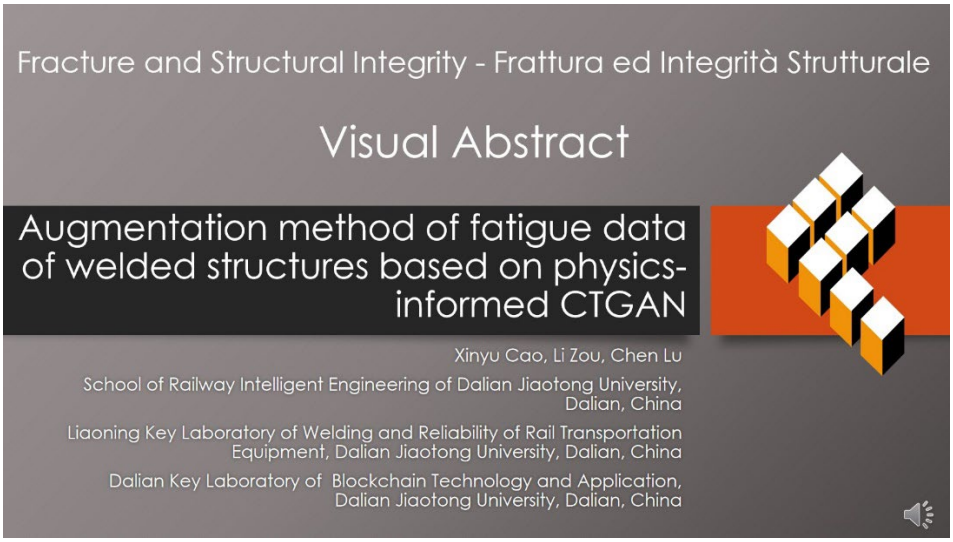
Augmentation method of fatigue data of welded structures based on physics-informed CTGAN

Xinyu Cao
School of Railway Intelligent Engineering of Dalian Jiaotong University, Dalian, China
cxy14157676@163.com

Li Zou*
School of Railway Intelligent Engineering of Dalian Jiaotong University, Dalian, China
Liaoning Key Laboratory of Welding and Reliability of Rail Transportation Equipment, Dalian Jiaotong University, Dalian, China
Dalian Key Laboratory of Blockchain Technology and Application, Dalian Jiaotong University, Dalian, China
lizou@djtu.edu.cn

Chen Lu
School of Railway Intelligent Engineering of Dalian Jiaotong University, Dalian, China
19912060618@163.com



Citation: Cao, X., Zou, L., Lu, C., Augmentation method of fatigue data of welded structures based on physics-informed CTGAN, Frattura ed Integrità Strutturale, 72 (2025) 162-178.

Received: 13.01.2025
Accepted: 19.02.2025
Published: 27.02.2025
Issue: 04.2025

KEYWORDS. Fatigue life prediction, Data augmentation, CTGAN, Variable amplitude loading.

INTRODUCTION

In modern engineering practice, welded structures in aerospace, automotive industry, bridge construction and other fields typically perform under variable amplitude loading and are very susceptible to fatigue failure [1]. Fatigue damage increases cumulatively during the service of welded structures until the structure fails by fracture. The randomness and complexity in variable amplitude loading elevate the challenges associated with predicting fatigue life. Consequently, accurate prediction of fatigue life under variable amplitude loads is crucial for ensuring the project's safe operation and mitigating financial losses.

Up to now, two primary categories have been established for methods predicting fatigue life under variable amplitude loading: cumulative damage theory [2] and machine learning methods [3]. Among them, the theory of fatigue accumulation damage is categorized into two principal divisions: namely, the linear theory of damage accumulation and the nonlinear theory of damage accumulation. The most common Miner linear cumulative damage model [4] is applied extensively in engineering as it is easy to calculate. However, it overlooks the impacts of loading sequence and loading effects in its consideration. There is often a significant discrepancy between the actual damage and the predicted damage of welded structures under variable amplitude loading. Therefore, experts and scholars have proposed nonlinear fatigue damage accumulation theories rooted in damage curves, loading interactions, continuum damage mechanics, energy-based approaches, and physical property degradation to address this issue. For example, Ye et al. [5] proposed a nonlinear fatigue damage accumulation model, referred to as the Ye model, which is based on the dynamic degradation of material toughness resulting from fatigue-induced damage. This model is prevalent in engineering applications due to its straightforward form and concise physical explanation. Nonetheless, the model overlooks the interaction between loads, leaving ample room for enhancing the precision of life prediction. Several scholars have improved the model to address this issue. Lv et al. [6] integrated the influence of loading interaction into the Ye model through the introduction of a two-level loading ratio. Wang et al. [7] demonstrated the impact of load interactions on fatigue damage progression by factoring in the square of the load ratio between successive loading stages, thereby improving prediction precision. Peng et al. [8] considered both the influence of loading sequence and the interplay between two loads when assessing residual life.

The fatigue damage models discussed above are based on specific physical mechanisms, yet they generally do not account for the uncertainties arising from various influencing factors during the fatigue analysis of welded structures [9]. Thus, methods of machine learning have been employed. For instance, Gan et al. [10] applied a data-driven model, grounded in the Kernel Extreme Learning Machine (KELM), to predict the residual lifespan of welded materials subjected to two-level loading conditions. The model autonomously learns the best correlation from the training samples, effectively describing the fatigue damage mechanism. Liu et al. [11] utilized three algorithmic frameworks—specifically, Random Forest (RF), ExtremeGradient Boosting (XGBoost), and Gradient Boosting Machines (GBM)—for forecasting the fatigue life of high-strength steels. Among these, the utilizing gradient boosting demonstrated the highest precision in estimating the fatigue lifetime of high-strength steels under extremely high cycle conditions. Matin et al. [12] used multiple machine learning algorithms to evaluate the factors influencing piston aluminum alloy specimens and the interactions that affect fatigue life values. It took into account the effect of various inputs on fatigue life. Zou et al. [13] established a method for predicting the fatigue life of welded joints, leveraging the whale optimization algorithm alongside the Support Vector Machine (SVM). It took into account the sequence and interactions of loads to estimate the fatigue life under conditions of two-level loading. However, the accuracy of fatigue life predictions made by machine learning models is often hampered by the scarcity of available fatigue samples. Obtaining a sufficient number of fatigue samples is challenging due to the large number and difficulty of various conditions required for fatigue testing, leading to inadequate accuracy in predicting fatigue life under variable loading.

Data augmentation is a method to increase the amount and improve the quality of data by transforming, augmenting or synthesizing the original data. Data augmentation methods such as Generative Adversarial Networks (GAN) [14] have emerged. GAN is a form of deep learning model, proposed by Goodfellow in 2014, with the ability to solve problems associated with limited sample sizes. Since GAN is proposed, many variants have emerged, which are widely used in fields such as computer vision, medicine, and natural language processing. In the realm of fatigue life analysis and prediction, their utilization is still in its nascent phases. For example, He et al. [15] utilized data produced by a table GAN within a machine learning framework for predicting multiaxial fatigue life. The inclusion of synthetic data enhanced the predictive capacity of the machine learning models in estimating life expectancy, the findings suggested. Sun et al. [16] employed a cyclical GAN to augment a dataset of 20 multiaxial fatigue data points, expanding it into thousands of comparable samples. This well balances the time cost of large sample sizes with prediction accuracy. He et al. [17] introduced a deep learning architecture that combines a GAN with a physical model for the purpose of predicting multiaxial fatigue life. When equipped with

suitable physical constraints, the model outperforms the neural network in terms of prediction accuracy. Primarily, these models serve the purpose of predicting multiaxial fatigue life under conditions of constant amplitude loading. However, welded structures frequently encounter variable amplitude loading scenarios in practical engineering applications, resulting in fatigue data that exhibits complexity and dispersion. Therefore, it is still a problem in fatigue research applications to enhance effective training samples under the variable amplitude loading.

