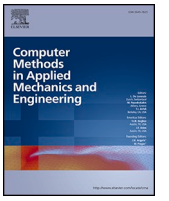
Computer Methods in Applied Mechanics and Engineering 423 (2024) 116868

Contents lists available at ScienceDirect

Computer Methods in Applied Mechanics and Engineering

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



Machine learning aided uncertainty quantification for engineering structures involving material-geometric randomness and data imperfection

Qihan Wang [a], Di Wu [b], Guoyin Li [c], Zhenyu Liu [d], Jingzhong Tong [e], Xiaojun Chen [a], Wei Gao [a,*]

[a] Centre for Infrastructure Engineering and Safety, School of Civil and Environmental Engineering, The University of New South Wales, Sydney, NSW 2052, Australia
[b] School of Civil and Environmental Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia
[c] School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW, Australia
[d] State Key Lab of CAD&CG, School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China
[e] College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China

HIGHLIGHTS

- A machine learning-aided uncertainty quantification framework is proposed for engineering structures.
- The effects of material and geometric randomness on structural performance are quantified simultaneously.
- Data imperfections, i.e., noise and outliers within observations, are considered within the proposed framework.
- A novel machine learning technique is developed to handle the datasets with imperfection.
- The applicability and computational efficiency of the proposed approach are well demonstrated.

ARTICLE INFO

Keywords:
Material and geometric uncertainty
Data imperfection
Capped extended support vector regression
Machine learning
Engineering application

ABSTRACT

In real-world engineering, uncertainty is ubiquitous within material properties, structural geometry, load conditions, and the like. These uncertainties have substantial impacts on the estimation of structural performance. Furthermore, information or datasets in real life commonly contain imperfections, e.g., noise, outliers, or missing data. To quantify these impacts induced by uncertainties on structural behaviours and reduce the effects of data imperfections simultaneously, a machine learning-aided stochastic analysis framework is proposed. A novel supervised machine learning technique, namely the Capped Extended Support Vector Regression (CX-SVR) technique, is developed to effectively suppress the effects of outliers and noise in datasets. Its inherent convexity in optimization and capped strategy theoretically supports the accuracy of CX-SVR, especially in handling datasets with imperfections. Once the effective surrogate model is established, subsequent analyses, like sampling-based methods, can circumvent the cumbersome physical model, which is potentially the nest of computational burden and errors in engineering applications. The high robustness of the proposed approach can be summarized in four main aspects: unrestrictive selection of the system inputs and their statistical information, 'perfect' or 'imperfect' data, enough statistical information (including statistical moments, probability

* Corresponding author.
E-mail address: w.gao@unsw.edu.au (W. Gao).

https://doi.org/10.1016/j.cma.2024.116868
Received 18 April 2023; Received in revised form 12 November 2023; Accepted 18 February 2024
Available online 26 February 2024

density functions, and cumulative distribution functions) of the system outputs, and physical problems from various engineering fields.

1. Introduction

The blossom of technology facilitates heterogeneous access in engineering practices to tremendous data of high diversity [1–3]. However, uncertainty exists within almost any information [4–7]. Appropriate consideration of the system uncertainty and relevant quantification is always an essential task for engineering applications.

In addition to the inherent uncertainty, it is undeniable that real-world data collection systems are imperfect [8–10]. These imperfections may present as wrong data, missing data, outliers in data, lack of variability in data, etc., and indeed the datasets that appear perfect may contain noise [8]. Moreover, data imperfections can also occur in numerical simulation due to errors in every step during the numerical simulation. An ill-posed or ill-defined initial condition may result in simulation results that are completely different from the actual physical system being simulated. Generally, the sources of the data imperfection in numerical simulation results can be summarized as inaccurate or incomplete input data, improper simulation setup, mathematical modelling errors, numerical errors, inappropriate simulation assumptions, human errors, software limitations, etc. Regardless of various sources from real-world engineering or numerical simulation, inaccurate, incomplete, and even contradictory data may lead to wrong estimations without appropriate consideration.

Engineering mechanic problems by considering the continuous random variation of system properties do not yield easy analytical solutions. The challenge to the analyst is formidable, even for the most simplified structural configurations. Thus, to seek possible access to tackle such sophisticated problems, over the past decades, uncertainty quantification has been continuously investigated without any rupture [11–13]. The stochastic finite element method (SFEM) is developed to initially estimate adequate information about the statistical moments (e.g., means, standard deviations) of the concerned structural response [14,15]. Then, SFEM has been successfully applied to a wide variety of problems, such as solid, structural, and fluid mechanics, acoustics, heat transfer, etc. [16–22]. Several variants of the SFEM have been developed and three of these are the most used and accepted: the Monte Carlo simulation (MCS) method [23,24], the perturbation method [25,26], and the spectral stochastic finite element method (SSFEM) [27,28]. Each method adopts a different approach to represent, solve, and study the randomness of the system [29]. The MCS method is the most general and straightforward approach for SFEM, while tremendous computational power is required to achieve credible estimations. Another widely applied branch of the SFEM is the perturbation method [30–33]. Within the framework of the perturbation method, randomness is introduced into the system via Taylor series expansions, and the accuracy of the perturbation method increases with the number of terms used to calculate the response variables. Due to the computational costs, the perturbation method is widely used to obtain the means and covariance, yet rarely high-order moments, of the structural responses. The method is also limited to the values of random variables that do not exhibit large variations. Furthermore, the SSFEM, mainly concerned with representing the random material properties of a structure [29], uses spectral methods, such as the Karhunen-Loève (KL) expansion or polynomial chaos expansion (PCE), to reduce the computational power used in other methodologies such as MCS [34–38]. The investigation of system uncertainty has been developed for decades, while the consideration of data imperfections is familiar for data analysts [39–41], yet merely in structural analysis.

To conquer such a real-life engineering-stimulated challenge, a novel machine learning-aided strategy is proposed to provide a possible solution. Effective surrogate model construction from imperfection-involved datasets can be one of the primary issues. Accordingly, a novel supervised machine learning algorithm, namely the Capped Extended Support Vector Regression (CX-SVR) technique, is developed. By succeeding the merits of the Extended Support Vector Regression (X-SVR) technique [7,42,43], the proposed CX-SVR technique can be formulated by solving a convex optimization problem, which means the optimal solution can be theoretically guaranteed. Then, by integrating with the capped strategy [44,45], the outliers in training datasets can be removed and the noise can be suppressed effectively. This feature greatly benefits the proposed framework in handling datasets with imperfections, e.g., missing data, outliers, noise, etc. Furthermore, the kernelized strategy extends the proposed CX-SVR technique to tackle nonlinear problems.

The established surrogate model alternatively describes the relationship, which used to be underpinned, implicit, and sophisticated in most engineering applications, between the system uncertainties and the structural responses of interest in an explicit mathematical expression. High computational efficiency and explicit mathematical expression of the established surrogate model greatly benefit the subsequent sampling-based analysis, sensitivity analysis, optimization programming, etc. Another significant advantage of the proposed scheme is that once the surrogate model has been established, the subsequent analyses can circumvent the intricate calculation process of the original physical model, and correspondingly avoid the potential for resulting errors and additional computational burdens. In addition, the proposed framework can serve with high robustness, mainly in four aspects: unrestrictive selection of the system inputs and corresponding statistical information (e.g., statistical moments, distribution types), ‘perfect’ or ‘imperfect’ system outputs, enough statistical information, involving not only the mean, standard deviation, but also probability density function (PDF), cumulative distribution function (CDF), of the concerned system outputs, and physical problems from various engineering fields. Convincingly, the proposed stochastic uncertainty quantification strategy integrating material, geometric randomness, and data imperfections, in conjunction with the newly developed regression technique can greatly benefit real-world engineering applications, over the stages of design, analysis, service life, maintenance, and even recycling.

