Heliyon 10 (2024) e26429

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# A novel machine learning-based imputation strategy for missing data in step-stress accelerated degradation test

Yaqiu Li [a,d], Qijie Zhou [a,d], Ye Fan [b], Guangze Pan [a,c,*], Zongbei Dai [a], Baimao Lei [a]

[a] *China Electronic Product Reliability and Environmental Testing Research Institute, No. 76, West Zhucun Avenue, Guangzhou, China*
[b] *Beijing Institute of Structure and Environment Engineer, No.1, South Dahongmen Avenue, Beijing, China*
[c] *Guangdong Provincial Key Laboratory of Electronic Information Products Reliability Technology, No. 76, West Zhucun Avenue, Guangzhou, China*
[d] *Key Laboratory of Active Medical Devices Quality & Reliability Management and Assessment, No. 76, West Zhucun Avenue, Guangzhou, China*

ARTICLE INFO

*Keywords:*
Missing data imputation
Accelerated degradation test
Support vector machine
Radial basis function

ABSTRACT

The presence of missing data is a significant data quality issue that negatively impacts the accuracy and reliability of data analysis. This issue is especially relevant in the context of accelerated tests, particularly for step-stress accelerated degradation tests. While missing data can occur due to objective factors or human error, high missing rate is an inevitable pattern of missing data that will occur during the conversion process of accelerated test data. This type of missing data manifests as a degradation dataset with unequal measuring intervals. Therefore, developing a more appropriate imputation method for accelerated test data is essential. In this study, we propose a novel hybrid imputation method that combines the LSSVM and RBF models to address missing data problems. A comparison is conducted between the proposed model and various traditional and machine learning imputation methods using simulation data, to justify the advantages of the proposed model over the existing methods. Finally, the proposed model is implemented on real degradation datasets of the super-luminescent diode (SLD) to validate its performance and effectiveness in dealing with missing data in step-stress accelerated degradation test. Additionally, due to the generalizability of the proposed method, it is expected to be applicable in other scenarios with high missing data rates.

## 1. Introduction

Accelerated degradation test is a highly desirable measure to shorten new product introduction time given the pressure on industry today, as its advantages of obtaining richer information for reliability assessment within limited time and costs by exposing the products to harder-than-normal stress [1]. In accelerated test, the product samples are usually subjected to successively higher stress levels in predetermined stages, and thus follow a time-varying stress profile [2]. The test is terminated when a certain number of failures or degradation data are observed, then it is necessary to transform the collected data at higher stress into equivalent failure-to-time data at normal stress for lifetime distribution fitting and further reliability estimation [3–5]. However, the presence of missing data throughout the entire test can adversely impact the performance of data analysis methods and introduce bias in evaluation results, posing a significant challenge in accelerated test research.

* Corresponding author. China Electronic Product Reliability and Environmental Testing Research Institute, No. 76, West Zhucun Avenue, Guangzhou, China.
*E-mail address:* panguangze@126.com (G. Pan).

https://doi.org/10.1016/j.heliyon.2024.e26429
Received 10 April 2023; Received in revised form 25 November 2023; Accepted 13 February 2024
Available online 18 February 2024

The lack of degradation data from accelerated tests primarily arises from objective and human factors, including sensor failure and errors made by testers. The data gaps can occur at the beginning, middle, or end of a time-ordered dataset, resulting in data absence patterns $M_1$, $M_2$ and $M_3$ shown in Fig. 1. Additionally, the intervals of transformed data obtained from higher stress levels may differ from those of the observed data at normal stress, resulting in regular yet unequal gaps throughout the degradation dataset, referred to the missing data pattern $M_4$ in Fig. 1. For studies involving data processing for accelerated testing, evaluation of product degradation, as well as remaining lifetime prediction, complete data sets are essential as input. The presence of missing data remarkably complicates the processing of performance degradation data, rendering many traditional methods incapables of conducting statistical analysis on incomplete datasets. For instance, time-series based remaining lifetime prediction methods usually require the data set to be both complete and equidistant.

To address the issue of missing data, there are two potential approaches: enhancing traditional data processing methods to handle degraded data with missing values or transforming the data with missing values into complete datasets. The former option poses implementation challenges, whereas the latter is regarded as a more feasible and realistic approach. However, the practicability of imputing missing data often depends on the underlying missing mechanism, which are categorized as follows: MAR (Missing At Random), MCAR (Missing Completely At Random), and MNAR (Missing Not At Random) [6]. Rubin has designed a probabilistic framework for missing data, he supposes that the complete data as consisting of two components, the observed data and the missing data ($Y_{obs}$ and $Y_{mis}$, respectively), and defines a binary variable $M$ that denotes whether a value on a particular variable is observed or missing (i.e., $M = 1$ if a value is observed, and $M = 0$ if a value is missing). A more precise description of each missing data category is listed in Table 1 [7]. Generally, if the missing mechanism is MNAR, which means that the missing data is influenced by both the observed and missing data, it is not feasible to interpolate the missing data. Fortunately, performance degradation data often adhere to a continuous degrading process, allowing for the assumption that the missing degradation data depends solely on the observed data rather than the missing portion. Consequently, the imputation methods are promising to handle the missing data issue of degradation data.

Missing data imputation can be performed by several models, such as the mean model [8], regression model [9,10], cluster model [11,12], hot-deck model [13] and nearest neighbor model [14,15]. According to the number of imputations, the imputation methods can be classified into single and multiple imputations. The single imputation methods often fill each missing value with either the average of its predicted distribution or a randomly drawn value from that distribution, and various statistical analysis are able to be performed on the complete data set following imputation. However, few single imputation methods take uncertainty into account and they have a common drawback of distorting the sample data distributions. The multiple imputation method entails generating multiple alternative values for each missing value, resulting in several complete data sets. These complete data sets are then evaluated using the same approach to yield multiple evaluations. Finally, the results from these evaluations are combined to obtain a final estimate of the target variable. Multiple imputation methods offer improved handling of the uncertainty associated with missing data and maintain the distribution of data samples. However, these methods are computationally intensive and often yield less accurate values.

During the last decades, machine learning techniques are extensively employed for anomaly detection, assessment prediction, and missing data imputation, leveraging their excellent capability to extract valuable information from data, such as the Neural Network (NN) model, Support Vector Machine (SVM) model, Naive Bayes (NB) model and so on [16]. Although NN models, such as radial basis

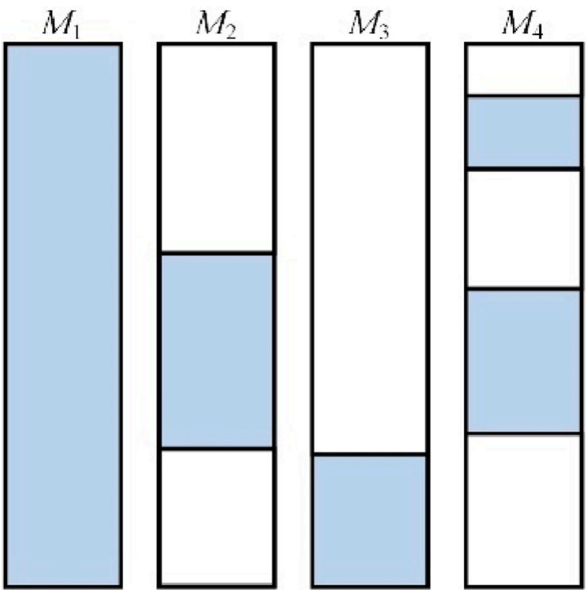



**Fig. 1.** Four prototypical missing data patterns of degradation data. The shaded areas represent the location of the missing values in the dataset. $M_1$, $M_2$, *and* $M_3$ represent missing data occurring in the front, middle and back parts of the dataset respectively, $M_4$ represents missing data caused by transforming.

**Table 1**
Description and mathematical expression of missing data types.

| Category | Precise Description | Mathematical Expression |
|---|---|---|
| MAR | Data are missing at random when the probability of missing data on a variable $Y$ is related to some other measured variables in the analysis model but not to the values of $Y$ itself. | $f(M\|Y) = f(M\|Y_{obs}, \varphi)$ |
| MCAR | Data are missing completely at random when the probability of missing data on a variable $Y$ is unrelated to other measured variables and is unrelated to the values of $Y$ itself. | $f(M\|Y) = f(M\|\varphi)$ |
| MNAR | Data are missing not at random (MNAR) when the probability of missing data on a variable $Y$ is related to the values of $Y$ itself, even after controlling for other variables. | $f(M\|Y) = f(M\|Y_{obs}, Y_{mis}, \varphi)$ |

$\varphi$ is a parameter that describes the relationship between $M$ and the data.

function (RBF), utilize their strong non-linear approximation characteristics to overcome the limitation of insufficient data supply [17–19], they suffer from drawbacks such as high sample requirements, slow convergence, susceptibility to local minimum trapping, and limited generalization ability. These shortcomings of the NN model can be mitigated by employing other machine learning techniques, such as SVM, which utilizes a kernel to transform the data from the input space into a higher-dimensional feature space, enabling linear separability of the problem. As a result, SVM exhibits improved generalization capability and is available for the small sample sizes compared to conventional neural network models [20]. Besides, the least squares support vector machine (LSSVM) incorporates an additional sum squared error term in its objective function for the optimization problem, which decreases the computation time of the convex optimization problem and improves the performance, leading to high precision and fast convergence compared to other SVM models [21].