This work proposes a novel augmented model for fatigue data of welded structures subjected to two-step loading that integrates physical mechanisms. The cumulative damage model-Peng model is integrated into the CTGAN as a physical loss, enabling the generated fatigue data to adhere to the relevant physical mechanisms. The validity of the augmented model is confirmed through testing on machine learning models. The problem of insufficient fatigue data for residual fatigue life prediction under two-step loading is solved. The accuracy of the machine learning models for fatigue life prediction of welded structures is further improved.

BASIC THEORY

Ye and its modified model

From a macro-physics standpoint, the process of fatigue damage accumulation can be interpreted as a progressive degradation and decline in the structural properties. Therefore, the alterations in material's macroscopic characteristics can be considered as damage variables to measure fatigue damage. Ye et al. [5] discovered through extensive fatigue testing that the most significant change in the material's fatigue damage history is its toughness. Consequently, a nonlinear cumulative damage model was introduced, emphasizing the dissipation of material toughness. The fatigue damage evolution equation in Ye model is shown as Eqn.(1):

$$D_N \approx -\frac{\ln\left(1-\frac{n}{N_f}\right)}{\ln\left(N_f\right)} \tag{1}$$

where N_f denotes the fatigue life under stress σ, n represents the number of cycles under stress σ, and D_N signifies the cumulative damage variable after n cycles of stress σ.

Fig 1. displays the fatigue damage curves obtained under two-step loading conditions.

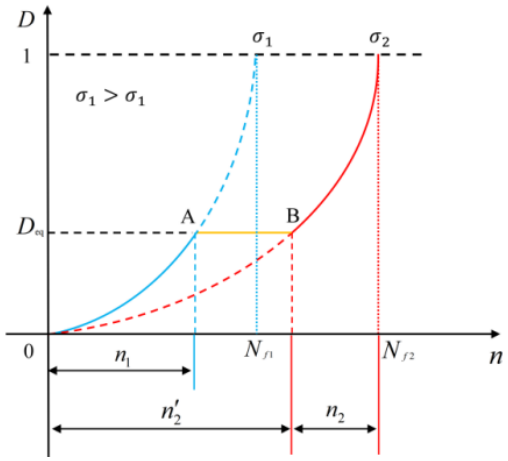


Figure 1: Fatigue damage curves under two-step loading.

Based on the Ye model, after n_1 cycles of σ_1, the fatigue cumulative damage value is:

$$D_A = -\frac{\ln\left(1-\frac{n_1}{N_{f1}}\right)}{\ln\left(N_{f1}\right)} \tag{2}$$

It is equivalent to the damage incurred by n_2' cycles of σ_2 as illustrated in Eqn.(3).

$$D_B = -\frac{\ln\left(1-\frac{n_2'}{N_{f2}}\right)}{\ln\left(N_{f2}\right)} \tag{3}$$

According to the principle of equivalent damage, it can be obtained that the corresponded damage at A and B is the same, i.e.

$$D_A = D_B \tag{4}$$

By substituting Eqn. (2) and Eqn. (3) into Eqn. (4), we obtain:

$$\frac{n_2'}{N_{f2}} = 1-\left(1-\frac{n_1}{N_{f1}}\right)^{\frac{\ln\left(N_{f2}\right)}{\ln\left(N_{f1}\right)}} \tag{5}$$

Typically, fatigue damage arises when the cumulative damage *D* attains a critical limit, which is defined as 1 in the Ye model. Currently, the formulation for cumulative damage criterion is presented in Eqn. (6).

$$\frac{n_2'}{N_{f2}} + \frac{n_2}{N_{f2}} = 1-\left(1-\frac{n_1}{N_{f1}}\right)^{\frac{\ln\left(N_{f2}\right)}{\ln\left(N_{f1}\right)}} + \frac{n_2}{N_{f2}} = 1 \tag{6}$$

Thus the predicted value of the remaining life by using Ye model for two-step loading could be reached as shown in Eqn. (7):

$$\frac{n_2}{N_{f2}} = \left(1-\frac{n_1}{N_{f1}}\right)^{\frac{\ln\left(N_{f2}\right)}{\ln\left(N_{f1}\right)}} \tag{7}$$

Despite the simplicity of its form and the clear physical meaning of the Ye model, it fails to reflect the influences of load interactions under conditions of variable amplitude loading. Stress ratios that describe load interactions were used by many existing models in nonlinear cumulative damage theory. Peng et al [8] proposed the improved Ye model, accounting for both the sequence of applied loads and the mutual influence between different loads on residual life. Eqn. (8) delineates the estimated fatigue life remaining under two-step loading, according to Peng's model.

$$\frac{n_2}{N_{f2}} = \left(\frac{1}{N_{f2}}\right)^{\left(\frac{\ln\left(1-\frac{n_1}{N_{f1}}\right)}{\ln\left(N_{f1}\right)}\right)^{\frac{\sigma_2}{\sigma_1}}} \tag{8}$$

In summary, the cumulative damage models generally have a clear and explicit physical definition. However, due to the insufficient consideration of uncertainty of fatigue life influencing factors and the complexity of the formula, its application in engineering is limited. Traditional cumulative damage models' limitations are increasingly being tackled effectively through the utilization of machine learning techniques now [18]. Due to the limited number of fatigue test samples, obtaining high-precision prediction models under small sample conditions is a bottleneck problem.

CTGAN

A generative model known as GAN comprises two neural networks: a generator and a discriminator, engaged in a competitive learning dynamic. The optimization objective function of the GAN is presented in Eqn. (9):

$$\min_G \max_D V(G,D) = E_{x\sim p_{data}(x)}\left[\log D(x)\right] + E_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \tag{9}$$

where x stands for the real data, z signifies the potential noise, $E(*)$ indicates the expected value of the distribution function, $p_{data}(x)$ embodies the distribution of authentic samples, $p_z(z)$ represents the distribution of the noise defined in the lower dimension, $D(x)$ denotes the result of the judgement of the discriminator on the real data x , and $G(z)$ denotes the fake data that was generated based on the random data z in the generator. The discriminator D needs to match the real samples as much as possible, i.e., maximize $\log D(x)$, while the generator G needs to maximize the loss of D, i.e., minimize $\log D(x)$. Ultimately, a Nash equilibrium is reached between the generator and the discriminator, allowing the generator to create realistic data.

Conditional Tabular GAN (CTGAN) [19] uses a GAN-based approach to model samples from tabular data distribution. CTGAN employs a Variational Gaussian Mixture model (VGM) for each continuous variable, with the intention of determining the most suitable k gaussian models to depict the data through the application of the expectation maximization algorithm. Additionally, it compels the generator to produce samples with discrete variable distributions that closely resemble the training data and incorporates a condition vector as part of the input. The input to CTGAN comprises a condition vector, which direct the generator to create samples that belong to designated categories. The condition vector, which is encoded in one-hot format to represent all discrete columns, selects conditions by sampling from the training dataset. The generator's loss function ensures that the samples produced by the generator fulfill the specified condition. Incorporating the cross-entropy between condition vector and generated samples within loss function accomplishes this. In this work, CTGAN is employed for data augmentation to tackle the challenge posed by limited fatigue data.

FATIGUE LIFE PREDICTION METHOD BASED ON DATA AUGMENTATION

Fatigue data generation method based on physics-informed generative adversarial networks

Investigating fatigue performance typically necessitates a substantial number of repeated tests. However, it is currently difficult to obtain a large number of training samples due to the complexity and randomness of fatigue testing. The shortage of samples in the training dataset affects both the precision and the ability of the model to generalize. This work proposes a CTGAN generative model based on physics-informed to solve the problem of fewer training samples under two-step loading. This enables the machine learning models to effectively capture the relationship between inputs and outputs.

In this work, five fatigue life prediction models, Miner law [4], Ye model [5] and its improved model LV model [6], Wang model [7], Peng model [8], are selected to predict fatigue life under two-step loading for five welded materials [20-24]. Specific details related to the literature dataset here can be found in section Experimental results and analysis. The mean absolute percentage error (*MAPE*) quantifies the precision of life prediction, as defined by the equation below:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i' - y_i\right| \times 100 \tag{10}$$

The predicted fatigue life for the i-th sample is denoted as y_i' , while the actual fatigue life for the same sample is y_i . Tab. 1 presents the average absolute percentage error for the five materials.

