired concept generation[J]. Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AI EDAM, 2010, 24(4): 521.

[4] Shimizu S, Tosha K, Tsuchiya K. New data analysis of probabilistic stress-life (P–S–N) curve and its application for structural materials[J]. International Journal of Fatigue, 2010, 32(3): 565-575.

[5] Haixiang G, Yijing L, Shang J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications, 2017, 73: 220-239.

[6] Habtemariam B, Tharmarasa R, Thayaparan T, et al. A multiple-detection joint probabilistic data association filter[J]. IEEE Journal of Selected Topics in Signal Processing, 2013, 7(3): 461-471.

[7] Rezaii T Y, Tinati M A. Distributed multi-target tracking using joint probabilistic data association and average consensus filter[J]. Annals of telecommunications-annales des télécommunications, 2011, 66(9-10): 553-566.

[8] Samet H. K-nearest neighbor finding using MaxNearestDist[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 30(2): 243-252.

[9] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[10] Hasnat A, Ghosh A, Khatun A, et al. Pattern classification of fabric defects using a probabilistic neural network and its hardware implementation using the field programmable gate array system[J]. Fibres & Textiles in Eastern Europe, 2017.

[11] Zhang Y, Lu Z, Li J. Fabric defect classification using radial basis function network[J]. Pattern Recognition Letters, 2010, 31(13): 2033-2042.

[12] Deep K, Singh K P , Kansal M L, et al. A real coded genetic algorithm for solving integer and mixed integer optimization problems[J]. Applied Mathematics & Computation, 2009, 212(2):505-518.

[13] Datta D, José Rui, Figueira. A real-integer-discrete-coded particle swarm optimization for design problems[J]. Applied Soft Computing, 2011, 11(4):3625-3633.

[14] Datta D, Figueira, José Rui. A real–integer–discrete-coded differential evolution[J]. Applied Soft Computing, 2013, 13(9):3884-3893.

[15] Liu J, Chen B, et al. Torsional stiffness calculation of double-enveloping cycloid drive[J]. Journal of Advanced Mechanical Design, Systems, and Manufacturing, 2012, 6(1): 2-14.

[16] Meng Y , Wu C , Ling L . Mathematical modeling of the transmission performance of 2K–H pin cycloid planetary mechanism[J]. Mechanism & Machine Theory, 2007, 42(7):776-790.

[17] Kaymaz I. Application of kriging method to structural reliability problems[J]. Structural Safety, 2005, 27(2): 133-151.

[18] Nitish Srivastava. Improving Neural Networks with Dropout[D]. Toronto: University of Toronto, 2013.

[19] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[20] Keskar N S, Mudigere D, Nocedal J, et al. On large-batch training for deep learning: Generalization gap and sharp minima[J]. arXiv preprint arXiv:1609.04836, 2016.