In practical application, the prediction accuracy of LSSVM is affected by the kernel function and the parameters of the kernel function [22]. At present, there is no definite theory and method to support how to determine the kernel function and the parameters of the kernel function. Moreover, researchers have observed that the output residual terms of SVM-like models contain valuable information that can be further learned. Based on this observation, hybrid models are promising to refine the predictions of the LSSVM by retrain the residual terms with other machine learning techniques [23].

Based on the above description, this paper proposes hybrid intelligent models to impute the missing data value in step-stress accelerated degradation test. The proposed models combine the LSSVM with a RBF neural network model. It is worth mentioning that this is the first work that investigate the impact of missing data on degradation evaluation and explore the effectiveness of imputation techniques in the field of accelerated testing. The main innovations and research ideas of this paper are given.

- The RBF and LSSVM methods are introduced respectively to train the model using observed data and subsequently impute the missing data through prediction. A comparative analysis is conducted to determine the applicability of these two models, based on which the framework of the hybrid imputation model is proposed.
- By integrating the LSSVM and RBF models, the proposed hybrid model achieves a more comprehensive and accurate imputation of the missing data. The LSSVM models the degradation trend, ensuring that the imputation data follows a consistent trend with the observed data, while the RBF is utilized to estimate the residual series of the missing data, thereby maintaining a similar data dispersion as the observed data.
- Data imputation strategies and improved process (missing data with unequal measuring intervals) are specifically designed for possible data missing mechanisms in accelerated degradation test. The performance of the proposed model is validated by comparing with several traditional methods.

The paper is organized as follows. Section 2 illustrates the theoretical foundations and detailed process of the proposed imputation method. Section 3 presents the influences of rate and measuring interval of missing data, comparing the effects of several traditional techniques and the proposed method. Section 4 shows an experiment case. Finally, conclusions and future work directions are discussed in Section 5.

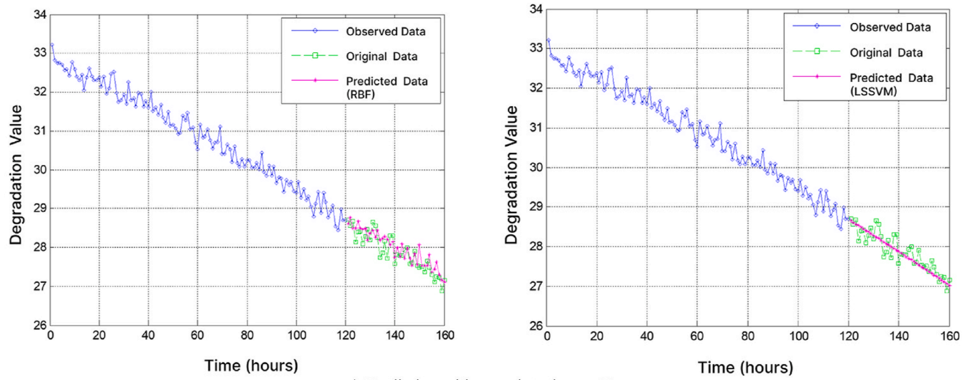
## 2. Methodology

### 2.1. Machine learning techniques

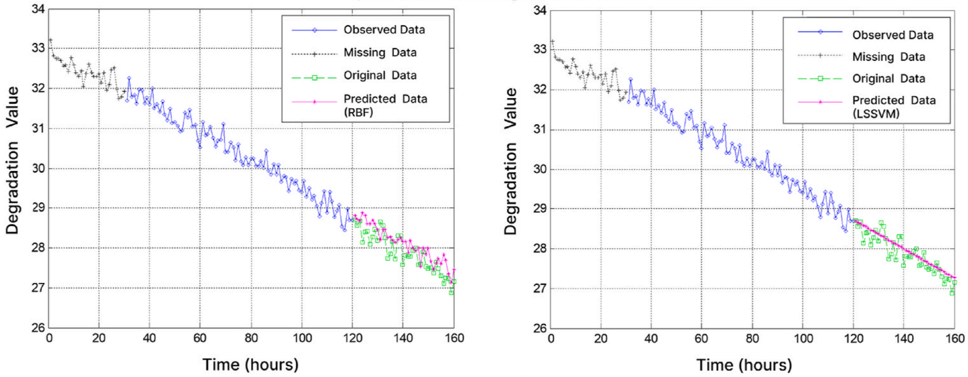
In recent years, machine learning techniques are widely used for predictive analytics. However, these techniques also possess significant potential for effectively handling incomplete data.

#### 2.1.1. Radial basis function (RBF) neural network

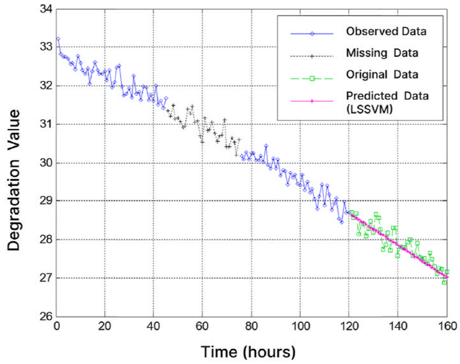
The radial basis function (RBF) neural network model possesses a high ability for function approximation, particularly in dealing with nonlinear data problems [24]. The model does not require a large number of samples and has a strong generalization ability for each input and output sample. The "basis" of the hidden layer neurons in the RBF neural network model is the radial basis function, which denotes a scalar function with radial symmetry. Its mathematical expression is $R(||x\text{-}c||)$, which is a monotonic function of the



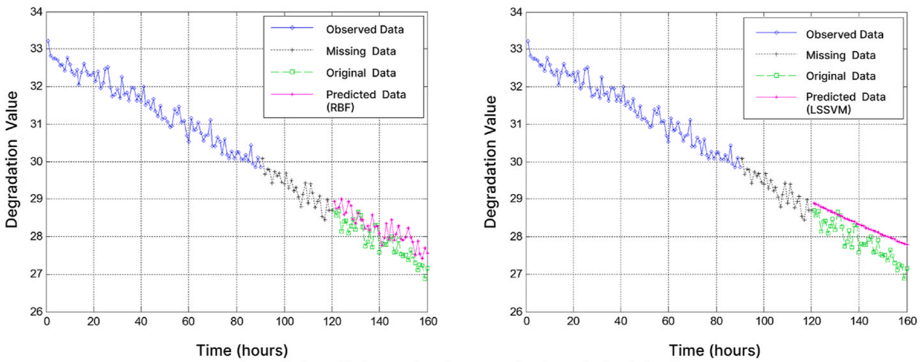

a) Prediction with complete dataset $Y_0$




b) Prediction with incomplete dataset $Y_1$ (30 data missing in front)




c) Prediction with incomplete dataset $Y_2$ (30 data missing in middle)




d) Prediction with incomplete dataset $Y_3$ (30 data missing in back)

**Fig. 2.** Prediction results of RBF and LSSVM models for complete dataset and incomplete datasets with missing data in the front, middle and back of the dataset.

Euclidean distance between any point in space ($x$) and the center ($c$), as expressed in Eq. (1) and Eq. (2).

$$R(\|x-c\|) = \exp\left\{-\|x-c\|^2 / \left(2\sigma^2\right)\right\} \tag{1}$$

$$\sigma = d_{\max} \Big/ \sqrt{2h} \tag{2}$$

where c is the kernel function center, σ is variance and is the width parameter of the function, $d_{\max}$ is the maximum Euclidean distance between all central vectors, and $h$ is the number of hidden layer neurons.

#### 2.1.2. Least squares support vector machine (LSSVM)

Least Squares Support Vector Machine (LSSVM) is an improved machine learning algorithm of SVM proposed by Suykens [25]. It has the advantages of quick learning speed, good generalization ability, and does not suffer from overfitting of other neural network models and long training time of SVM. The regression algorithm of the LSSVM can be represented as Eq. (3) [26]:

$$y(x) = \omega^T \varphi(x) + b \tag{3}$$

Where $\omega$ is the weight vector, $b$ is the bias vector, and $\varphi(x)$ is the nonlinear mapping function. The optimization problem can be expressed as:

$$\min J(\omega, e) = \frac{1}{2}\omega^T \omega + \gamma \sum_{i=1}^{l} e_i^2 \tag{4}$$

with the constraints as Eq. (5),

$$y_i = \omega^T \varphi(x_i) + b + e_i, i = 1, \cdot\cdot\cdot, l \tag{5}$$

where $\omega^T\omega$ controls the complexity of the decision function, $\gamma$ is the regularization parameter, which controls the degree of penalty beyond the error sample. $e$ is the error vector.

By introducing the Lagrange function, Eq. (4) can be transformed as Eq. (6)

$$y(x) = \sum_{i=1}^{k} \delta_i \times K(x - x_i) + b \tag{6}$$

where $\delta_i$ is the Lagrangian multiplier and $K(x, x_i)$ is the symmetric kernel function satisfying the Mercer condition, if we choose the RBF as the kernel function of LSSVM, thus $K(x, x_i) = R(||x\text{-}c||)$.