Materials	Miner model [4] *MAPE* (%)	Ye model [5] *MAPE* (%)	LV model [6] *MAPE* (%)	Wang model [7] *MAPE* (%)	Peng model [8] *MAPE* (%)
Butt joint	37.18	35.12	26.35	18.76	16.48
Corner joint	24.86	22.16	14.43	16.42	20.84
Al-2024-T42	73.10	65.75	41.27	26.80	19.74
Al-7070-T7451	67.64	66.15	59.67	60.26	54.29
Ti-6Al-4V	92.08	84.17	68.43	53.18	46.85

Table 1: MAPE values of five traditional cumulative damage models.

According to Tab. 1, the Peng model demonstrates a notably smaller prediction error compared to the other models for materials such as butt joint, Al-2024-T42, Al-7075-T7451, and Ti-6Al-4V. The Peng model exhibits a marginally higher error on corner joint as opposed to the LV and Wang models compared to other models. In general, it seems that the Peng model has the smallest prediction error. In this work, the Peng model is selected as the physical loss component to be integrated into the generator loss function of CTGAN. The physical loss component serves as a regularizer within the network, aiding in the improvement of the training procedure. The generator is enabled by the physical loss term to produce data that adheres to physical constraints. The generator's overall loss function in this work is formulated as follows:

$$L_{generator} = L_{adversarial} + L_{discrete} + L_{mmd} + L_{physics} \tag{11}$$

where $L_{adversarial}$ is the adversarial loss function, $L_{discrete}$ is the discrete column loss function, L_{mmd} is the Max Mean Discrepancy (MMD) loss function [25] and $L_{physics}$ is the physics loss function. The adversarial loss function and discrete column loss function in CTGAN are shown below:

$$L_{adversarial} = -\frac{1}{N}\sum_{i=1}^{N}\log D\left(G\left(x_i\right)\right) \tag{12}$$

$$L_{discrete} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log\left(y'_{i,k}\right) \tag{13}$$

where N denotes the quantity of data, x_i signifies the input data, $G\left(x_i\right)$ represents the synthetic data generated based on the input, and $D\left(G\left(x_i\right)\right)$ denotes the discriminator's probability of deeming the synthetic data as authentic. K indicates the count of potential values for the discrete column, $y_{i,k}$ signifies the actual probability of the kth value in the discrete column for the i-th sample, and $y'_{i,k}$ denotes the probability of the same kth value generated by the generator for the i-th sample. The generator's adversarial loss function primarily strives to decrease the discriminator's likelihood of classifying the generated sample as fake. In essence, it aims to elevate the probability that discriminator will deem generated sample as authentic. In CTGAN, the loss function for discrete columns typically utilizes cross-entropy to diminish the discrepancy between the probability distribution of the generated discrete data and that of the real data. It ensures a tight correspondence between the probability distribution of the authentic data and that of the synthetic discrete data.
This work introduces two additional loss functions to the base CTGAN: the MMD loss function and the physical loss function. Below is the representation of MMD loss function:

$$L_{mmd}\left(P,Q\right) = \left\| \frac{1}{N}\sum_{i=1}^{N}\phi\left(y_i\right) - \frac{1}{M}\sum_{j=1}^{M}\phi\left(y_j\right) \right\|^2 \tag{14}$$

where P denotes the probability distribution of the data produced by the generator and Q represents the probability distribution of the real data. N and M are the number of samples obtained by sampling from P and Q, respectively. y_i

and y_j are the samples obtained by sampling from P and Q, respectively. $\phi(*)$ serves as a feature mapping function, transforming the samples into a higher-dimensional feature space. MMD aims to minimize the distance separating two probability distributions, ensuring that generator's sample distribution closely approximates genuine sample distribution. Often, it is integrated into GAN as a component of the generator's loss function, with the goal of decreasing the discrepancy between generated and true values.

The physical loss function is derived from the Peng fatigue life prediction model and is applied to this loss component as detailed below:

$$L_{physics}=\left(\left(\frac{1}{N_{f2}}\right)^{\left(\frac{\ln\left(1-\frac{n_1}{N_{f1}}\right)}{\ln\left(N_{f1}\right)}\right)^{\frac{\sigma_2}{\sigma_1}}}\right)\cdot N_{f2} \tag{15}$$

where σ_1 denotes the first stress level, σ_2 denotes the second stress level, N_{f1} is the fatigue life under σ_1, N_{f2} is the fatigue life under σ_2, and n_1 is the number of cycles under the first level of stress. Incorporating physical loss function allows the generated fatigue data to adhere to physical laws, resulting in generated residual fatigue life values that are closer to the real data n_2. It also reflects the fatigue behavior features and the prior training process is more stable.

In CTGAN, the discriminator's loss function is usually based on the adversarial training principle. The cross-entropy loss function is commonly employed, and it is expressed as follows:

$$L_{discri\min ator}=-\frac{1}{N}\sum_{i=1}^{N}\left[label_i\log\left(D\left(x_i'\right)\right)+\left(1-label_i\right)\log\left(1-D\left(x_i'\right)\right)\right] \tag{16}$$

where, x_i' is the input generated data, $label_i$ is the real label, the real data takes the value of 1, the generated fake data takes the value of 0, and $D\left(x_i'\right)$ is the probability of judging that the data is real data. This loss function boosts the discriminator's capacity to differentiate accurately between genuine and synthesized samples. At the same time, it minimizes the probability of incorrectly classifying a generated sample as real.

In contrast to original GAN, the input for CTGAN generator in this work consists of real experimental data, instead of random variables. The physics loss $L_{physics}$ is combined with the MMD loss L_{mmd} to make the generated data by the generator closer to the real values, adhering to the physical principles governing fatigue data under variable amplitude loading during training phase. Fatigue data exhibits high discreteness. CTGAN's initial discrete column loss $L_{discrete}$ guarantees that the synthesized fatigue data corresponds to the real data distribution under specified conditions, enhancing the generator network's efficiency and elevating the quality of the generated fatigue data.

Fatigue life prediction model based on machine learning models

A fatigue data augmentation model for welded structures fused with physical mechanism is developed to target the problem of less fatigue data in machine learning models under two-step loading. The proposed model's overall architecture is illustrated in Fig. 2. This model generates data that reflects fatigue behavior under variable amplitude loading, making the generated fatigue data consistent with the physical results. It can be seen that the overall experimental process is divided into four main parts. First we get the fatigue data. The data attributes included are specifically shown therein. Then the data is input into the CTGAN model. And it has been validated effectively on four machine learning models KELM, SVM, RF, and Back Propagation (BP), respectively. Finally its effect on fatigue prediction accuracy is evaluated by two indicators.

The specific steps of the framework are outlined as follows:

Step1: Gather literature and laboratory test data on welded structures subjected to variable amplitude loading to create a fatigue dataset for these structures. The attributes of the fatigue dataset are: σ_1 and σ_2 are the first and second stress levels,

respectively, N_{f1} and N_{f2} are the fatigue life at the first and second stress levels, respectively, n_1 is the number of cycles under σ_1 , and n_{2p} is the number of cycles under σ_2, namely, the residual fatigue life.