The remainder of this manuscript is organized as follows. The stochastic uncertainty quantification involving material, geometric

randomness, and data imperfections is introduced in Section 2. In Section 3, the algorithms of the newly proposed regression technique are thoroughly presented. Then, Section 4 illustrates the proposed machine learning-aided generalized stochastic uncertainty quantification framework. To demonstrate the applicability and computational efficiency of the proposed approach, two engineering-stimulated applications: fracture analysis for a holed plate and bandgap analysis for a 3D lattice-based elastic metamaterial (EMM) are thoroughly investigated in Section 5. At last, some conclusions are drawn in Section 6.

2. Stochastic uncertainty quantification involving material-geometric randomness and data imperfection

2.1. System uncertainty: material and geometric randomness

Given a complete probability space (Ω, Ξ, P), it is characterized by the sample space Ω, σ-algebra of events Ξ, and the probability measure $P : F \rightarrow [0, 1]$, $P(\Omega) = 1$. Then, a physical problem is defined on a random domain $\Omega(\boldsymbol{\xi}^{\mathrm{R}})$, in which $\boldsymbol{\xi}^{\mathrm{R}}$ refers to a finite set of uncorrelated random variables with known probability distributions,

$$\boldsymbol{\xi}^{\mathrm{R}} \in \Omega := \left\{ \boldsymbol{\xi}^{\mathrm{R}} \in \Re^{n} \middle| \xi_j^{\mathrm{R}} \sim f_{\xi_j^{\mathrm{R}}}(x), \text{for } j = 1, 2, ..., n \right\} \tag{1}$$

where n denotes the dimension of the random variables and $f_{\xi_j^{\mathrm{R}}}(x)$ denotes the PDF of the jth random variable ξ_j^{R}. In this research, the random variables $\boldsymbol{\xi}^{\mathrm{R}}$ constitute a parameterization of the material properties and geometry, simultaneously.

Considering a mechanical system whose structural performance can be modelled by a set of governing equations, typically partial differential equations (PDEs), and utilizing some suitable solution scheme, the computational model can be generally expressed as,

$$\mathbf{y} = \mathbf{F}(\mathbf{x}) \tag{2}$$

where $\mathbf{x}$ denotes a vector of input parameters of the model. These parameters can be related to the system geometry, material constitutive behaviour or the applied loading conditions. $\mathbf{y}$ denotes the vector of response of interest which may generally involve the displacement, strain, stress, spatial, temporal variations, or their associated components. The computational model, $\mathbf{F}$, in real-world engineering applications, is commonly sophisticated, and cannot be modified by the analyst but only run for a given set of input parameters.

Taking the stochastic dynamic analysis as an example, the governing equations are established as,

$$\mathbf{M}(\boldsymbol{\xi}^{\mathrm{R}})\ddot{\mathbf{U}}^{\mathrm{R}} + \mathbf{C}(\boldsymbol{\xi}^{\mathrm{R}})\dot{\mathbf{U}}^{\mathrm{R}} + \mathbf{K}(\boldsymbol{\xi}^{\mathrm{R}})\mathbf{U}^{\mathrm{R}} = \mathbf{P}(\boldsymbol{\xi}^{\mathrm{R}}) \tag{3}$$

where $\mathbf{M}(\boldsymbol{\xi}^{\mathrm{R}})$, $\mathbf{C}(\boldsymbol{\xi}^{\mathrm{R}})$, and $\mathbf{K}(\boldsymbol{\xi}^{\mathrm{R}}) \in \Re^{D_{\mathrm{f}} \times D_{\mathrm{f}}}$ denote the random mass, damping, and stiffness matrices, respectively. $\mathbf{U}^{\mathrm{R}}$, $\dot{\mathbf{U}}^{\mathrm{R}}$, and $\ddot{\mathbf{U}}^{\mathrm{R}} \in \Re^{D_{\mathrm{f}}}$ denotes the random displacement, velocity, and acceleration vectors, respectively. $\mathbf{P}(\boldsymbol{\xi}^{\mathrm{R}}) \in \Re^{D_{\mathrm{f}}}$ denotes the random load vector. D_{f} denotes the degree of freedom of the structure. By involving the random inputs, Eq. (3) is a stochastic structural dynamic problem. To the best knowledge of the authors, there are no theoretical supports or algorithms to directly solve it. Theoretically, there are infinite sets of possible realizations of the random variables, which can lead to infinite calculations. Thus, it is computationally intractable to solve Eq. (3) for all possible solutions. Instead, the statistical characteristics, e.g., the statistical moments (means, standard deviations), PDF, and CDF, of the structural responses appear more meaningful in engineering applications.

The geometric uncertainty that cooperates with the material uncertainty may lead the system to be a mesh-varying random system, of higher chaotic performance. The response sensitivity concerning shape variables is acknowledged to be more difficult to compute [46,47]. Thus, the underpinned relationship from geometry and material properties to structural response is often more challenging to describe than a single source of uncertainty. Since the generation of the mesh for random geometry is out of the scope of this research, an automatic mesh generation strategy is adopted [48].

2.2. Data imperfection: imperfection of the system output datasets

In practice, most of the datasets possess imperfections because real-world data collection systems are imperfect [49]. It may be inaccurate, incomplete, and possibly contradictory as obtained from a variety of sources, which leads to wrong results. Thus, imperfect data is a generic problem, in which information extraction and decision-making are difficult tasks. The components that could be considered in discussing imperfect datasets are shown in Fig. 1.

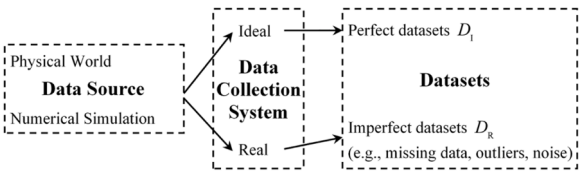


Fig. 1. Ideal versus real data collection systems.

In Fig. 1, the data source shown as the leftmost block includes the physical world and numerical simulation. This data source feeds two parallel paths. The ideal path as the upper one leads from the data source, through an ideal data acquisition system, into perfect (i.e., error-free) datasets D_{I}. The lower path leads from the same data source through a real data acquisition system, generating the real datasets D_{R}, which is imperfect. Imperfections in the datasets corresponding to differences between D_{I} and D_{R} can be divided into five kinds, as summarized in Appendix A.