#### 2.1.3. Impact of missing degradation data on degradation predictions

Missing data issue is crucial to ensure the robustness and accuracy of the degradation behavior prediction. Therefore, prior to constructing the imputation model, a comparison analysis is conducted between RBF and LSSVM models to assess the impact of missing data.

Observed data missing due to monitoring sensors failure or record errors may occur at any time during the whole accelerated test. The influence of missing data in the front, middle and back of the dataset are compared and analyzed by four groups of simulated performance degradation data.

##### 2.1.3.1. Supposing that the degradation model is an approximate linear monotonic descent model as Eq. (7)

$$Y = at + b + \varepsilon, t \in N, a \in R \text{ and } a < 0, b \in R \text{ and } b > 0, \varepsilon \sim N(0, 1) \tag{7}$$

Random sampling is conducted to collect the complete degradation dataset $Y_0$ and the size of dataset is 120. Missing datasets $Y_1$, $Y_2$, $Y_3$ are consists of the remaining 90 data after removing 30 data values from the front, middle and back of $Y_0$ respectively. RBF and LSSVM methods are used for predicting the next 40 values, as shown in Fig. 2. (a)-(d).

Mean Absolute Percentage Error (MAPE) measures the relative magnitude of deviations between predicted and true values. MAPE is often preferred over other fit error metrics, such as Mean Squared Error (MSE) or Root Mean Squared Error (RMSE), as it is less sensitive to the influence of extreme values. The value of MAPE can be obtained as Eq.(8),

$$MAPE = = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i' - y_i}{y_i} \right| \tag{8}$$

In practical, a smaller MAPE value indicates a higher level of accuracy in the prediction model. The MAPE values for the predictions generated by the RBF and LSSVM models for each dataset are calculated. Moreover, the MAPE of the complete dataset ($Y_0$) serves as the benchmark to assess the extent of decline in precision resulting from data missing situations in $Y_1$, $Y_2$, $Y_3$. These MAPE values, along with the declining precision analysis, are presented in Table 2, providing a comprehensive overview of the impact of data missing scenarios on prediction accuracy.

As results shows, missing data will always lead to a decrease in the precision of evaluation or prediction models. Specifically, when the missing data is located at the beginning of the dataset, the MAPE for RBF-based time series predictions decreased by 30.7% compared to that based on complete data, while the MAPE for LSSVM predictions only decreased 11.7% in this scenario. In cases where the missing data is situated in the middle of the dataset, the impact on LSSVM prediction accuracy was relatively less pronounced. However, missing data in the middle makes RBF-based time series prediction unreliable due to the fact that missing data pattern in this scenario fails to meet the requirement of equally interval inputs, which is necessary for this kind of prediction model. When the missing data occurred at the end of the dataset, it had the most significant impact on prediction accuracy. In this situation, the MAPE for RBF and LSSVM predictions decreased by 68.0% and 166.7% respectively.

The results demonstrate that the LSSVM model is more successful in preserving the declining trend of the original data. However, it struggles to accurately reflect the uncertainty of the observed data, which may result in an underestimation of data uncertainty. On the other hand, while the RBF model requires strict adherence to equal interval inputs, it effectively maintains the discrete characteristics of the original data. It is worth noting that a relatively smaller SPREAD parameter in the RBF model can lead to improved data dispersion.

### *2.2. Hybrid imputation model based on LSSVM and RBF*

According to the above analysis, this paper merges LSSVM and RBF methods into an improved model, in which the LSSVM algorithm is used for modeling the trend of degradation data and meanwhile a margin of random residual error is controlled by the RBF algorithm so that the imputation results can reflect the uncertainty associated with missing data. The framework of the proposed model is shown in Fig. 3.

The detailed modeling procedure is as follows.

(1) Collecting testing data to form the observed degradation dataset $\{\mathbf{T}_{obs}, \mathbf{Y}_{obs}\}$, in which $\mathbf{T}_{obs} = \{t_{obs\_1}, t_{obs\_2}, \ldots, t_{obs\_n}\}$ denotes the test timing points and $\mathbf{Y}_{obs} = \{y_{obs\_1}, y_{obs\_2}, \ldots, y_{obs\_n}\}$ denotes to the observed degradation value.

(2) Training the LSSVM model with the vector $\mathbf{T}_{obs}$, $\mathbf{Y}_{obs}$ as input and output respectively to derive the trend model $f(t)$ of degradation data as Eq. (9).

$$f(t) = \sum_{i=1}^{n_{obs}} \alpha_i \psi(t, t_{obs\_i}) + b \tag{9}$$

(3) Calculating the trend term series of missing data ($\mathbf{Q}_{mis}$) by substituting the timing points of missing data ($\mathbf{T}_{mis}$) into Eq. (7), where $\mathbf{Q}_{mis} = \{q_{mis\_1}, q_{mis\_2}, \ldots, q_{mis\_m}\}$ and $\mathbf{T}_{mis} = \{t_{mis\_1}, t_{mis\_2}, \ldots, t_{mis\_m}\}$.

(4) The trend term series of observed data $\mathbf{Q}_{obs} = \{q_{obs\_1}, q_{obs\_2}, \ldots, q_{obs\_m}\}$ can be obtained by substituting $\mathbf{T}_{obs}$ into Eq. (7), hence the residual term series of observed data $\mathbf{E}_{obs} = \{e_{obs\_1}, e_{obs\_2}, \ldots, e_{obs\_n}\}$ can be calculated as Eq. (10):

$$e_{obs\_i} = y_{obs\_i} - q_{obs\_i}, i = 1, 2, \cdots, n \tag{10}$$

(5) Training the RBF model with the vector $\mathbf{T}_{obs}$ as input and $\mathbf{E}_{obs}$ as output, then taking the vector $\mathbf{T}_{mis}$ as input into trained RBF model to predict the residual term of missing data, obtaining the residual term series $\mathbf{E}_{mis} = \{e_{mis\_1}, e_{mis\_2}, \ldots, e_{mis\_m}\}$.

(6) The estimated trend term and residual term are continuously added as new data into the training set $\{\mathbf{E}_{obs}, \mathbf{T}_{obs}\}$ for updating the RBF model, in order to enhance the prediction precision.

(7) Obtaining the Missing data ($\mathbf{Y}_{mis}$) by merging the trend term and residual term as Eq. (9) shows. Finally, the complete degradation dataset is formed by combining the observed data series and predicted missing data series as Eq. (11).

$$y_{mis\_i} = q_{mis\_i} + e_{mis\_i}, i = 1, 2, \cdots, m \tag{11}$$

## 3. Imputation strategies for missing data patterns

### *3.1. Strategies for missing data with different missing rate*

The missing data rate is one of the most vital factors for evaluation as well as imputation. Datasets with high missing rates may

**Table 2**
MAPE of prediction result of RBF and LSSVM model.

| Dataset | RBF Model | | LSSVM Model | |
|---|---|---|---|---|
| | MAPE (%) | Decline in Precision | MAPE (%) | Decline in Precision |
| $Y_0$ | 0.75 | – | 0.60 | – |
| $Y_1$ | 0.98 | 30.7% | 0.67 | 11.7% |
| $Y_2$ | / | / | 0.61 | 1.67% |
| $Y_3$ | 1.26 | 68.0% | 1.60 | 166.7% |

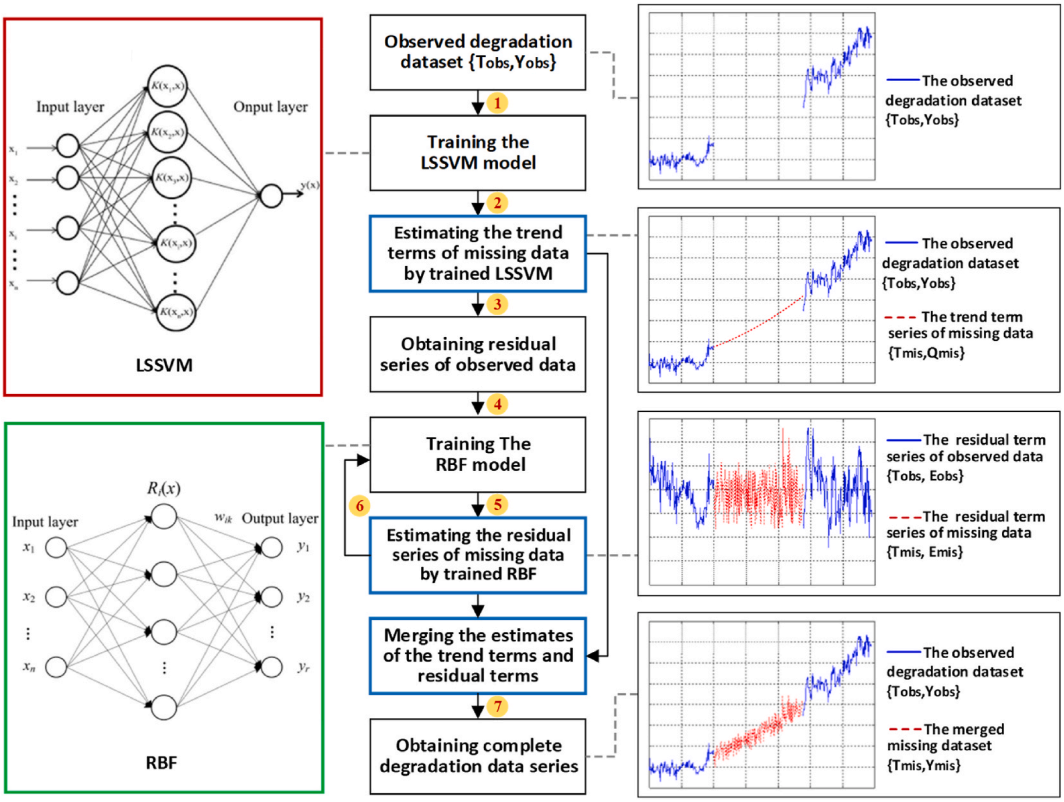



**Fig. 3.** The framework of hybrid imputation method based on LSSVM-RBF model and an illustrative example for each modeling step.

cause instability in the imputation process, resulting in huge errors in final prediction and evaluation. In this section, the influences of missing data rate at different levels are analyzed by several methods with simulation data.