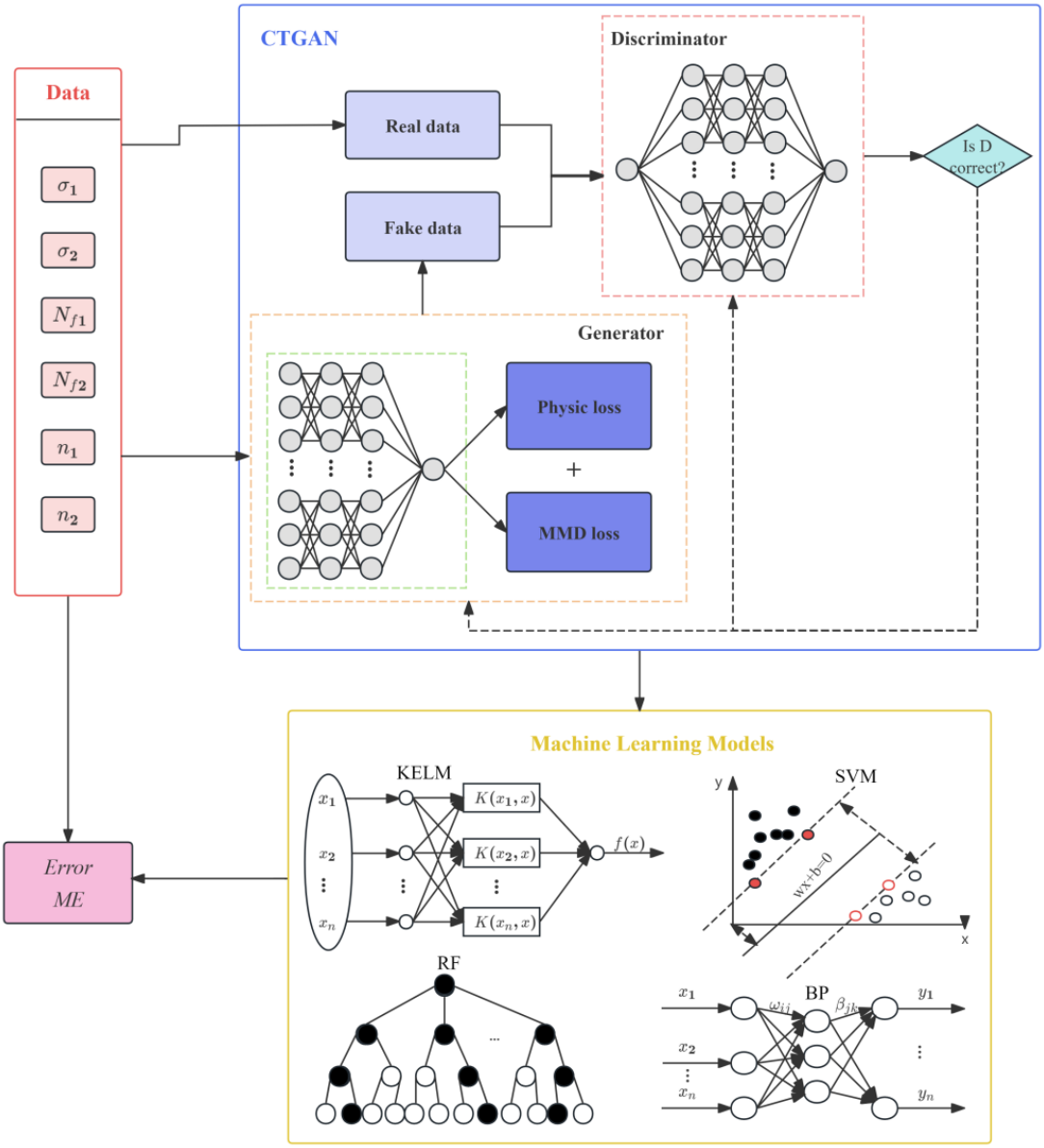


Figure 2: The whole experimental framework for fatigue life prediction of welded structures under two-step loading based on physics-informed CTGAN.

Step2: The CTGAN model was used to data augmentation based on the original welded structure fatigue dataset that has been built. The fatigue attributes σ_1, σ_2, n_1, N_{f1}, N_{f2}, n_{2p} are input into the CTGAN model. The trained model finally obtains fatigue data with outputs σ_{1m}, σ_{2m}, n_{1m}, n_{2m}. Where σ_{1m} and σ_{2m} are the generated first and second stress levels, respectively, n_{1m} is the generated number of cycles under σ_{1m} , and n_{2m} is the generated number of cycles under σ_{2m}. Filtering in fatigue data from σ_1, σ_2, n_{1m}, N_{f1}, N_{f2}, n_{2m} and the data were normalized.

Step3: The machine learning models were trained using the normalized data values σ_1, σ_2, n_{1m}, N_{f1}, N_{f2}, n_{2m} as inputs, with the normalized fatigue life n_{2m} serving as the output. Subsequent training sessions were then conducted for these models.

Step4: Experiments were conducted on fatigue data from the test set using machine learning models obtained after data augmentation and original data training, respectively. The work examined how the generated samples influenced the predictive accuracy of the machine learning models. Furthermore, the proposed model's performance was validated by comparing it with other conventional physical models, namely the Miner law, the Ye model, and its improved model, the Peng model.

Within the four outlined steps, Step 2 encompasses the following two specific subprocesses:

Step2.1: Fatigue data at the same stress level as the original dataset were chosen from the output of the CTGAN generation model that is when σ_{1m} equals σ_1 and σ_{2m} equals σ_2. And insert the corresponding N_{f1} and N_{f2} in the four columns of the output data which are the fatigue life under the first stage load and the second stage load. Finally, the complete data from σ_1, σ_2, n_{1m}, N_{f1}, N_{f2}, n_{2m} is obtained.

Step2.2: The welded structures' fatigue dataset was split into training and test sets. A part of the fatigue test data was selected as the test set and the remaining data was used as the training set. Following this, the validated augmented data was combined with original training data to create a new training sample set. The data was at the same time normalized and the normalization was calculated by the formula:

$$x_{new} = \frac{x - x_{\max}}{x_{\max} - x_{\min}} \tag{17}$$

where x denotes the pre-normalization value of the sample data, $x_{\max}$ and $x_{\min}$ represent the maximum and minimum values of the fatigue test samples for a given data attribute, respectively, and x_{new} is the value of the sample data after normalization.

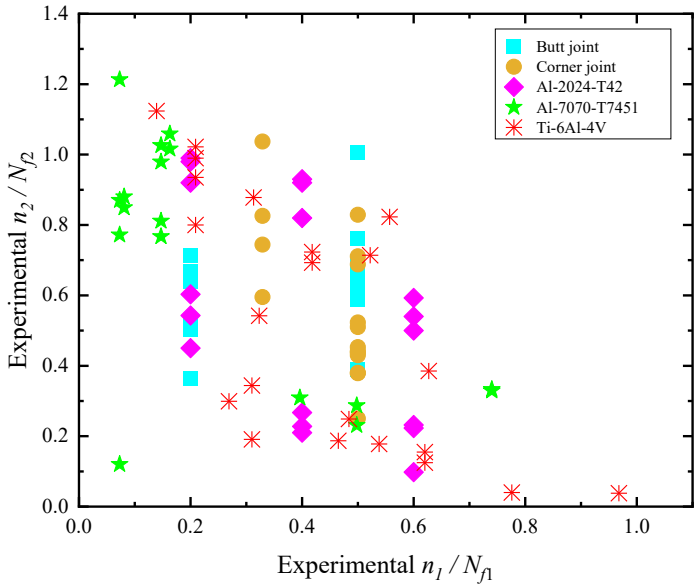


Figure3: Fatigue sample data.

EXPERIMENTAL RESULTS AND ANALYSIS

Fatigue data

In this work, two welded joints and three welded materials are selected for experiments on the proposed framework, aluminum alloy butt joint [20,21] and corner joint [20,21], Al-2024-T42 [22], Al-7070-T7451 [23], and Ti-6Al-4V [24] respectively. These welded joints and materials are widely utilized in aerospace, marine, automotive, and various other

industrial sectors. All experimental materials were tested for fatigue in a stress-controlled mode at room temperature. Additional information regarding the other test data is available in the accompanying references. The distribution of data from five types of experimental data is represented in terms of loading cycle ratios in Fig. 3.