Throughout this research, simple missing data, coded missing data, and disguised missing data are treated as a manner of gross errors. Gross errors, noisy data, or outliers refer to the data being considered meaningless, due to the existence of too much variation. The missing data or corrupt data, which refers to any data that is not machine-readable, are also treated as outliers. The term 'noise' refers to the unexplained variability within a data sample, which is often considered as random data. Correspondingly, some common sources of data imperfections in engineering applications can be summarized: (1) Measurement errors: errors can occur due to inaccuracies in measuring devices or human errors during data collection. (2) Incomplete data capture: data may be incomplete due to issues such as system failure or power outages. (3) Recording errors: errors can occur during data recording, such as typos or incorrect units of measurement. (4) Data inconsistencies: data may be inconsistent due to differences in data sources, changes in measurement techniques, theorems, assumptions, calculation platforms, etc. Moreover, improper data processing, noise in the data collection system, etc., may also lead to data imperfections. These data imperfections degrade the quality of the data and subsequently affect the decision-making to various extents. By taking these data imperfections into account, the original uncertainty quantification problem would become of higher applicability in real-world engineering.

3. Capped extended support vector regression (CX-SVR)

To investigate the effects of the material and geometric randomness for real-life engineering applications where data are often collected with imperfections, a machine learning-aided uncertainty quantification strategy is proposed. Within the proposed framework, a supervised machine learning technique, namely the Capped Extended Support Vector Regression (CX-SVR) technique, is newly developed to generate an effective surrogate model based on datasets with imperfections.

3.1. Linear capped extended support vector regression (CX-SVR)

Given the training datasets with inputs $\mathbf{A} = [\mathbf{x}_1,\ \mathbf{x}_2,\ ...,\ \mathbf{x}_i,\ ...,\ \mathbf{x}_m]^{\mathrm{T}} \in \Re^{m\times n}$ and output $\mathbf{y} = [y_1,\ y_2,\ ...,\ y_i,\ ...,y_m]^{\mathrm{T}} \in \Re^{m}$, the targeted hyperplane model is defined as,

$$\widehat{f}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} - \delta \tag{4}$$

where $\mathbf{w} = [w_1,\ w_2,\ ...,\ w_j,\ ...,\ w_n]^{\mathrm{T}} \in \Re^{n}$ and $\delta \in \Re$ denote the normal to the hyperplane and bias, respectively. m and n denote the number of training samples and the dimension of the inputs, respectively. Then, by implementing the ε-insensitive loss function and elastic-net penalty which contains both L_1 and L_2-norms penalty, the linear regression function can be established by solving the optimization problem in the form of,

$$\begin{aligned} &\min_{\mathbf{w},\delta,\boldsymbol{\xi},\boldsymbol{\xi}^*} : \frac{1}{2}\| \mathbf{w} \|_2^2 + \lambda \| \mathbf{w} \|_1 + \frac{c}{2}\left(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi} + \boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{\xi}^*\right) \\ &\text{s.t.} \begin{cases} \mathbf{A}\mathbf{w} - \delta\mathbf{e}_m - \mathbf{y} \le \varepsilon\mathbf{e}_m + \boldsymbol{\xi} \\ \mathbf{y} - \mathbf{A}\mathbf{w} + \delta\mathbf{e}_m \le \varepsilon\mathbf{e}_m + \boldsymbol{\xi}^* \\ \boldsymbol{\xi}, \boldsymbol{\xi}^* \ge \mathbf{0}_m \end{cases} \end{aligned} \tag{5}$$

where λ and c denote two tuning parameters, which balance the L_1 and L_2-norms of $\mathbf{w}$, and the flatness of hyperplane and the amount up to which deviations larger than the tolerance ε, respectively. ε denotes the tolerable deviation between the observations $\mathbf{y}$ and model prediction $\widehat{\mathbf{y}} = \widehat{f}(\mathbf{x})$. $\boldsymbol{\xi} = [\xi_1,\ \xi_2,\ ...,\ \xi_i,\ ...,\ \xi_m]^{\mathrm{T}} \in \Re^m$ and $\boldsymbol{\xi}^* = [\xi_1^*,\ \xi_2^*,\ ...,\ \xi_i^*,...,\ \xi_m^*]^{\mathrm{T}} \in \Re^m$ denote two non-negative vectors which collect slack variables ξ_i and ξ_i^*. These slack variables are used to allow certain constraints to be violated. $\mathbf{e}_m$ and $\mathbf{0}_m \in \Re^m$ denote ones and zeros vectors in the dimension of m, respectively.

Then, a decomposition strategy is used to eliminate the computation of the L_1-norm of $\mathbf{w}$. Two non-negative variables $\mathbf{p}$ and $\mathbf{q} \in \Re^n$ are defined in the form of,

$$p_j := \left(w_j\right)_+ = \begin{cases} 0, w_j \le 0 \\ w_j, w_j > 0 \end{cases} \textit{ and } q_j := \left(w_j\right)_- = \begin{cases} -w_j, w_j < 0 \\ 0, w_j \ge 0 \end{cases}, \textit{ for } j = 1, 2, ..., n \tag{6}$$

It is indicated by the definition in Eq. (6) that $w_j = p_j - q_j$ and $p_j q_j = 0$ can be promised $\forall j$. Thus, the computation of L_1 and L_2-norms of $\mathbf{w}$ can be alternatively calculated as,

$$\begin{aligned} \| \mathbf{w} \|_1 &= |w_1| + |w_2| + ... + |w_n| \\ &= p_1 + q_1 + p_2 + q_2 + ... + p_n + q_n \\ &= \mathbf{e}_n^{\mathrm{T}}(\mathbf{p} + \mathbf{q}) \end{aligned} \quad \textit{and} \quad \begin{aligned} \| \mathbf{w} \|_2^2 &= \| \mathbf{p} - \mathbf{q} \|_2^2 \\ &= \| \mathbf{p} \|_2^2 + \| \mathbf{q} \|_2^2 - 2\mathbf{p}^{\mathrm{T}}\mathbf{q} \\ &= \| \mathbf{p} \|_2^2 + \| \mathbf{q} \|_2^2 \end{aligned} \tag{7}$$

Subsequently, the optimization problem in Eq. (5) can be simplified as,

$$\min_{\mathbf{p},\mathbf{q},\delta,\boldsymbol{\xi},\boldsymbol{\xi}^*} : \frac{1}{2}\left(\| \mathbf{p} \|_2^2 + \| \mathbf{q} \|_2^2\right) + \lambda \mathbf{e}_n^{\mathrm{T}}(\mathbf{p}+\mathbf{q}) + \frac{c}{2}\left(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi} + \boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{\xi}^*\right)$$
$$\text{s.t.} \begin{cases} \mathbf{A}(\mathbf{p}-\mathbf{q}) - \delta\mathbf{e}_m - \mathbf{y} \le \varepsilon\mathbf{e}_m + \boldsymbol{\xi} \\ \mathbf{y} - \mathbf{A}(\mathbf{p}-\mathbf{q}) + \delta\mathbf{e}_m \le \varepsilon\mathbf{e}_m + \boldsymbol{\xi}^* \\ \mathbf{p}, \mathbf{q} \ge \mathbf{0}_n; \boldsymbol{\xi}, \boldsymbol{\xi}^* \ge \mathbf{0}_m \end{cases} \tag{8}$$

Till now, the original optimization problem of the Extended Support Vector Regression (X-SVR) technique [7,42,43] has been achieved. However, a common challenge in real-life engineering is noisy data with extreme outliers, as presented in Fig. 2.