Supposing that the degradation model as Eq. (12),

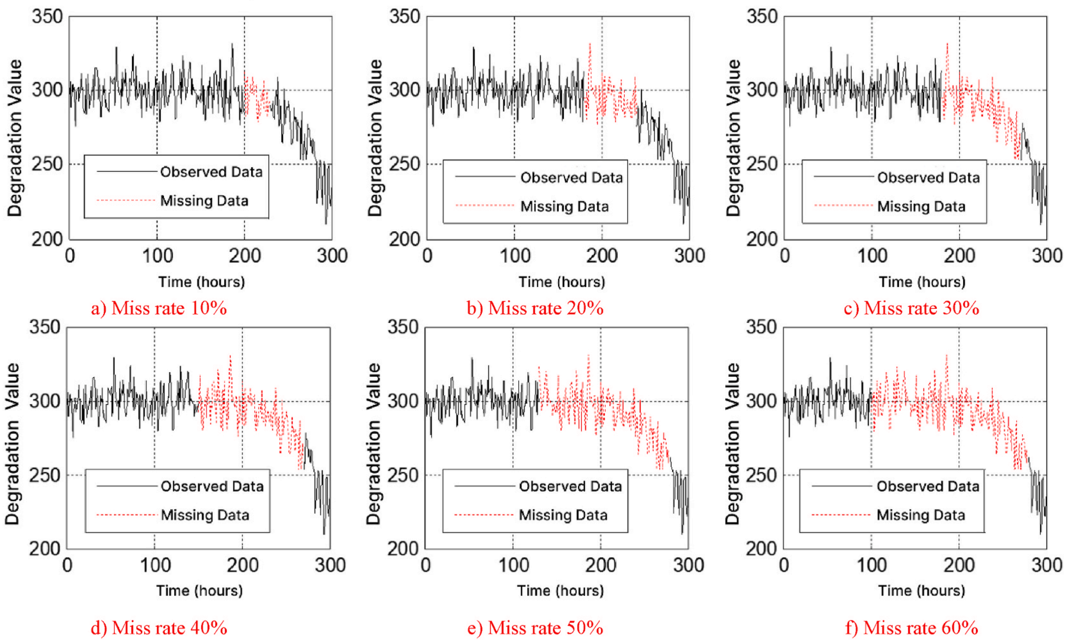



**Fig. 4.** Degradation dataset at different missing rate.

$$Y = 300 - 0.0095 \times e^{0.03t} + \varepsilon, t \in N, \varepsilon \sim N(0,100) \tag{12}$$

Random sampling is conducted to collect the complete degradation dataset $Y_0$, whose size is 300. The missing datasets $\{Y_i\}$, $i = 1, \ldots, 6$, are obtained by removing the continuous date from the middle part of each dataset to fit the missing rate of 10%, 20%, 30%, 40%, 50% and 60%. The degradation datasets at different missing rate are shown in Fig. 4. (a)-(f).

Imputation method based on mean model, regression model, EM algorithm, regression-RBF and LSSVM-RBF models are conducted, and their imputation effects are evaluated according to the metrics of root mean square error (RMSE) and relative variance error ($\varepsilon_{\sigma^2}$) respectively as Eq. (13) and Eq. (14).

$$RMSE(X,h) = \sqrt{\frac{1}{10}\sum_{i=1}^{10}\left(h(x_i) - y_i\right)^2} \tag{13}$$

$$\varepsilon_{\sigma^2} = \frac{\sigma'^2 - \sigma^2}{\sigma^2} \tag{14}$$

where $X$ denotes the timepoint set of degradation data, $h(x)$ denotes the models aforementioned, $\sigma^2$ and $\sigma'^2$ denote the variances of original and imputation data separately. The RMSE quantifies the average magnitude of discrepancies between predicted and actual values within a dataset, providing a valuable assessment of the alignment between a model's predictions and observed data points. Additionally, the $\varepsilon_{\sigma^2}$ serves as the evaluation metric for gauging the effectiveness of interpolated data in accurately representing the variability present in the original dataset. The evaluation results are shown in Fig. 5.

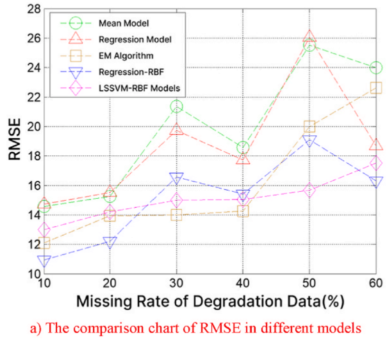
From Fig. 5. (a), it is evident that the RMSE of each imputation method tends to increase as the missing rate gradually rises. Specifically, when the missing rate is below 20%, the regression-RBF model exhibits the smallest RMSE, followed by the EM algorithm and the LSSVM-RBF model. However, at missing rates of 30% and 40%, the EM algorithm achieves the smallest RMSE, followed by LSSVM-RBF and the regression-RBF. As the missing rate exceeds 50%, the RMSE of the EM algorithm experiences a notable increase. At the missing rate of 60%, the RMSE of the EM algorithm surpasses that of the regression model, which becomes the third smallest, closely resembling that of the regression-RBF as well as LSSVM-RBF. Notably, the mean imputation method consistently exhibits the largest RMSE.

The comparison of $\varepsilon_{\sigma^2}$ in Fig. 5. (b) reveals that among the five methods, regression imputation demonstrates the closest resemblance to the original data variance, exhibiting remarkable stability across all missing rates. Following closely are the imputation method based on regression-RBF and LSSVM-RBF, both displaying comparable proximity to the original data variance and exhibiting a high level of stability. When the missing rate is below 40%, the interpolated data variance of the EM algorithm performs nearly as well as that of the regression method. However, as the missing rate surpasses 50%, the variance of the interpolated data from the EM algorithm experiences significant deviations. Notably, the mean imputation method showcases the largest error in the relative variance error.

Based on the analysis above, the effects of handling the influence of the missing rate of five models are evaluated as Table .3 shows.

### 3.2. Strategies for missing data with unequal measuring intervals

Equally spaced data serves as a fundamental requirement for conducting time series analysis. Consequently, the imputation of data with unequally measuring intervals into equally spaced data becomes a prerequisite for regression and prediction tasks. In the case of the step-stress accelerated test, the degradation data will inevitably change to data with unequal interval in time when folded to the same stress level, which may bring difficulties to subsequent data analysis. Essentially, this situation can be regarded as a special missing data problem, which means imputation method can be conducted to improve the dataset quality.



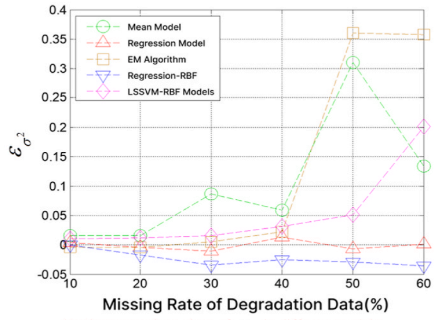




**Fig. 5.** RMSE and $\varepsilon_{\sigma^2}$ evaluation of mean model, regression model, EM algorithm, regression-RBF and LSSVM-RBF models at different missing rate levels.

**Table 3**
Effects of 5 models in dealing with missing data at different rates.

| Evaluation method | Missing rate | Mean Model | Regression Model | EM Model | Regression-RBF Model | LSSVM-RBF Model |
|---|---|---|---|---|---|---|
| RMSE | 10%~40% | Medium | Medium | Good | Good | Good |
| | 50% | Bad | Bad | Medium | Medium | Good |
| | 60% | Bad | Medium | Bad | Good | Good |
| $\varepsilon_{\sigma^2}$ | 10%~40% | Medium | Good | Good | Medium | Good |
| | 50% | Bad | Good | Bad | Medium | Medium |
| | 60% | Medium | Good | Bad | Good | Medium |

### 3.2.1. Determination of the imputation point

The primary goal of imputation for unequal measuring interval data is that the amount of interpolated data should be as small as possible, to ensure that the proportion of original observed data are large enough to maintain statistical characteristics.

Supposing two datasets whose measuring interval is $d_1$ and $d_2$ respectively, and $d_1 > d_2$, there may be two possible cases.

- Case1: $d_1 = kd_2, k \in N$.