Parameter settings

In this work, all experiments were made on a personal computer with Windows 10 (64bit) operating system, hardware platform parameters of Intel(R) Core (TM) i5-7300HQ CPU, 2.50GHz main frequency and 16G RAM. The CTGAN model employs the Adam optimization algorithm for training the neural network. The discrete feature embedding dimension is set to 128, while the hidden layer dimensions for both the generator and discriminator are (256, 256). The learning rates for the generator and discriminator are both set to 2e-4 , with a learning rate decay of 1e-6 for both. The batch size during training is 600, and the number of epochs is 2000. 300 fatigue data for each material is generated. The relevant parameters of the machine learning models are presented in Tab. 2:

Machine learning models	Parameters setting
KELM	optimal regularization coefficient C: 20, kernel function parameter S: 1, kernel function: rbf
SVM	penalty factor c: 4.0, radial basis function parameter g: 0.8, kernel function: rbf
RF	number of decision trees t: 100, minimum number of leaves l: 5
BP	learning rate： 0.01, error threshold: 1e-6

Table 2: Parameter setting of machine learning models.

In this work, a part of aluminum butt and corner joints, Al-2024-T42, Al-7075-T7451 and Ti-6Al-4V are selected as the test set. A part of the data of the test set is shown in Tab. 3. The rest of the data is the original training set. The generated fatigue dataset is then integrated with the original training set to produce the augmented training set.

Materials	σ_1 / *MPa*	σ_2 / *MPa*	n_1 / *cycles*	N_{f1} / *cycles*	N_{f2} / *cycles*	n_2 / *cycles*
Butt joint	104	74	109900	549300	1540100	795800
Butt joint	74	89	770100	1540100	880500	581400
Corner joint	93	73	309900	619800	1546100	386120
Corner joint	73	83	509200	1546100	952300	708200
Al-2024-T42	200	150	30000	150000	430000	233400
Al-2024-T42	150	200	258000	430000	150000	89000
Al-7070-T7451	176	133	2000	27300	61400	47400
Al-7070-T7451	176	85	2000	27300	225800	27100
Al-7070-T7451	133	85	5000	61400	225800	198600
Ti-6Al-4V	647	517	18000	37200	143633	35700
Ti-6Al-4V	595	517	40000	64467	143633	22300
Ti-6Al-4V	517	595	30000	143633	64467	63800

Table 3: A part of the test set.

Accuracy evaluation

The evaluation indicators in this work are absolute percentage error *Error* and mean absolute percentage error *ME* to assess the performance of the model. Its calculation formula is:

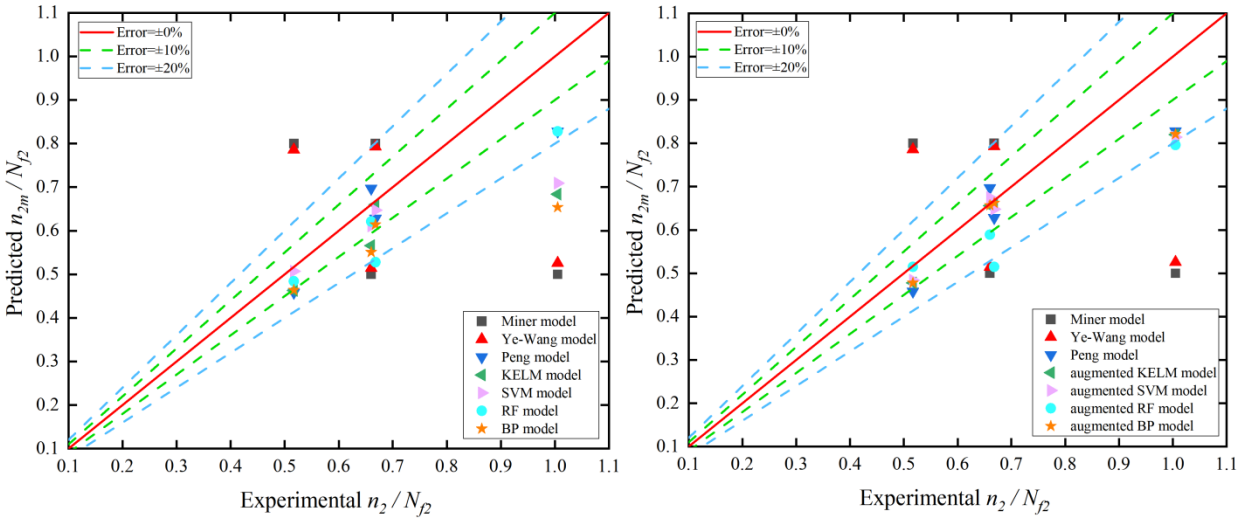
$$Error = \frac{\left| \frac{n_{2m}}{N_{f2}} - \frac{n_2}{N_{f2}} \right|}{\frac{n_2}{N_{f2}}} \times 100\% \tag{18}$$

$$ME = \frac{1}{m}\sum_{m=1}^{m} Error \tag{19}$$

where $\frac{n_2}{N_{f2}}$ is the experimental cycle ratio under stress σ_2, $\frac{n_{2m}}{N_{f2}}$ is the predicted cycle ratio under stress σ_2, and m is the sample size of the dataset. The smaller the values of *Error* and *ME* mean that the model performs better.

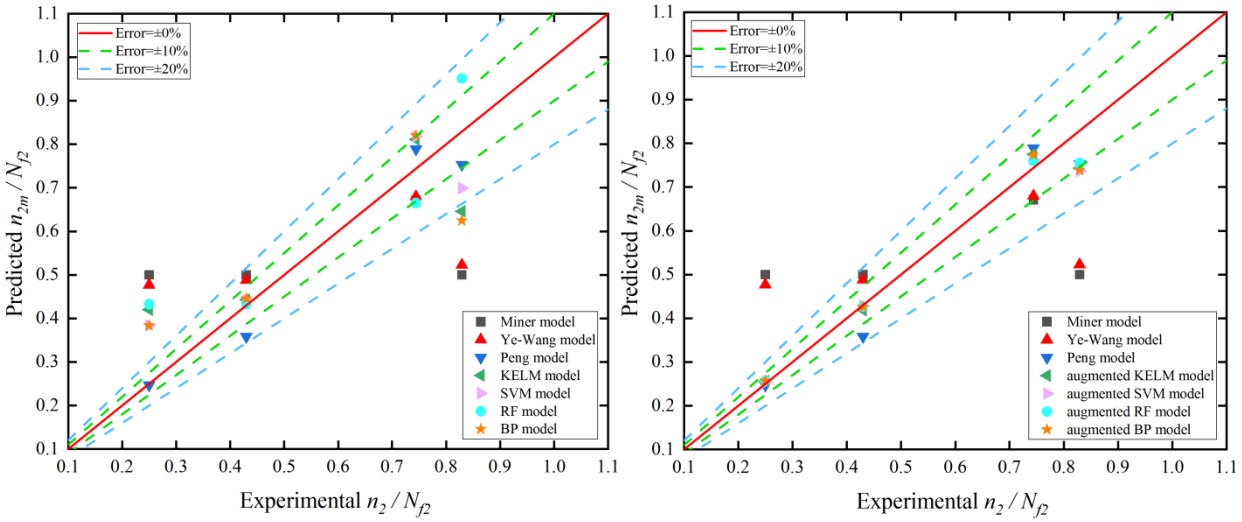
Aluminum alloy welded joints

The experimental data utilized for the tests comprised the fatigue test results of welded joints from the aluminum alloy bodies of high-speed trains in papers [20,21]. The welded joints used in the test contain two types of welding: butt joint and corner joint. The samples' base material is ENAW6005, an aluminum alloy employed in the construction of CRH2 high-speed train bodies. Based on fatigue testing outcomes, the butt joint fatigue life of 549,300, 880,500, and 1,540,100 cycles at stress levels of 104, 89, and 74 MPa, respectively. Similarly, the corner joint's fatigue life were 619,800, 952,300, and 1,546,100 cycles at stress levels of 93, 83, and 73 MPa, respectively. The error between experimental and predicted values for butt joint and corner joint are shown in Fig 4 and 5.



(A) Butt joint without data augmentation (B) Butt joint with data augmentation

Figure 4: Comparison of predicted and experimental cycle ratios for aluminum butt joint.