When there is noise or outliers in training datasets, squared L_2-norm distance used in X-SVR will exaggerate the effects of outliers. Consequently, the construction of optimum regression hyperplanes would be significantly affected. Inspired by the capped strategy used in the L_1-norm support vector machine [44,45], a novel variation of the X-SVR technique, namely the Capped Extended Support Vector Regression (CX-SVR) is introduced.

Then, the problem in Eq. (8) is reformulated as the following optimization problem,

$$\min_{\mathbf{p}_{(t)},\mathbf{q}_{(t)},\delta_{(t)},\boldsymbol{\xi}_{(t)},\boldsymbol{\xi}^*_{(t)}} : \frac{1}{2}\left(\| \mathbf{p}_{(t)} \|_2^2 + \| \mathbf{q}_{(t)} \|_2^2\right) + \lambda \mathbf{e}_n^{\mathrm{T}}\left(\mathbf{p}_{(t)} + \mathbf{q}_{(t)}\right) + \frac{c}{2}\left(\boldsymbol{\xi}_{(t)}^{\mathrm{T}}\mathbf{D}_{(t)}\boldsymbol{\xi}_{(t)} + \boldsymbol{\xi}_{(t)}^{*\mathrm{T}}\mathbf{D}_{(t)}^*\boldsymbol{\xi}_{(t)}^*\right)$$
$$\text{s.t.} \begin{cases} \mathbf{x}\left(\mathbf{p}_{(t)} - \mathbf{q}_{(t)}\right) - \delta_{(t)}\mathbf{e}_m - \mathbf{y} \le \varepsilon\mathbf{e}_m + \boldsymbol{\xi}_{(t)} \\ \mathbf{y} - \mathbf{x}\left(\mathbf{p}_{(t)} - \mathbf{q}_{(t)}\right) + \delta_{(t)}\mathbf{e}_m \le \varepsilon\mathbf{e}_m + \boldsymbol{\xi}_{(t)}^* \\ \mathbf{p}_{(t)}, \mathbf{q}_{(t)} \ge \mathbf{0}_n; \boldsymbol{\xi}_{(t)}, \boldsymbol{\xi}_{(t)} \ge \mathbf{0}_m \end{cases} \tag{9}$$

where the subscript (t) denotes the tth iteration. The matrices $\mathbf{D}_{(t)}$ and $\mathbf{D}_{(t)}^*$ are used to 'discard' outliers or noisy data points. When the points exceed the soft margin and are statistically far from the hyperplane, they will be suspected as outliers and their contributions in regression will be eliminated or degraded. More specifically, two diagonal matrices $\mathbf{D}_{(t)}$ and $\mathbf{D}_{(t)}^* \in \Re^{m\times m}$ contain diagonal elements as $d_{(t),\,i}$ and $d^*_{(t),\,i}$ (for $i = 1,\ 2,\ \ldots,\ m$) in the tth iteration. For the first iteration, i.e., $t = 1$, $\mathbf{D}_{(1)}$ and $\mathbf{D}_{(1)}^*$ are initialized as two identity matrices $\mathbf{I}_m \in \Re^{m\times m}$, which means that $\forall i$, $d_{(t),\,i} = d^*_{(t),\,i} = 1$. For later iterations, i.e., $t > 1$,

$$d_{(t+1),\,i} = \begin{cases} smallval,\ if\ suspected\ outlier \\ 1\ or\ \dfrac{1}{\left|\xi_{(t),\,i}\right|},\ otherwise \end{cases} \tag{10}$$

$$d^*_{(t+1),\,i} = \begin{cases} smallval,\ if\ suspected\ outlier \\ 1\ or\ \dfrac{1}{\left|\xi^*_{(t),\,i}\right|},\ otherwise \end{cases} \tag{11}$$

in which 'smallval' denotes a small constant, and the two most straightforward options built-in CX-SVR to determine its value are given: (1) set to a small constant, e.g., $1e^{-5}$, (2) weaken the weights by a multiplier less than 1, e.g., $d_{(t+1),\,i} = 0.5 \cdot d_{(t),\,i}$. Moreover, for those unsuspected points, it should be mentioned that by replacing the weights with the L_1-norm of $\xi_{(t),\,i}$ or $\xi^*_{(t),\,i}$, the objective function

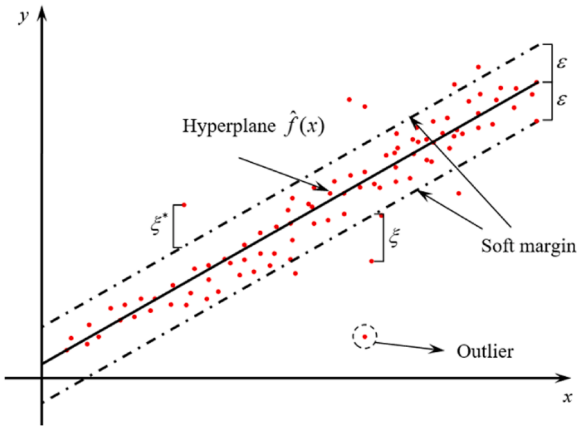


Fig. 2. Diagram of a one-dimensional linear regression model.

in Eq. (10) can be regarded as the L_1-norm distance. Therefore, the proposed CX-SVR technique can customize the empirical risk in L_1-norm or L_2-norm.

As for the criteria for outliers, within the proposed CX-SVR technique, there are two main built-in manners, including z-score [51] and quartile analyses [52]. The z-score is often used to measure the variance of an observation from the mean in terms of standard deviation, assuming a normal distribution [51]. Accordingly, the z-score of the ith point in the tth iteration can be calculated in the form of

$$z_{(t),\,i} = \frac{\xi_{(t),\,i} - \mu\left(\boldsymbol{\xi}_{(t)}\right)}{\sigma\left(\boldsymbol{\xi}_{(t)}\right)} \tag{12}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation of $(\cdot)$, respectively. The observations with greater absolute z-scores are marked as outliers. Referring to the 'three-sigma rule' or 68-95-99.7 rule [53], approximately 68% of the values lie within 1 standard deviation from the mean, 95% of values are within 2 standard deviations, and 99.7% are within 3. The observations with greater absolute z-scores are marked as outliers, and the commonly adopted threshold τ of z-score in CX-SVR included 2.5, 3, and 3.5. Mathematically, it is expressed that if

$$\left|z_{(t),\,i}\right| > \tau \tag{13}$$

then the ith observation is suspected to be an outlier.

Another built-in strategy is the quartile analysis [52]. The interquartile range in the ith iteration ($IQR_{(t)}$) is preliminarily defined as,

$$IQR_{(t)} = Q_3\left(\boldsymbol{\xi}_{(t)}\right) - Q_1\left(\boldsymbol{\xi}_{(t)}\right) \tag{14}$$

where $Q_1(\cdot)$ and $Q_3(\cdot)$ denote the first quartile and third quartile, corresponding to the 25th and 75th percentiles of the datasets $(\cdot)$, respectively. The values falling outside the $k\times IQR_{(t)}$ are suspected to be outliers. A commonly used value for the multiplier k is 1.5, which is also adopted in CX-SVR. Hence, the criteria by quartile analysis for suspected outliers is that if

$$\xi_{(t),\,i} < Q_1\left(\boldsymbol{\xi}_{(t)}\right) - k \times IQR_{(t)} \;or\; \xi_{(t),\,i} > Q_3\left(\boldsymbol{\xi}_{(t)}\right) + k \times IQR_{(t)} \tag{15}$$

then the ith observation is suspected to be an outlier.