In this case, only the first dataset needs to be interpolated $k$-1 data between every two data with the measuring interval of $d_2$, as Fig. 6. (a) shows. Since $k \geq 2$, it means that the first dataset should be considered as a missing dataset with 50% or more missing rate when the imputation is conducted.

- Case 2: $d_1 \neq kd_2, k \in N$.

In this case, the maximum convention, $m$, needs to be calculated firstly, then imputation should to be performed to both datasets. $d_1/m$-1 and $d_2/m$-1 data should be interpolated between every two data for the first and second dataset respectively, unifying the measuring interval of two dataset as $m$, as Fig. 6. (b) shows. Likely, the missing rate of these two datasets in this case is above 50%.

If there are more than two datasets available, the imputation points can be determined by extending these two cases above.

### 3.2.2. Datasets with unequal measuring intervals

Random sampling is conducted according to Eq. (12) to collect the complete degradation dataset with 3 replications while the size of each dataset is 300, namely $Y_1$, $Y_2$, $Y_3$. Then, by removing corresponding data these three datasets are transformed into 'incomplete' datasets $\{Y_1^{'}, Y_2^{'}, Y_3^{'}\}$ with measuring intervals of 2, 3 and 4 respectively, as shown in Fig. 7. (a)-(f).

Since the maximum convention of the three measuring intervals (2, 3, 4) is 1, it is necessary to perform imputation to convert them into dataset with measuring interval of 1. As a result, imputation information of the three datasets can be seen in Table .4.

### 3.2.3. Imputation for incomplete dataset

The minimal data missing rate is close to 50%, as shown in Table .4, indicating that it is necessary to choose imputation methods that can still work in the case of substantial data missing. According to Table .3, the regression model, regression-RBF model, and LSSVM-RBF model exhibit superior performance in data interpolation, particularly at high missing rates. Consequently, case in this chapter is utilized to conduct a further comparison of the imputation performance of these three models.

Dataset with a measuring interval of 3 ($Y_2^{'}$) is taken as an example to illustrate the imputation effects by comparing the imputation data with original data $Y_2$ and 'incomplete' data $Y_2^{'}$. The comparison results of the regression imputation method are shown in Fig. 8. (a) and (b).

Similarly, the comparison results of regression-RBF and LSSVM-RBF imputation methods are shown in Fig. 9. (a)–(c) and Fig. 10.

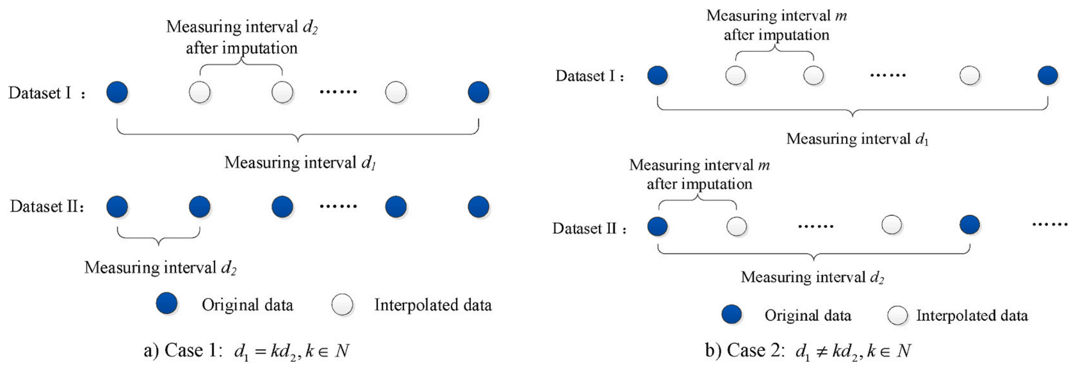
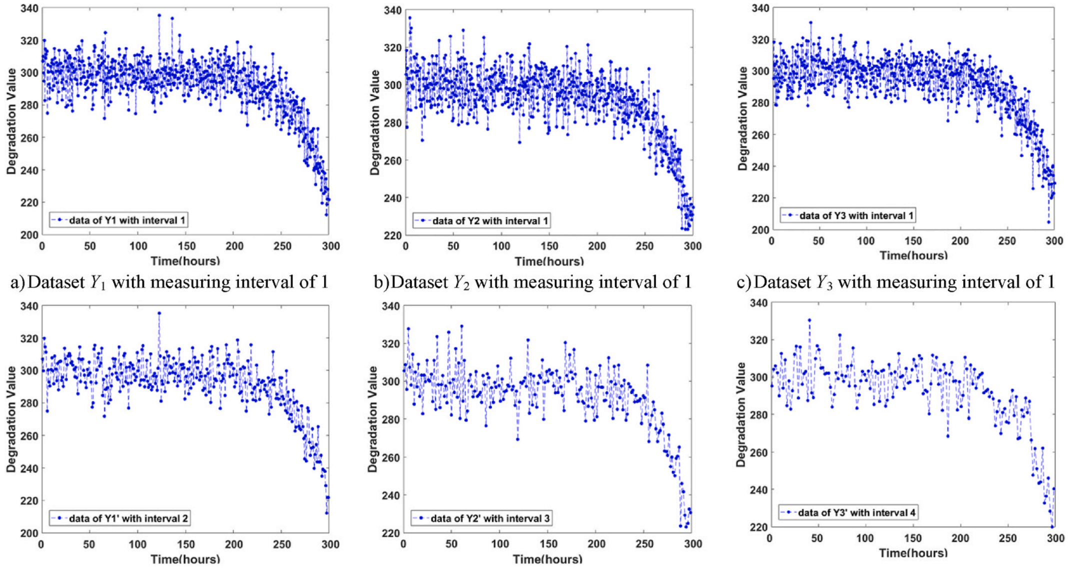



**Fig. 6.** Diagram of determination progress of the imputation point.




a) Dataset $Y_1$ with measuring interval of 1 b) Dataset $Y_2$ with measuring interval of 1 c) Dataset $Y_3$ with measuring interval of 1

d) Dataset $Y_1'$ with measuring interval of 2 e) Dataset $Y_2'$ with measuring interval of 3 f) Dataset $Y_3'$ with measuring interval of 4

**Fig. 7.** The original datasets $Y_1$, $Y_2$, $Y_3$ and missing datasets $Y_1'$, $Y_2'$, $Y_3'$ with measuring interval = 2,3,4.

**Table 4**
Imputation data information.

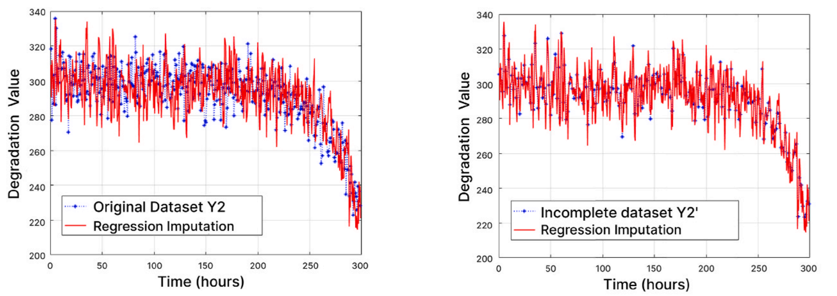
| Dataset | Measuring Interval | | Data Amount | | | Percentage of Imputation data |
|---|---|---|---|---|---|---|
| | Original (hours) | Imputation (hours) | Original | Imputation | Total | |
| $Y_1'$ | 2 | 1 | 150 | 149 | 299 | 49.8% |
| $Y_2'$ | 3 | 1 | 100 | 198 | 298 | 66.4% |
| $Y_3'$ | 4 | 1 | 173 | 516 | 689 | 74.9% |

(a)-(c), and the situation about trend and residual terms of imputation are presented as well.

### 3.2.4. *Evaluation of imputation methods*

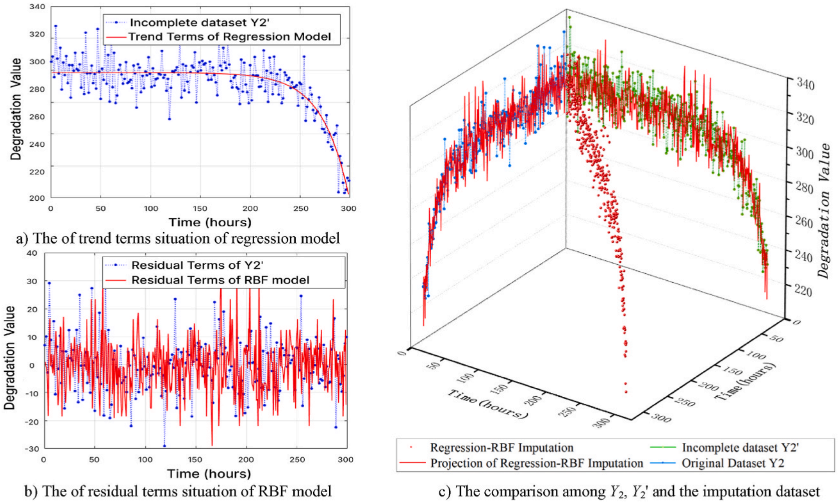
By repeating the simulation approach mentioned in chapter 3.3.2, 10 sets each of data with measuring intervals of 2, 3, 4 ($\{Y_1^i\}, i = 1,2,\cdots,10$, $\{Y_2^j\}, j = 1,2,\cdots,10$, $\{Y_3^k\}, k = 1,2,\cdots,10$) are obtained to evaluate the effects of imputation methods considering RMSE and relative variance error as the metrics. The results of evaluation are shown in Fig. 11. (a)-(f).