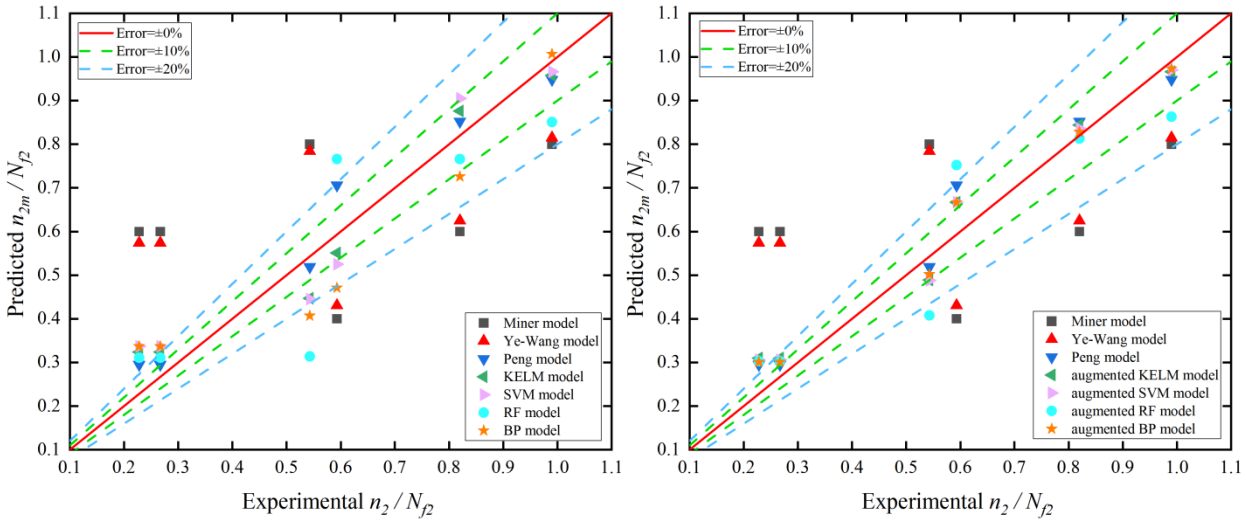
(A) Corner joint without data augmentation (B) Corner joint with data augmentation

Figure 5: Comparison of predicted and experimental cycle ratios for aluminum corner joint.

As observed in the predicted results of the welded joints presented in Figs. 4 and 5. For the butt joint, most of the four machine learning models utilizing data augmentation predicted points within the 10% error band, with three points nearly aligning with the 0% error band. In contrast, the majority of predictions made by the Miner and Ye models falls outside the 20% error band. For corner joint, four machine learning models predictions with data augmentation are comparatively closer to the 0% error band. The machine learning models exhibited more consistent predictive performance compared to traditional models, displaying minimal variance in their results.

Aluminum alloy materials Al-2024-T42

The Al-2024-T42 material [22] offers advantages such as high strength and excellent temperature tolerance. It is used to manufacturing a variety of components that hold high loads and is mainly used in aerospace applications. Fully reversed fatigue loads of different amplitudes were applied to the polished thin plate samples, setting the experimental frequency at 25 Hz and the stress ratio R=-1. The results of fatigue testing indicate that, under an applied stress of 150 MPa, Al-2024-T42 has a fatigue life of 430,000 cycles, whereas at 200 MPa, its fatigue life is 150,000 cycles. The error between the experimental and predicted values of Al-2024-T42 is shown in Fig 6.



(A) Al-2024-T42 without data augmentation (B) Al-2024-T42 with data augmentation

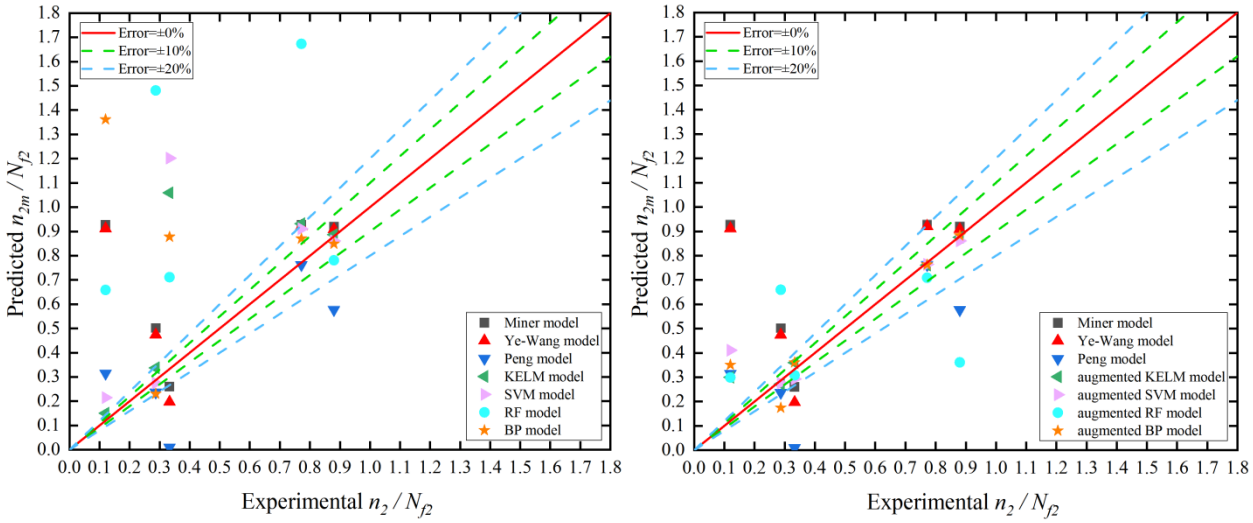
Figure 6: Comparison of predicted and experimental cycle ratios for Al-2024-T42.

As shown in the prediction results in Fig. 6, most of the predicted values from the four machine learning models utilizing data augmentation are concentrated within the 20% error band. Conversely, the majority of predictions from the Miner and Ye models lie beyond the 20% error band. And there are five points in the augmented machine learning models that are closer to the fatigue experiment results and better stability relative to the Peng model.

Aluminum alloy materials Al-7050-T7451

Al-7050-T7451 is a high strength aluminum alloy material. It has good corrosion resistance and process-ability and is widely used in aerospace and automotive industries. The Al-7050-T7451 material from the paper [23] was tested for single stress amplitude change in flexural fatigue at room temperature. The results of fatigue testing reveal that, when subjected to applied stresses of 176 MPa, 133 MPa, and 85 MPa, the fatigue life corresponds to 27,300 cycles, 61,400 cycles, and 225,800 cycles, respectively. The error between the experimental and predicted values of Al-7050-T7451 is shown in Fig 7.

The prediction results in Fig. 7 show that most of the predicted values from the four machine learning models using data augmentation are concentrated within the 10% error band. Conversely, the majority of the predicted values from both the Miner and Ye models lie outside and are distant from the 20% error band. Furthermore, it is evident that the KELM and SVM models demonstrate superior prediction performance compared to the RF and BP models among the four machine learning approaches.