Generally, the tolerance of soft margin ε and coefficient τ for z-score analysis or k for quartile analysis, control the criterion of whether an observation is suspected or unsuspected to be an outlier or a noisy point. Moreover, not limited to these two methods, other outlier detection techniques, such as the clustering approach, isolation forest method, etc., can also be integrated.

Then, for each iteration, the established optimization problem of the CX-SVR method in Eq. (9) is alternatively expressed to achieve a simplified formulation,

$$\begin{aligned} &\min_{\widehat{\mathbf{z}}_{(t)},\delta_{(t)}} : \frac{1}{2}\left(\widehat{\mathbf{z}}_{(t)}^{\mathrm{T}}\widehat{\mathbf{C}}\widehat{\mathbf{z}}_{(t)} + \delta_{(t)}^2\right) + \lambda\widehat{\mathbf{a}}^{\mathrm{T}}\widehat{\mathbf{z}}_{(t)} \\ &\text{s.t.}(\widehat{\mathbf{A}} + \mathbf{I}_{2m+2n})\widehat{\mathbf{z}}_{(t)} + \left(\varepsilon\mathbf{I}_{2m+2n} + \delta_{(t)}^2\widehat{\mathbf{G}}\right)\widehat{\mathbf{b}} + \widehat{\mathbf{d}} \geq \mathbf{0}_{2m+2n} \end{aligned} \tag{16}$$

where the matrices $\widehat{\mathbf{C}}$, $\widehat{\mathbf{G}}$, and $\widehat{\mathbf{A}} \in \Re^{(2n+2m)\times(2n+2m)}$ are defined as,

$$\widehat{\mathbf{C}} = \begin{bmatrix} \mathbf{I}_{2n} & & \\ & c\mathbf{D}_{(t)} & \\ & & c\mathbf{D}_{(t)}^* \end{bmatrix}, \widehat{\mathbf{G}} = \begin{bmatrix} \mathbf{0}_{2n\times 2n} & \mathbf{0}_{2n\times m} & \mathbf{0}_{2n\times m} \\ \mathbf{0}_{m\times 2n} & \mathbf{I}_m & \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times 2n} & \mathbf{0}_{m\times m} & -\mathbf{I}_m \end{bmatrix}, \widehat{\mathbf{A}} = \begin{bmatrix} \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times 2m} \\ -\mathbf{A} & \mathbf{A} & \mathbf{0}_{m\times 2m} \\ \mathbf{A} & -\mathbf{A} & \mathbf{0}_{m\times 2m} \end{bmatrix} \tag{17}$$

and the vectors $\widehat{\mathbf{a}}$, $\widehat{\mathbf{b}}$, $\widehat{\mathbf{d}}$, and $\widehat{\mathbf{z}}_{(t)} \in \Re^{2n+2m}$ are defined as:

$$\widehat{\mathbf{a}} = \begin{bmatrix} \mathbf{e}_n \\ \mathbf{e}_n \\ \mathbf{0}_{2m} \end{bmatrix}, \widehat{\mathbf{b}} = \begin{bmatrix} \mathbf{0}_{2n} \\ \mathbf{e}_m \\ \mathbf{e}_m \end{bmatrix}, \widehat{\mathbf{d}} = \begin{bmatrix} \mathbf{0}_{2n} \\ \mathbf{y} \\ -\mathbf{y} \end{bmatrix}, \widehat{\mathbf{z}}_{(t)} = \begin{bmatrix} \mathbf{p}_{(t)} \\ \mathbf{q}_{(t)} \\ \boldsymbol{\xi}_{(t)} \\ \boldsymbol{\xi}_{(t)}^* \end{bmatrix} \tag{18}$$

The square of the bias parameter δ^2 is added to the objective function to provide the benefits of optimizing the orientation and location of the hyperplane simultaneously. In addition, the constraints that $\mathbf{p}_{(t)}$ and $\mathbf{q}_{(t)}$ are non-negative in Eq. (9) have been reinforced by Eq. (16). Alternatively, by using the Lagrange method with the Karush-Kuhn and Tucker (KKT) condition [50], the problem in Eq. (16) can be solved through its dual formulation, which can be expressed in the form of,

$$\min_{\boldsymbol{\varphi}_{(t)}} : \frac{1}{2}\boldsymbol{\varphi}_{(t)}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\varphi}_{(t)} - \mathbf{m}^{\mathrm{T}}\boldsymbol{\varphi}_{(t)} \quad \text{s.t.} \boldsymbol{\varphi}_{(t)} \geq \mathbf{0}_{2m+2n} \tag{19}$$

where $\boldsymbol{\varphi}_{(t)} \in \Re^{2n+2m}$ denotes the Lagrange multiplier vector for the tth iteration; the matrices $\mathbf{Q} \in \Re^{(2n+2m)\times(2n+2m)}$ and $\mathbf{m} \in \Re^{2n+2m}$ are defined as,

$$\mathbf{Q} = (\widehat{\mathbf{A}} + \mathbf{I}_{2m+2n})\widehat{\mathbf{C}}^{-1}(\widehat{\mathbf{A}} + \mathbf{I}_{2m+2n})^{\mathrm{T}} + \widehat{\mathbf{G}}\widehat{\mathbf{b}}\widehat{\mathbf{b}}^{\mathrm{T}}\widehat{\mathbf{G}} \tag{20}$$

$$\mathbf{m} = \lambda(\widehat{\mathbf{A}} + \mathbf{I}_{2m+2n})\widehat{\mathbf{C}}^{-1}\widehat{\mathbf{a}} - \varepsilon\widehat{\mathbf{b}} - \widehat{\mathbf{d}} \tag{21}$$

It can be noticed the constraints are significantly simplified into purely non-negative constraints for the optimization variable $\boldsymbol{\varphi}_{(t)}$. Moreover, the optimization problem in Eq. (19) possesses a quadratic objective and affine inequality constraints, which is a common standard form of the quadratic programming problem. As a convex optimization problem, the global optimal solution can be theoretically guaranteed for the established optimization problem.

Therefore, the global optimum of the proposed CX-SVR technique can be efficiently obtained by solving the associated dual problem through any available quadratic programming solver. Let $\boldsymbol{\varphi}^* \in \Re^{2n+2m}$ be the solution of Eq. (19), then the variables $\widehat{\mathbf{z}}_{(t)}$ and $\delta_{(t)}$ can be calculated as,

$$\widehat{\mathbf{z}}_{(t)} = \widehat{\mathbf{C}}^{-1}\left[(\widehat{\mathbf{A}} + \mathbf{I}_{2m+2n})^{\mathrm{T}}\boldsymbol{\varphi}^* - \lambda\widehat{\mathbf{a}}\right] \tag{22}$$

$$\delta_{(t)} = \widehat{\mathbf{b}}^{\mathrm{T}}\widehat{\mathbf{G}}\boldsymbol{\varphi}^* \tag{23}$$