In terms of RMSE, the proposed LSSVM-RBF model consistently exhibits significantly smaller RMSE values compared to the other two methods. This indicates that the proposed model achieves a higher level of stability and has fewer imputation points with considerably larger errors. On the other hand, the Regression-RBF model shows less substantial improvement to the regression model than the data with equally interval case discussed in Section 3.1. For the relative variance error ($\varepsilon_{\sigma^2}$), the performance of the Regression-RBF model for data with unequal intervals under high missing rates is similar to that of the LSSVM-RBF. The Regression




a) The comparison between $Y_2$ and the imputation dataset b) The comparison between $Y_2'$ and the imputation dataset

**Fig. 8.** Imputation results of regression method.

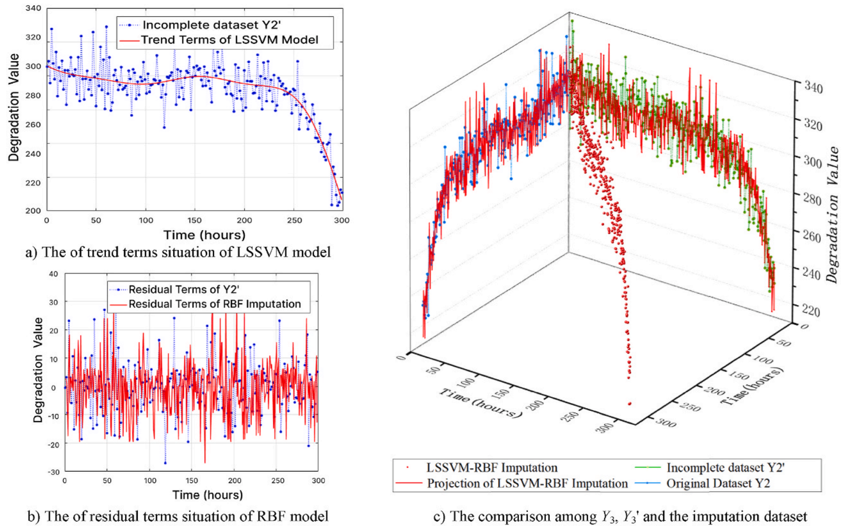



a) The of trend terms situation of regression model

b) The of residual terms situation of RBF model

c) The comparison among $Y_2$, $Y_2$' and the imputation dataset

**Fig. 9.** Imputation results of regression-RBF method.




a) The of trend terms situation of LSSVM model

b) The of residual terms situation of RBF model

c) The comparison among $Y_3$, $Y_3$' and the imputation dataset

**Fig. 10.** Imputation results of LSSVM-RBF method.

model maintains the smallest $\varepsilon_{\sigma^2}$, however, this advantage is not highly pronounced.

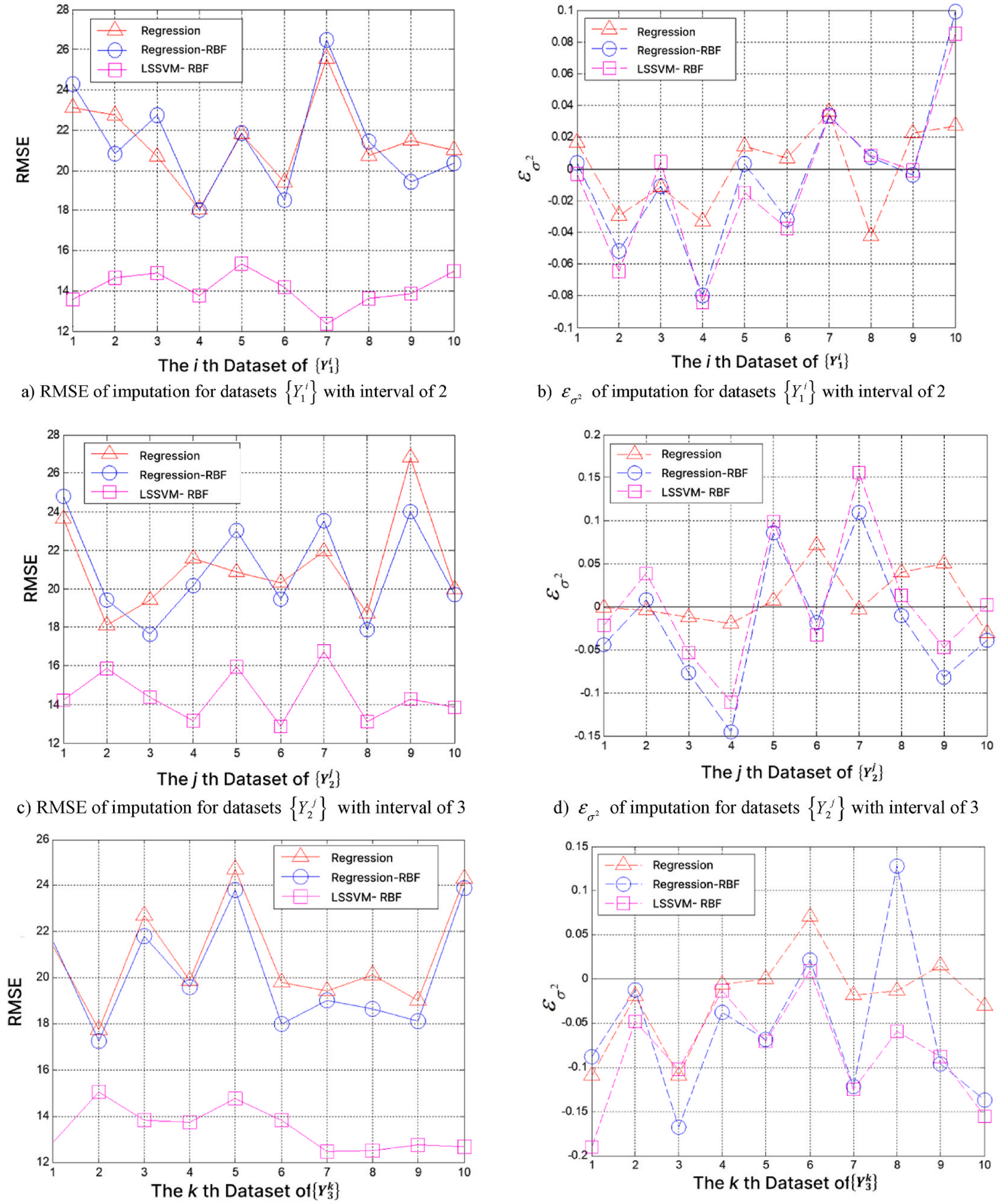
In summary, for scenarios involving data with unequally intervals under high missing rates, the model performance can be ranked in descending order as follow: LSSVM-RBF, Regression, and Regression-RBF model.

## 4. Experiment and discussion

Super-luminescent diode (SLD) is a semiconductor optical component which is now widely used in automotive, medical and radio fields due to its excellent characteristics of high output power, good stability, long life and high reliability. The stability of SLD light source is mainly influenced by the injection current and core temperature, thus temperature is chosen as the applied stress condition for the accelerated test to verify the affection of the imputation method proposed in this paper.

### 4.1. Experiment setup

Step-stress accelerated test is performed at 3 levels of temperature (60 °C, 70 °C, 80 °C) with retention of 1500 h, 1000 h and 500 h respectively. The output optical power of SLD is recorded every 3 h during the test, as shown in Fig. 12. (a) and (b). The differences in the initial value of optical power at 3 stress levels is the result of temperature drift, but it will not affect the subsequent data imputation.




a) RMSE of imputation for datasets $\{Y_1^i\}$ with interval of 2 b) $\varepsilon_{\sigma^2}$ of imputation for datasets $\{Y_1^i\}$ with interval of 2

c) RMSE of imputation for datasets $\{Y_2^j\}$ with interval of 3 d) $\varepsilon_{\sigma^2}$ of imputation for datasets $\{Y_2^j\}$ with interval of 3

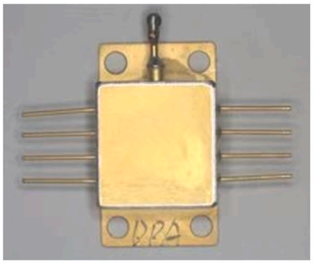
e) RMSE of imputation for datasets $\{Y_3^k\}$ with interval of 3 f) $\varepsilon_{\sigma^2}$ of imputation for datasets $\{Y_3^k\}$ with interval of 3

**Fig. 11.** Evaluation on RMSE and relative variance error of three imputation method.

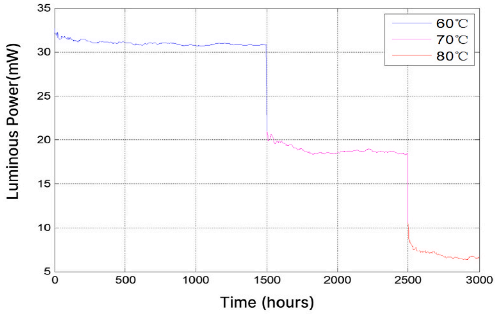
### *4.2. Data processing*

Severe stress conditions typically result in an accelerated degradation of products, which can be quantified using an acceleration factor. The acceleration factor represents the ratio by which the degradation under accelerated stress conditions increases relative to the degradation observed under base stress conditions in the same time period. Alternatively, it can be understood as the time required to achieve the same extent of degradation under accelerated stress conditions being reduced to $1/a$ times that of the base stress conditions. As a result, the measuring interval at accelerated stress levels will expand $a$ times when these data are converted in to the base stress level.