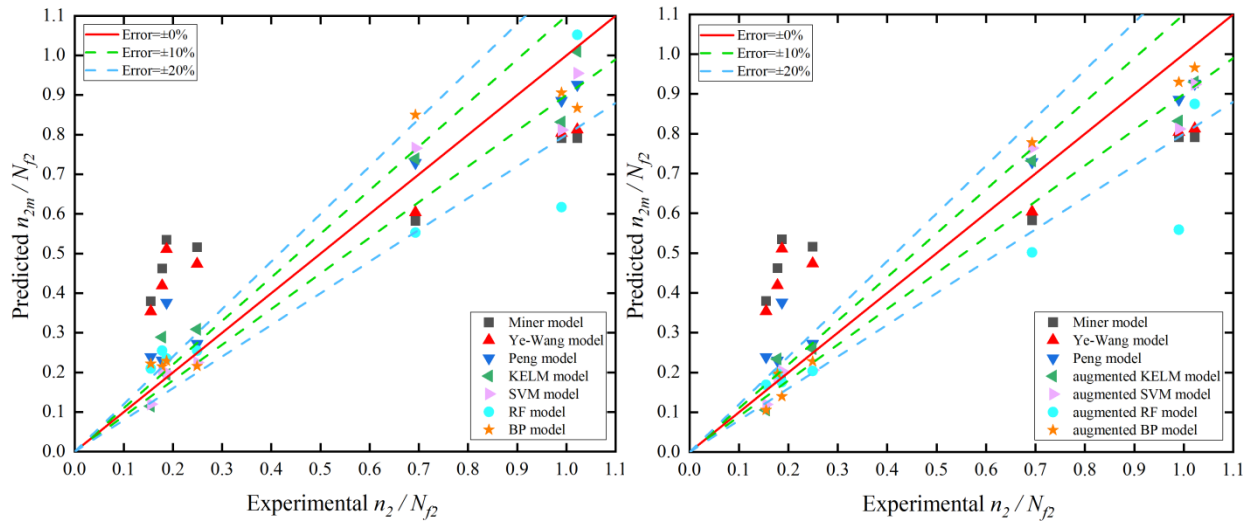


(A) Al-7050-T7451 without data augmentation (B) Al-7050-T7451 with data augmentation

Figure 7: Comparison of predicted and experimental cycle ratios for Al-7050-T7451.

Titanium alloy materials Ti-6Al-4V

Ti-6Al-4V is the pioneering titanium alloy material that has been successfully developed and implemented. Its exceptional heat and corrosion resistance qualities make it predominantly utilized in aviation engines, rockets, and various other industries. The aero-engine compressor blade material Ti-6Al-4V titanium alloy from the literature [24] was tested in a room temperature environment. Its variable amplitude loading fatigue test consists of two types of loading: high-to-low and low-to-high. The load levels for the high-low loading were 595-517 MPa and 647-517 MPa, while for the low-high loading, they were 517-595 MPa and 517-647 MPa, respectively. Where the fatigue life at 647, 517, and 595 MPa stresses were 37200, 143633, and 64467 cycles respectively. Fig. 8 displays the error between the experimental and predicted values for Ti-6Al-4V.



(A) Ti-6Al-4V without data augmentation (B) Ti-6Al-4V with data augmentation

Figure 8: Comparison of predicted and experimental cycle ratios for Ti-6Al-4V.

As shown in the prediction results in Fig. 8, the majority of the predicted values from the four machine learning models utilizing data augmentation fall within the 20% error band. Four of the data are closer to being within the 10% error band. Alternatively, the majority of the forecasts generated by the Miner and Ye models fall outside the 20% error band and are notably remote from it. It is apparent that the model augmented with data demonstrates a reduced error rate. The proposed augmented model is pertinent for predicting the fatigue life of both welded aluminum alloy structures and welded structures made from various other alloy materials.

Materials	Miner model *ME* (%)	Ye model *ME* (%)	Peng model *ME* (%)	KELM model *ME* (%)	SVM model *ME* (%)	RF model *ME* (%)	BP model *ME* (%)
Butt joint	37.26	35.16	10.13	14.47	10.47	12.74	11.47
Corner joint	41.49	37.47	8.34	25.61	20.63	24.81	23.09
Al-2024-T42	68.95	63.26	11.86	16.31	19.38	24.14	22.18
Al-7070-T7451	158.7	157.88	62.41	193.47	218.02	226.98	195.28
Ti-6Al-4V	93.81	82.79	31.23	20.11	18.49	23.99	20.82

Table 4: ME values of machine learning models with unaugmented data on the testing set.

Materials	Miner model *ME* (%)	Ye model *ME* (%)	Peng model *ME* (%)	KELM model *ME* (%)	SVM model *ME* (%)	RF model *ME* (%)	BP model *ME* (%)
Butt joint	37.26	35.16	10.13	6.95	7.68	10.94	7.32
Corner joint	41.49	37.47	8.34	4.9	4.72	4.68	4.68
Al-2024-T42	68.95	63.26	11.86	13.23	12.5	18.59	11.13
Al-7070-T7451	158.7	157.88	62.41	47.54	48.68	72.74	49.27
Ti-6Al-4V	93.81	82.79	31.23	14.35	12.78	18.45	14.24

Table 5: ME values of machine learning models with augmented data on the testing set.

For the four machine learning models with data augmentation, the *ME* for almost all materials are near 10%, with the exception of Al-7070-T7451, which is slightly higher. It is worth noting that the ME values for butt and corner joints are particularly low after data augmentation, with minimum errors of 6.82% and 4.65% on the KELM and RF、BP models, respectively. When compared to the conventional Miner, Ye, and Peng models, as well as the original machine learning models, the machine learning models that incorporate data augmentation exhibit a notable increase in prediction accuracy. It has been improved to some extent for the low fatigue data leading to poor accuracy of machine learning predictive models.

Stability validation of the augmented model on machine learning models

In order to avoid the uncertainty associated with the fatigue test data extracted above, five-fold cross-validation is added to this section. Thereby the stability and generalizability of the data augmentation model is verified on the machine learning models. *ME* is used here as an evaluation indicator to validate the generalization ability of the augmented model on four machine learning models. The results are shown in Tab. 6 and Tab. 7.

Materials	Miner model *ME* (%)	Ye model *ME* (%)	Peng model *ME* (%)	KELM model *ME* (%)	SVM model *ME* (%)	RF model *ME* (%)	BP model *ME* (%)
Butt joint	37.26	35.16	10.13	18.86	18.56	26.92	24.81
Corner joint	41.49	37.47	8.34	18.17	18.88	27.67	25.62
Al-2024-T42	68.95	63.26	11.86	24.51	30.75	42.27	31.09
Al-7070-T7451	158.7	157.88	62.41	96.81	97.72	137.68	102.01
Ti-6Al-4V	93.81	82.79	31.23	45.81	40.96	75.49	44.82

Table 6: ME values of machine learning models with unaugmented data on the testing set.

Materials	Miner model *ME* (%)	Ye model *ME* (%)	Peng model *ME* (%)	KELM model *ME* (%)	SVM model *ME* (%)	RF model *ME* (%)	BP model *ME* (%)
Butt joint	37.26	35.16	10.13	5.11	5.14	5.68	5.19
Corner joint	41.49	37.47	8.34	5.33	5.34	5.38	5.43
Al-2024-T42	68.95	63.26	11.86	5.74	5.32	13.76	5.82
Al-7070-T7451	158.7	157.88	62.41	10.32	7.66	18.21	10.66
Ti-6Al-4V	93.81	82.79	31.23	7.03	6.91	12.71	9.55

Table 7: ME values of machine learning models with augmented data on the testing set.

As can be seen from Tab. 6 and Tab. 7, the ME on the five-fold cross-validated machine learning models with augmented data overall less than 10%, and the errors are greatly reduced. And the KELM and SVM models have better prediction results relative to the RF and BP models. The data-augmented machine learning models are still more advantaged compared to the traditional Miner, Ye and Peng models. Comparing with Tab. 5, the ME values in Tab. 7 are further reduced in general. The five-fold cross-validation reduces the randomness of model performance evaluation and provides more reliable performance estimates. The results show the stability and generalization ability of data-augmented machine learning models.

CONCLUSION

Limited fatigue data frequently impacts the precision and universal applicability of machine learning models for life prediction. To tackle this challenge, this work introduces a physics-based Generative Adversarial Networks (GAN) model aimed at generating fatigue data under two-step loading. The generated data served as input for machine learning algorithms, enabling predictions of the fatigue life of two types of welded joints and three welded materials under variable amplitude loading conditions.

The model combines a traditional model with machine learning models to better characterize fatigue behavior relative to a simple GAN model. The life prediction Peng model is integrated within the loss function of Conditional Tabular GAN (CTGAN) to guarantee that the generated data conforms to the physical relationships between stress and life. The final valid data that aligns with the characteristics of the original dataset is obtained through a careful selection process. The Peng model can consider the load sequence and the interaction between loads, which makes generated data meet the characteristics of fatigue data under two-step loading. Meanwhile, it effectively solves the limitation that machine learning models rely on large samples. The experimental findings indicate that the generated data notably augment the models' predictive accuracy.