The coefficients $\mathbf{w}_{(t)}$, $\boldsymbol{\xi}_{(t)}$, and $\boldsymbol{\xi}^*_{(t)}$ can be calculated as,

$$\mathbf{w}_{(t)} = \mathbf{p}_{(t)} - \mathbf{q}_{(t)} = \widehat{\mathbf{z}}_{(t)}(1:n) - \widehat{\mathbf{z}}_{(t)}((n+1):2n) \tag{24}$$

$$\boldsymbol{\xi}_{(t)} = \widehat{\mathbf{z}}_{(t)}((2n+1):(2n+m)) \tag{25}$$

$$\boldsymbol{\xi}^*_{(t)} = \widehat{\mathbf{z}}_{(t)}((2n+m+1):(2n+2m)) \tag{26}$$

The iterations can be stopped for a maximum number of iterations $\bar{t}$ or according to convergence study with acceptable tolerance. Let $\mathbf{p}^*$, $\mathbf{q}^*$, and δ^* be the solutions after iterations, the linear regression function by the proposed CX-SVR technique can be obtained in the form of,

$$\widehat{f}(\mathbf{x}) = \mathbf{x}^{\mathbf{T}}(\mathbf{p}^* - \mathbf{q}^*) - \delta^* \tag{27}$$

3.2. Kernelized capped extended support vector regression (CX-SVR)

Through the kernel mapping strategy, the proposed linear CX-SVR technique can be extended to tackle nonlinear problems. The input datasets $\mathbf{x}$ would be transferred from the low-dimension input space to a higher-dimension Euclidian space or even infinite-dimension Hilbert space through the mapping function $\phi(\mathbf{x})$. The adopted empirical kernelization can be expressed as,

$$\mathbf{x}_i = \left[x_{i,1},\ x_{i,2},\ \ldots,\ x_{i,n}\right]^{\mathrm{T}} \mapsto \kappa(\mathbf{x},\ \mathbf{x}_i) = \begin{bmatrix} \phi(\mathbf{x}_1)\cdot\phi(\mathbf{x}_i) \\ \phi(\mathbf{x}_1)\cdot\phi(\mathbf{x}_i) \\ \vdots \\ \phi(\mathbf{x}_m)\cdot\phi(\mathbf{x}_i) \end{bmatrix}, \text{for } i = 1,\ 2,\ \ldots,\ m \tag{28}$$

Thus, for an arbitrary set of training inputs and given kernel function, the kernelized input datasets $\boldsymbol{\kappa}$ can be expressed as,

$$\begin{aligned} \boldsymbol{\kappa} &= [\kappa(\mathbf{x},\ \mathbf{x}_1),\ \kappa(\mathbf{x},\ \mathbf{x}_2),\ \ldots,\ \kappa(\mathbf{x},\ \mathbf{x}_m)] \\ &= \begin{bmatrix} \phi(\mathbf{x}_1)\cdot\phi(\mathbf{x}_1) & \phi(\mathbf{x}_1)\cdot\phi(\mathbf{x}_2) & \cdots & \phi(\mathbf{x}_1)\cdot\phi(\mathbf{x}_m) \\ \phi(\mathbf{x}_1)\cdot\phi(\mathbf{x}_1) & \phi(\mathbf{x}_2)\cdot\phi(\mathbf{x}_2) & \cdots & \phi(\mathbf{x}_2)\cdot\phi(\mathbf{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_m)\cdot\phi(\mathbf{x}_1) & \phi(\mathbf{x}_m)\cdot\phi(\mathbf{x}_2) & \cdots & \phi(\mathbf{x}_m)\cdot\phi(\mathbf{x}_m) \end{bmatrix} \in \Re^{m\times m} \end{aligned} \tag{29}$$

Equivalently, the kernelized nonlinear CX-SVR technique is solved by its dual formulation by using the Lagrange method with the KKT condition. Thus, by replacing the input datasets with the kernelized input datasets, the optimization problem in Eq. (19) can be alternatively expressed as,

$$\min_{\boldsymbol{\varphi}_{(t)}} : \frac{1}{2}\boldsymbol{\varphi}_{(t)}^{\mathrm{T}}\mathbf{Q}_{\kappa}\boldsymbol{\varphi}_{(t)} - \mathbf{m}_{\kappa}^{\mathrm{T}}\boldsymbol{\varphi}_{(t)} \quad \text{s.t.} \boldsymbol{\varphi}_{(t)} \geq \mathbf{0}_{2m+2n} \tag{30}$$

where $\mathbf{Q}_{\kappa}$ and $\mathbf{m}_{\kappa}$ are calculated from the kernelized input datasets, more specifically,

$$\mathbf{Q}_{\kappa} = (\widehat{\mathbf{A}}_{\kappa} + \mathbf{I}_{2m+2n})\widehat{\mathbf{C}}^{-1}(\widehat{\mathbf{A}}_{\kappa} + \mathbf{I}_{2m+2n})^{\mathrm{T}} + \widehat{\mathbf{G}}\widehat{\mathbf{b}}\widehat{\mathbf{b}}^{\mathrm{T}}\widehat{\mathbf{G}} \tag{31}$$

$$\mathbf{m}_{\kappa} = \lambda\left(\widehat{\mathbf{A}}_{\kappa} + \mathbf{I}_{(2m+2n)\times(2m+2n)}\right)\widehat{\mathbf{C}}^{-1}\widehat{\mathbf{a}} - \varepsilon\widehat{\mathbf{b}} - \widehat{\mathbf{d}} \tag{32}$$

in which $\widehat{\mathbf{A}}_{\kappa}$ is defined as,

$$\widehat{\mathbf{A}}_{\kappa} = \begin{bmatrix} \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times 2m} \\ -\boldsymbol{\kappa} & \boldsymbol{\kappa} & \mathbf{0}_{m\times 2m} \\ \boldsymbol{\kappa} & -\boldsymbol{\kappa} & \mathbf{0}_{m\times 2m} \end{bmatrix} \tag{33}$$

It is worth mentioning that the kernelized input matrix $\boldsymbol{\kappa}$ is in the dimension of $m \times m$. Thus, in kernelized CX-SVR, the dimension of input datasets equals the number of observations, i.e., $n = m$. Moreover, the convexity feature of the optimization program is still satisfactory in the kernelized CX-SVR technique.

After the construction of the kernelized CX-SVR prototype model, the capped strategy can be implemented on the established kernelized prototype model to remove extreme outliers and suppress the effect of the noise data. Eventually, by obtaining the solutions of the variables in the targeted hyperplane (i.e., $\mathbf{p}^*$, $\mathbf{q}^*$, and δ^*), the kernelized CX-SVR model can be formulated as,

$$\widehat{f}(\mathbf{x}) = \kappa(\mathbf{x}_{\text{train}}, \mathbf{x})(\mathbf{p}^* - \mathbf{q}^*) - \delta^* \tag{34}$$

The subscript 'train' is added in case of misunderstanding.

The proposed CX-SVR successes the kernel functions used in the X-SVR. Some kernel functions commonly used are summarized in Appendix B. Moreover, other Mercer's kernels can also be easily embedded into the proposed CX-SVR technique. The cross-validation strategy and Bayesian hyperparameter tuning are integrated within the proposed CX-SVR technique to avoid over-fitting and fulfil the feature of auto-tuning, respectively. As they are secondary contributions to this research, the detailed algorithms can be referred to in the works [54–57].