In order to determine the imputation timepoint for the transformed data at base stress level, we first calculating the average degradation rate (ADR) by employing a linear regression model to fit the performance (power) degradation data at each stress level on



a) Test sample of SLD




b) Degradation data of 3 stress levels

**Fig. 12.** Step stress accelerated degradation test of SLD.

the basis of Fig. 12. (b), as shown in Table .5.

$$\begin{cases} 70^{\circ}C \rightarrow 60^{\circ}C : a_1 = \dfrac{ADR_{70^{\circ}C}}{ADR_{60^{\circ}C}} = \dfrac{0.0012}{0.00052} \approx 2 \\ 80^{\circ}C \rightarrow 60^{\circ}C : a_2 = \dfrac{ADR_{80^{\circ}C}}{ADR_{60^{\circ}C}} = \dfrac{0.0032}{0.00052} \approx 6 \end{cases} \tag{15}$$

The $a_1$ and $a_2$ represent the acceleration factors at 70 °C and 80 °C respectively. Hence, the corresponding imputation timepoints of 70 °C and 80 °C transformed to 60 °C can be presented as Eq. (16):

$$\begin{cases} t_{70}' = 1500 + (t_{70} - 1500) \times a_1 \\ t_{80}' = 1500 + 1000 \times a_1 + (t_{80} - 2500) \times a_2 \end{cases} \tag{16}$$

in which $t_{70}$, $t_{80}$ represent to the measuring timepoints at 70 °C and 80 °C respectively, while $t_{70}'$, $t_{80}'$ represent the corresponding timepoints of them at 60 °C after transformation, the transformed data at 60 °C are shown in Fig. 13. According to the accelerated factors $a_1$ and $a_2$, the measuring data intervals of $t_{70}'$ and $t_{80}'$ are approximated 6 and 18 h respectively, in which the 'missing' data need to be interpolated to obtain the complete dataset with equal interval, the specific information of imputation data is shown in Table .6.

### 4.3. Results and discussion

Data imputation is performed by LSSVM-RBF model to deal with the high missing rate of transformed degradation dataset, the results can be seen in Fig. 14. (a)-(f).

The RMSEs for the transformed datasets at 70 °C and 80 °C are calculated to be 4.66 and 5.45, respectively. The corresponding values of $\varepsilon_{\sigma^2}$ are 0.08 and 0.05, respectively. Fig. 14 shows that the degradation trends and residual terms of the complete transformed dataset closely align with the original true observations. Regarding the non-equally spaced data at 70 °C, whose missing data rate of only 50% according to Table 6, and the proposed model maintains its expected performance under these conditions. However, for the 80 °C dataset, the missing data rate reaches 80%, fortunately the availability of sufficient data from the benchmark stress level at 60 °C and the transformed data at 70 °C surpasses the cases discussed in Chapters 2 and 3. Consequently, the proposed method maintains a high accuracy due to the abundance of valid data for model training. This exemplifies the excellent performance of the proposed LSSVM-RBF model in handling the imputation issue of unequal interval data from step-stress accelerated degradation test. Furthermore, the complete dataset with equal interval of SLD degradation can be more suitable for prediction after undergoing some other data preprocessing method such as supplying the affection of temperature drift.

## 5. Conclusions and future work

Incomplete data is the one of the most significant problems in current data analysis research. Even though it is difficult to solve this problem completely, researchers are constantly working to propose more appropriate methods for specific data missing scenarios. The aim of this study is to propose a hybrid imputation method that combines the LSSVM model with RBF neural networks for the purpose of addressing the potential issues related to missing data in step-stress accelerated degradation tests.

**Table 5**
Average degradation rate of SLD at each stress level.

| Stress Level | 60 °C | 70 °C | 80 °C |
|---|---|---|---|
| **ADR** | 0.00052 | 0.0012 | 0.0032 |

Then the acceleration factors at higher stress levels relative to 60 °C can be obtained as Eq. (15).

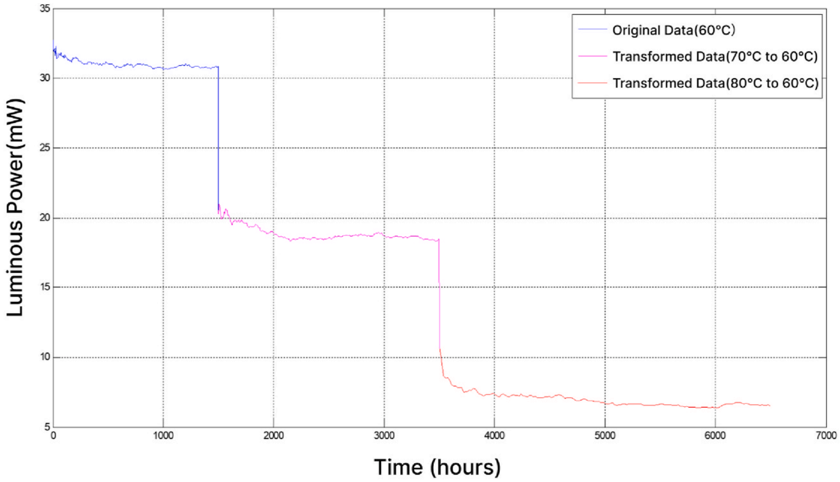



**Fig. 13.** Transformed degradation data at 60 °C of SLD output optical power.

**Table 6**
Imputation data information.

| Stress Level | Data Interval | | Data Amount | | | Percentage of Imputation data |
|---|---|---|---|---|---|---|
| | Transformed (hours) | Imputation (hours) | Transformed | Imputation | Total | |
| 70 °C | 6 | 3 | 334 | 333 | 667 | 49.9% |
| 80 °C | 18 | 3 | 167 | 830 | 997 | 83.2% |

In this paper, a two-step framework for hybrid imputation method is constructed. Firstly, we separate the missing data into trend and residual terms for individual imputation value, the non-linear approximation capability of LSSVM is leveraged to modeling the degradation trend, capturing its underlying patterns. Then, the RBF model is utilized to further extract data dispersion information in the residuals of the LSSVM, which ensures the imputation of missing data maintains a similar level of dispersion as the original observed data. The imputation strategy for accelerated degradation data with unequal intervals are improved by considering the acceleration factor, and the performance and validity of the proposed model are verified using the realistic test data for SLDs. The findings of the current study and future work direction are presented as follow.

- According to the obtained results, missing data will always lead to a decrease in the precision of evaluation or prediction models. Specifically, missing data occurring in the middle part of dataset has the least impact, although it may render some models unusable. On the other hand, missing data in the back part tends to have the most severe impacts on the performance of the models.
- Comparison results reveals that in the scenarios of missing data with equal measuring interval, the hybrid models generally outperform the traditional models in terms of imputation performance, particularly at high missing data rates. However, it should be noted that the traditional regression model excels in reproducing the original data dispersion. Besides, the proposed model demonstrates significantly better performance compared to the other models for missing data with unequal measuring intervals.
- To optimize the proposed model, hybrid model integrated by the other intelligent learning methods could be also examined.
- This paper investigates strategies for resolving missing data from accelerated tests with degradation trends, further research should focus on the missing data on the success/failure type of accelerated test which does not have a degradation trend.

## Data availability statement

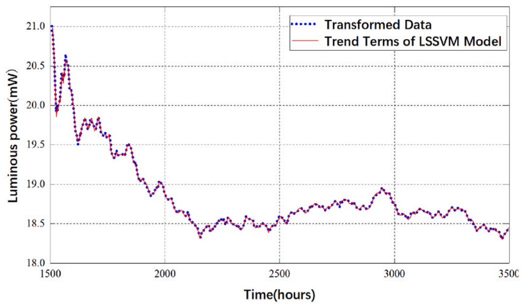
Data will be made available on request.

## CRediT authorship contribution statement

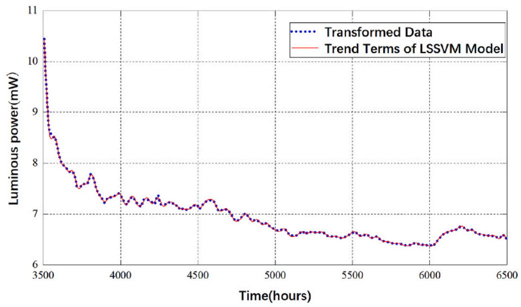
**Yaqiu Li:** Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft. **Qijie Zhou:** Formal analysis, Validation, Writing – original draft. **Ye Fan:** Data curation, Formal analysis. **Guangze Pan:** Conceptualization, Methodology, Resources, Writing – review & editing, Validation. **Zongbei Dai:** Investigation, Resources. **Baimao Lei:** Data curation, Formal analysis.