The experiments are validated on two welded joints and three welded materials. The prediction indicators absolute percentage error *Error* and mean absolute percentage error *ME* decreased obviously for each material. The *ME* values of both welded joints decreased to less than 10%, and the *ME* values of titanium alloy materials also decreased by almost 10% on average. The results show that it is not only suitable for aluminum alloy materials, but also apparently effective for titanium alloy materials. In comparison to the traditional Miner model, Ye model, and Peng model, the augmented machine learning model exhibits improved accuracy and stability. And the accuracy of model performance evaluation was improved using five-fold cross-validation. This model markedly improves the precision of fatigue life prediction and is highly appropriate for augmenting fatigue data under two-step loading conditions.

The data produced by CTGAN effectively addresses the challenge of limited fatigue samples in machine learning applications under variable amplitude loading, all while maintaining clear physical significance. Using generated data as input for machine learning to predict fatigue life holds significant potential in engineering, as it improves accuracy and addresses the challenge of data scarcity. Future studies ought to concentrate on comprehensive evaluations and assessments of the reliability of fatigue life predictions for welded materials subjected to multistage loading, to bolster the robustness and applicability of models.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation of China under Grant (52005071) and Liaoning Provincial Educational Department Project under Grant (2023JH2/101300236).

REFERENCES

[1] Schijve, J. (2003). Fatigue of structures and materials in the 20th century and the state of the art. International Journal of fatigue, 25(8), pp. 679-702. DOI: 10.1016/S0142-1123(03)00051-3.

[2] Gan, L., Wu, H., and Zhong, Z. (2023). Estimation of remaining fatigue life with an energy-based model considering the effects of loading sequence and load interaction. International Journal of Damage Mechanics, 32(3), pp. 340-361. DOI: 10.1177/10567895221120286

[3] Horňas, J., Běhal, J., Homola, P., Senck, S., Holzleitner, M., Godja, N., Pásztor, Z., Hegedüs, B., Doubrava, R., Růžek, R., and Petrusová, L. (2023). Modelling fatigue life prediction of additively manufactured Ti-6Al-4V samples using machine learning approach. International Journal of Fatigue, 169, 107483. DOI: 10.1016/j.ijfatigue.2022.107483.
[4] Miner, M. A. (1945). Cumulative damage in fatigue. J Appl Mech, 12(3), pp. A159-A164. DOI: 10.1115/1.4009458.
[5] Duyi, Y., and Zhenlin, W. (2001). A new approach to low-cycle fatigue damage based on exhaustion of static toughness and dissipation of cyclic plastic strain energy during fatigue. International Journal of Fatigue, 23(8), pp. 679-687. DOI: 10.1016/S0142-1123(01)00027-5.
[6] Lv, Z., Huang, H. Z., Zhu, S. P., Gao, H., and Zuo, F. (2015). A modified nonlinear fatigue damage accumulation model. International Journal of Damage Mechanics, 24(2), pp. 168-181. DOI: 10.1177/1056789514524075.
[7] Wang, X., Liu, M. Z., Cai, F. H., Liang, J. F., Du, J. W., and Su, X. (2018). Nonlinear fatigue damage accumulation model based on load interaction effects. Chinese Journal of Construction Machinery, 16(4), pp. 352-355. DOI: 10.15999/j.cnki.311926.2018.04.014.
[8] Peng, Z., Huang, H. Z., Zhu, S. P., Gao, H., and Lv, Z. (2016). A fatigue driving energy approach to high-cycle fatigue life estimation under variable amplitude loading. Fatigue & Fracture of Engineering Materials & Structures, 39(2), pp. 180-193. DOI: 10.1111/ffe.12347.
[9] Wang, H., Li, B., Gong, J., and Xuan, F. Z. (2023). Machine learning-based fatigue life prediction of metal materials: Perspectives of physics-informed and data-driven hybrid methods. Engineering Fracture Mechanics, 284, 109242. DOI: 10.1016/j.engfracmech.2023.109242
[10] Gan, L., Zhao, X., Wu, H., and Zhong, Z. (2021). Estimation of remaining fatigue life under two-step loading based on kernel-extreme learning machine. International Journal of Fatigue, 148, 106190. DOI: 10.1016/j.ijfatigue.2021.106190.
[11] Liu, X., Zhang, S., Cong, T., Zeng, F., Wang, X., and Wang, W. (2024). Very high-cycle fatigue life prediction of high-strength steel based on machine learning. Fatigue & Fracture of Engineering Materials & Structures, 47(3), pp. 1024-1035. DOI: 10.1111/ffe.14213.
[12] Azadi, M., and Matin, M. (2024). Shapley additive explanation on machine learning predictions of fatigue lifetimes in piston aluminum alloys under different manufacturing and loading conditions. Frattura ed Integrità Strutturale, 18(68), 357-370. DOI: 10.3221/IGF-ESIS.68.24.
[13] Zou, L., Yang, Y., Yang, X., and Sun, Y. (2023). Fatigue life prediction of welded joints based on improved support vector regression model under two-level loading. Fatigue & Fracture of Engineering Materials & Structures, 46(5), pp. 1864-1880. DOI: 10.1111/ffe.13969.
[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. 28th Conference on Neural Information Processing Systems (NIPS), Montreal, Canada, 08-13 December.
[15] Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, CANADA, 08-14 December.
[16] He, G., Zhao, Y., and Yan, C. (2022). Application of tabular data synthesis using generative adversarial networks on machine learning-based multiaxial fatigue life prediction. International Journal of Pressure Vessels and Piping, 199, 104779. DOI: 10.1016/j.ijpvp.2022.104779.
[17] Sun, X., Zhou, K., Shi, S., Song, K., and Chen, X. (2022). A new cyclical generative adversarial network based data augmentation method for multiaxial fatigue life prediction. International Journal of Fatigue, 162, 106996. DOI: 10.1016/j.ijfatigue.2022.106996.
[18] He, G., Zhao, Y., and Yan, C. (2023). A physics-informed generative adversarial network framework for multiaxial fatigue life prediction. Fatigue & Fracture of Engineering Materials & Structures, 46(10), pp. 4036-4052. DOI: 10.1111/ffe.14123.
[19] Chen, J., and Liu, Y. (2022). Fatigue modeling using neural networks: A comprehensive review. Fatigue & Fracture of Engineering Materials & Structures, 45(4), pp. 945-979. DOI: 10.1111/ffe.13640.
[20] Tian, J., Liu, Z. M., and He, R. (2012). Nonlinear fatigue-cumulative damage model for welded aluminum alloy joint of EMU. Journal of the China Railway Society, 34(3), pp. 40-43. DOI: 10.3969/j.issn.1001-8360.2012.03.007.
[21] He, R. (2008). Study on fatigue performance of aluminum alloy welded joint for high-speed train. Beijing Jiaotong University.
[22] Pavlou, D. G. (2002). A phenomenological fatigue damage accumulation rule based on hardness increasing, for the 2024-T42 aluminum. Engineering Structures, 24(11), pp. 1363-1368. DOI: 10.1016/S0141-0296(02)00055-X.

[23] Carvalho, A. L., Martins, J. P., and Voorlwad, H. J. (2010). Fatigue damage accumulation in aluminum 7050-T7451 alloy subjected to block programs loading under step-down sequence. Procedia Engineering, 2(1), pp. 2037-2043. DOI: 10.1016/j.proeng.2010.03.219.

[24] Fu, X. (2018). Research on fatigue damage model under multi-load effect and life prediction of compressor blade. Tianjin University.

[25] Zhou, Z., Zhong, Y., Liu, X., Li, Q., and Han, S. (2020). DC-MMD-GAN: A new maximum mean discrepancy generative adversarial network using divide and conquer. Applied Sciences, 10(18), pp. 6405. DOI: 10.3390/app10186405.