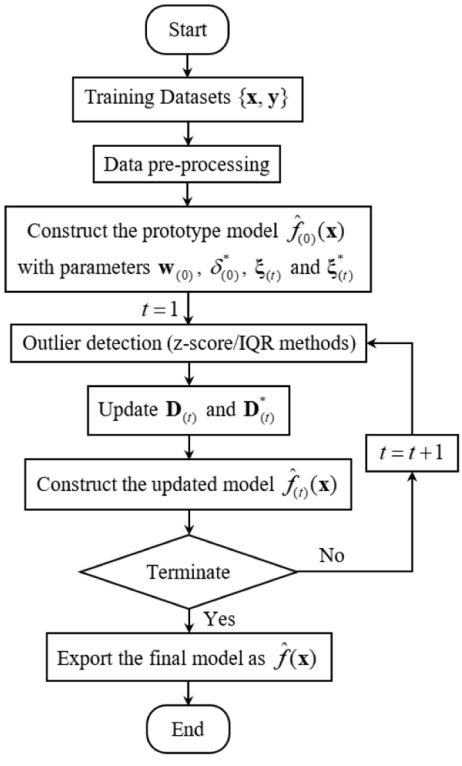


Fig. 3. The flowchart of the proposed CX-SVR technique.

3.3. Flowchart of the kernelized cx-svr technique

To more effectively illustrate the proposed kernelized CX-SVR technique, the detailed program is presented in the flowchart in Fig. 3.

From Fig. 3, the program of the proposed CX-SVR technique is well-demonstrated. Various data pre-processing techniques can be easily embedded into the proposed CX-SVR technique, such as the dimensionality reduction (e.g., principal component analysis), normalization method, clustering method, etc. These data pre-processing techniques can be used in a targeted manner according to the characteristics of the data or requirements of the tasks. Then, the CX-SVR prototype model is constructed, which is equivalent to the X-SVR model. Within this step, the cross-validation and Bayesian hyperparameter tuning would be integrated to avoid overfitting and fulfil the feature of hyperparameter auto-tuning, respectively. In addition to three hyperparameters in the linear CX-SVR model, the number of additional hyperparameters by introducing a specific kernel function into the kernelized CX-SVR model can vary from 0 to a large integral correspondingly.

Within the loop of the capped strategy, the outliers are detected, and for those suspected outliers, the corresponding components in the matrices $\mathbf{D}_{(t)}$ and $\mathbf{D}^{*}_{(t)}$ would be adjusted to eliminate or suppress the effects of the outliers and noise data. Then, the CX-SVR model would be reconstructed, until the termination condition is met. The terminate condition in this research is set by the maximum iteration number.

4. Stochastic uncertainty quantification strategy through the CX-SVR technique

A machine learning-aided uncertainty quantification strategy is proposed herein for engineering structures involving material, geometrical uncertainty, and data imperfection. As a data-driven method, the proposed scheme requires a sufficient supply of training datasets. Before the presentation of the proposed uncertainty quantification approach, the method in numerical simulation to generate the training datasets by implementing the brute Monte Carlo simulation (MCS) on the finite element analysis (FEA) model is presented as a flowchart in Fig. 4.

The consideration of both material and geometric uncertainty simultaneously leads to a mesh-varying random system. Quantifying the probabilistic performance of such a sophisticated system possesses several challenges:

(1) The re-meshing process would aggravate the computational costs for each calculation.
(2) The re-meshing process may lead to computational error, and then the system output datasets would be presented with imperfections.
(3) The underpinned relationship between the system geometric variables and the structural response normally is more difficult to depict, in comparison to the randomness in the material.
(4) The mesh-varying random system would present a more chaotic performance.
(5) Based on the proposed CX-SVR technique, the proposed generalized uncertainty quantification framework is presented in Fig. 5.

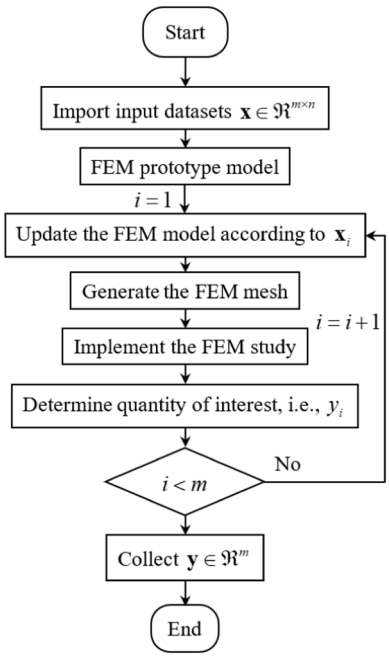


Fig. 4. The flowchart of the MCS on FEM.

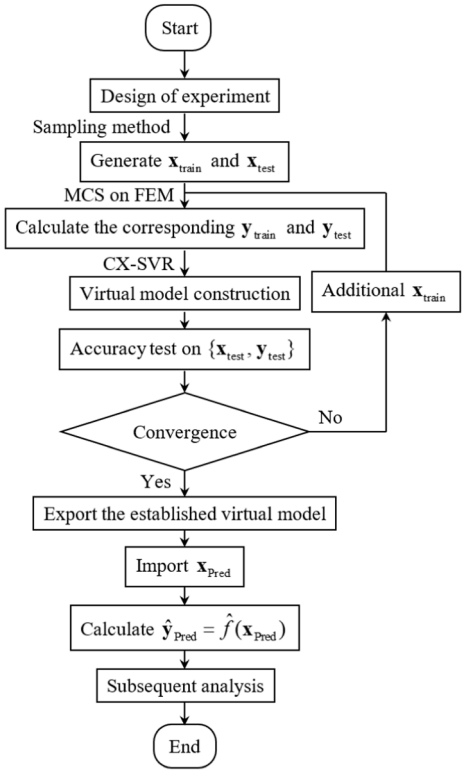


Fig. 5. The flowchart of the generalized uncertainty quantification through the proposed CX-SVR technique.

Fig. 5 illustrates the processes of the proposed machine learning-aided generalized uncertainty quantification framework. There are heterogeneous accesses to generate the database, such as historical records, information, communication techniques (sensors, monitors, actuators, etc.), experiments and so on. However, in this research, the database involving the training and testing datasets is generated by implementing the brute MCS method on the FEM model, as presented in Fig. 4. To consider the data imperfections in real-world engineering, the data imperfections, more specifically, noisy data and outliers, are added to the generated training datasets in Fig. 5, to test the performance of the developed CX-SVR technique in handling these out-of-system factors.

The calculation process in FEA has been inherently underpinned within the generated surrogate model. Thus, the MCS can be implemented on the established surrogate model in a much more efficient manner, instead of repetitively running the cumbersome FDA, as well as the re-meshing processes. Moreover, the computational errors or unstable results induced by the re-meshing process or FEA study can be effectively avoided.

High compatibility of the proposed framework should also be emphasized: various statistical information (e.g., means, standard deviations, distribution types, etc.) of the system inputs can be considered; there is no obvious limitation on the selection of the quantities of system inputs and outputs; a variety of physical problems from interdisciplinary or multidisciplinary fields can be investigated; multiple data pre-processing technique or machine learning techniques can be embedded or added; and a sufficient amount of statistical information (the statistical moments, PDF, CDF, etc.) of the concerned structural response can be effectively estimated. In addition, different to the previous generation of uncertainty quantification strategies, the proposed machine learning-aided generalized scheme possesses an inherent feature of information update. With the established surrogate model, the predictions on the newly collected system inputs following the updated statistical information can be easily achieved, without re-running the physical simulation.