## Declaration of competing interest

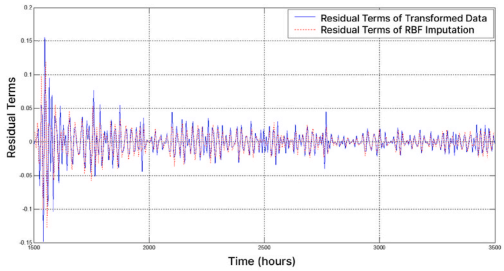
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to



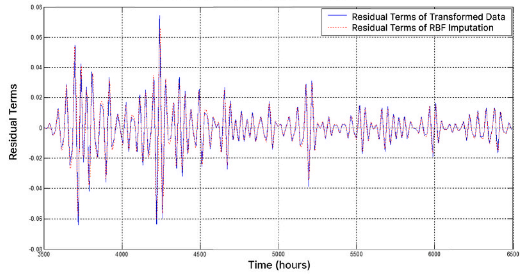

a) Trend term of imputation based on the LSSVM model for transformed data (70°C)



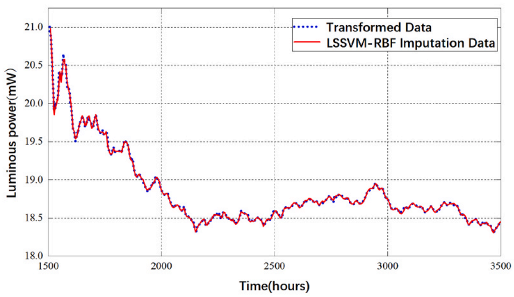

b) Trend term of imputation based on the LSSVM model for transformed data (80°C)



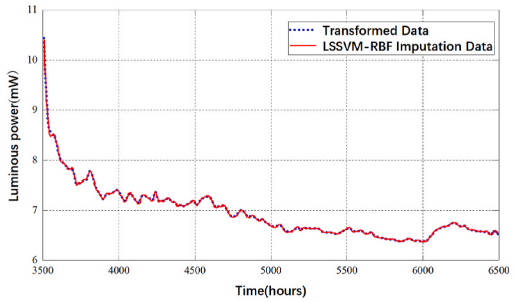

c) Residual term of imputation based on the RBF model for transformed data (70°C)




d) Residual term of imputation based on the RBF model for transformed data (80°C)




e) Complete degradation dataset for transformed data (70°C)




f) Complete degradation dataset for transformed data (80°C)

**Fig. 14.** Imputation results for the data transformed from 70 °C to 80 °C.

influence the work reported in this paper.

## Acknowledgements

This work was partially funded by the Guangdong Basic and Applied Basic Research Foundation (2021A1515110679) and the science and technology program of Guangzhou, China (No.202201010303 and No. 202201010584), the ministry of industry and information technology project (TC210804R-1).

## References

[1] W. Si, Y. Shao, W. Wei, Accelerated degradation testing with long-term memory effects, Ieee T Reliab 69 (2020) 1254–1266, https://doi.org/10.1109/TR.2020.2997404.

[2] M. LuValle, A theoretical framework for accelerated testing, in: N. Limnios, M. Nikulin (Eds.), Recent Advances in Reliability Theory: Methodology, Practice, and Inference, Birkhäuser, Boston, MA, 2000, pp. 419–433, https://doi.org/10.1007/978-1-4612-1384-0_27.

[3] Z. Wang, L. Zhao, Z. Kong, J. Yu, C. Yan, Development of accelerated reliability test cycle for electric drive system based on vehicle operating data, Eng. Fail. Anal. 141 (2022) 106696, https://doi.org/10.1016/j.engfailanal.2022.106696.

[4] Z. Ma, S. Wang, C. Ruiz, C. Zhang, H. Liao, E. Pohl, Reliability estimation from two types of accelerated testing data considering measurement error, Reliab. Eng. Syst. Saf. 193 (2020) 106610, https://doi.org/10.1016/j.ress.2019.106610.

[5] B. Mehmood, M. Akbar, R. Ullah, Accelerated aging effect on high temperature vulcanized silicone rubber composites under DC voltage with controlled environmental conditions, Eng. Fail. Anal. 118 (2020) 104870, https://doi.org/10.1016/j.engfailanal.2020.104870.

[6] A. Ngueilbaye, H. Wang, D.A. Mahamat, S.B. Junaidu, Modulo 9 model-based learning for missing data imputation, Appl. Soft Comput. 103 (2021) 107167, https://doi.org/10.1016/j.asoc.2021.107167.

[7] D.F. Heitjan, S. Basu, Distinguishing "missing at random" and "missing completely at random,", Am. Statistician 50 (1996) 207–213, https://doi.org/10.1080/00031305.1996.10474381.

[8] R.L. Brown, Efficacy of the indirect approach for estimating structural equation models with missing data: a comparison of five methods, Struct. Equ. Model. 1 (1994) 287–316, https://doi.org/10.1080/10705519409539983.
[9] S. Prasad, An exponential imputation in the case of missing data, J. Stat. Manag. Syst. 20 (2017) 1127–1140, https://doi.org/10.1080/09720510.2017.1407515.
[10] A.I. Al-Omari, C.N. Bouza, C. Herrera, Imputation methods of missing data for estimating the population mean using simple random sampling with known correlation coefficient, Qual. Quantity 47 (2013) 353–365, https://doi.org/10.1007/s11135-011-9522-1.
[11] J. Lin, N. Li, M.A. Alam, Y. Ma, Data-driven missing data imputation in cluster monitoring system based on deep neural network, Appl. Intell. 50 (2020) 860–877, https://doi.org/10.1007/s10489-019-01560-y.
[12] B.T. Keller, H. Du, A fully conditional specification approach to multilevel multiple imputation with latent cluster means, Multivariate Behav. Res. 54 (2019) 149–150, https://doi.org/10.1080/00273171.2018.1556085.
[13] R. Andridge, L. Bechtel, K.J. Thompson, Finding a flexible hot-deck imputation method for multinomial data, J Surv Stat Methodol 9 (2021) 789–809, https://doi.org/10.1093/jssam/smaa005.
[14] S. Faisal, G. Tutz, Multiple imputation using nearest neighbor methods, Inf. Sci. 570 (2021) 500–516, https://doi.org/10.1016/j.ins.2021.04.009.
[15] M. Quartagno, J.R. Carpenter, Multiple imputation for discrete data: evaluation of the joint latent normal model, Biom. J. 61 (2019) 1003–1019, https://doi.org/10.1002/bimj.201800222.
[16] D. Nd, R. Mm, Missing value imputation using stratified supervised learning for cardiovascular data, J Inform Data Min 1 (2016) 1–10, https://doi.org/10.4172/2229-8711.S1113.
[17] C. Gautam, V. Ravi, Data imputation via evolutionary computation, clustering and a neural network, Neurocomputing 156 (2015) 134–142, https://doi.org/10.1016/j.neucom.2014.12.073.
[18] K.J. Nishanth, V. Ravi, Probabilistic neural network based categorical data imputation, Neurocomputing 218 (2016) 17–25, https://doi.org/10.1016/j.neucom.2016.08.044.
[19] J. Shao, W. Meng, G. Sun, Evaluation of missing value imputation methods for wireless soil datasets, Personal Ubiquitous Comput. 21 (2017) 113–123, https://doi.org/10.1007/s00779-016-0978-9.
[20] G. Sharma, A. Panwar, I. Nasiruddin, R.C. Bansal, Non-linear LS-SVM with RBF-kernel-based approach for AGC of multi-area energy systems, IET Generation, Transm. Distrib. 12 (2018) 3510–3517, https://doi.org/10.1049/iet-gtd.2017.1402.
[21] G. Sharma, I. Nasiruddin, K.R. Niazi, R.C. Bansal, Automatic generation control (AGC) of wind power system: an least squares-support vector machine (LS-SVM) radial basis function (RBF) kernel approach, Elec. Power Compon. Syst. 46 (2018) 1621–1633, https://doi.org/10.1080/15325008.2018.1511003.
[22] C. Liu, P. Niu, G. Li, X. You, Y. Ma, W. Zhang, A hybrid heat rate forecasting model using optimized LSSVM based on improved GSA, Neural Process. Lett. 45 (2017) 299–318, https://doi.org/10.1007/s11063-016-9523-0.
[23] C. Cornelis, R. Jensen, G. Hurtado, D. Śle¸zak, Attribute selection with fuzzy decision reducts, Inf. Sci. 180 (2010) 209–224, https://doi.org/10.1016/j.ins.2009.09.008.
[24] S.A. Sarra, S. Cogar, An examination of evaluation algorithms for the RBF method, Eng. Anal. Bound. Elem. 75 (2017) 36–45, https://doi.org/10.1016/j.enganabound.2016.11.006.
[25] J.A.K. Suykens, J. De Brabanter, L. Lukas, J. Vandewalle, Weighted least squares support vector machines: robustness and sparse approximation, Neurocomputing 48 (2002) 85–105, https://doi.org/10.1016/S0925-2312(01)00644-0.
[26] Y. Zhang, R. Li, Short term wind energy prediction model based on data decomposition and optimized LSSVM, Sustain Energy Techn 52 (2022) 102025, https://doi.org/10.1016/j.seta.2022.102025.