5. Numerical investigation

To demonstrate the applicability of the proposed machine learning-aided uncertainty quantification strategy for problems with material-geometric randomness and data imperfections, two engineering applications have been fully investigated. In Section 5.1, stochastic brittle fracture analysis for a holed plate, involving structural material nonlinearity, material, and geometric randomness, and further considering the training output datasets with outliers, is thoroughly studied. Correspondingly, some statistical information on the FEM mesh of the holed plate presents the feature of randomness, which is also discussed in the same section. Then, the investigation of the probabilistic bandgap characteristics for a 3D elastic metamaterial (EMM) is implemented on 'perfect' (error-free) training datasets and training datasets with outliers and noise.

5.1. Brittle fracture analysis for a holed plate

The brittle fracture analysis for a holed plate made of a cement mortar is investigated herein by considering the system uncertainty within geometric and material properties simultaneously. The experimental specimen, presented in Fig. 6(a) and (b), is the prototype of this numerical investigation. The setup of the model is based on experimental data from previous research [58]. Loading is applied through displacement-controlled metal pins inserted into the two smaller holes. As the plate is loaded, a mixed-mode fracture is induced with a crack propagating from the predefined notch to the unsymmetrically placed hole in the centre of the plate. Fracture is modelled using a damage model that regularizes the sharp geometry of the crack by the phase field approximation [59].

The width and height of the plate are 65 and 120 mm, respectively. A plane stress condition is assumed, and the thickness of the plate is a unit constant. Two small holes for upper and lower pins are located 20 mm to the left, upper, and bottom edges, and the radius is 10 mm. Some other geometric characteristics of the holed plate, including the features of the initial notch (the height h_n, width w_n, and location l_n) and the middle hole (the radius r_h and location of the centre x_h and y_h), are marked in Fig. 6(d).

To properly resolve the phase field and achieve stable material behaviour, a high mesh density is required in the vicinity of the propagating crack. Eventually, the holed plate is discretised into 11,907 triangular elements, as shown in Fig. 6(c). The convergence study for the FEM mesh is implemented based on the deterministic model. For the deterministic model, the crack trajectory can be estimated as shown in Fig. 6(d), through the phase field damage method.

The considered system uncertainty contains geometric parameters (h_n, w_n, l_n, r_h, x_h, and y_h), and material parameters (Young's modulus E, Poisson's ratio ν, and density ρ). For surrogate model construction, the generated training input datasets are expected to be more evenly distributed, and thus, the Latin hypercube sampling (LHS) method is adopted to generate the training input datasets uniformly distributed within the bounds (lower bound i.e., LB, and upper bound i.e., UB), as summarized in Table 1.

By considering the system uncertainty within the material and geometric parameters, the whole system of the holed plate possesses the feature of randomness. This feature of randomness is presented in the domain, material resistance, and fracture performance of the holed plate. Through the brute MCS with 1e3 iterations, the probabilistic fracture performance of the holed plate during the loading process has been depicted in Fig. 7.

From Fig. 7, it can be noticed that the whole curves possess the feature of randomness. Moreover, the critical values (locally maximum) would be achieved at various loading steps under the probabilistic problem. Then, the random critical loads (marked as red points) at regions I and II (namely $P_{cr,\,1}$ and $P_{cr,\,2}$, respectively) are investigated. Under the same system uncertainty during the loading process of the holed plate, $P_{cr,\,1}$ presents a larger deviation in the load axis, while $P_{cr,\,2}$ in the displacement axis.

Through the proposed CX-SVR technique, the convergence study is implemented to determine the size of training samples for surrogate model construction for $P_{cr,\,1}$. R^2 and *RMSE* are estimated. The initial training size is set as 40 and increases to 500 gradually. For each size of the training sample, the surrogate model construction is repetitively implemented 20 times. The computational results are depicted in Fig. 8.

From Fig. 8, the proposed CX-SVR technique presents sharp convergence trends for both R^2 and *RMSE* in constructing the surrogate

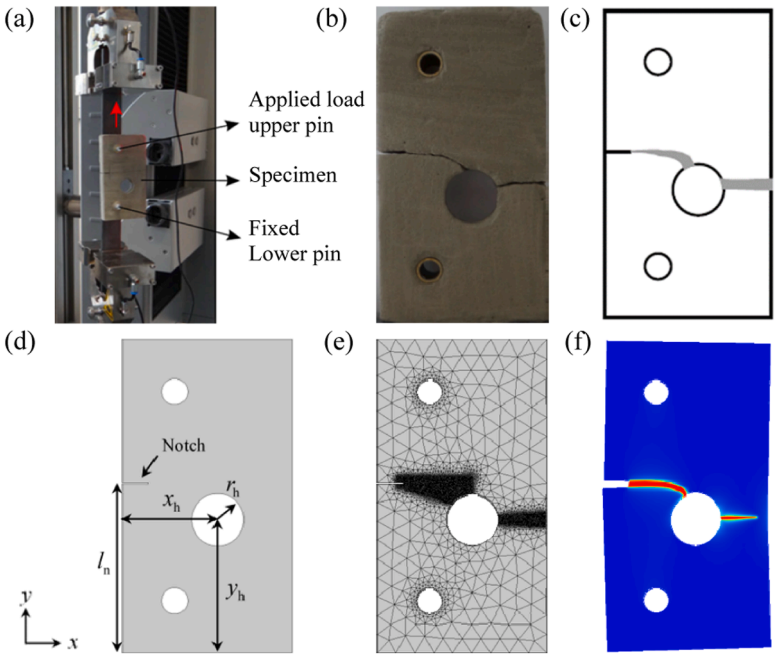


Fig. 6. The holed plate: (a) Experiment setup, (b) fractured specimen (c) experimentally observed crack patterns [58], (d) numerical model, (e) the adopted FEM mesh, and (f) the estimated crack phase field at the last parameter step for the deterministic model.

Table 1
The statistical information of system uncertainties for the plate.

Geometric Property of the Notch Bounds	h_n (mm)	w_n (mm)	l_n (mm)
LB	0.49	9.9	64.7
UB	0.51	10.1	65.3
Geometric Property of the Central Hole **Bounds**	r_h (mm)	x_h (mm)	y_h (mm)
LB	9.9	36.3	50.7
UB	10.1	36.7	51.3
Material Property **Bounds**	ρ (kg /m^3)	E (GPa)	ν
LB	1900	5.7	0.19
UB	2100	6.3	0.21

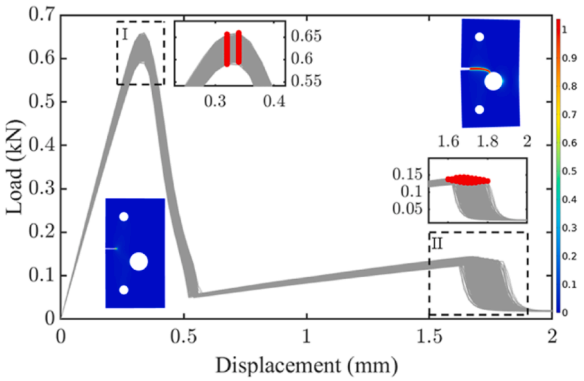


Fig. 7. The probabilistic load versus displacement curves.

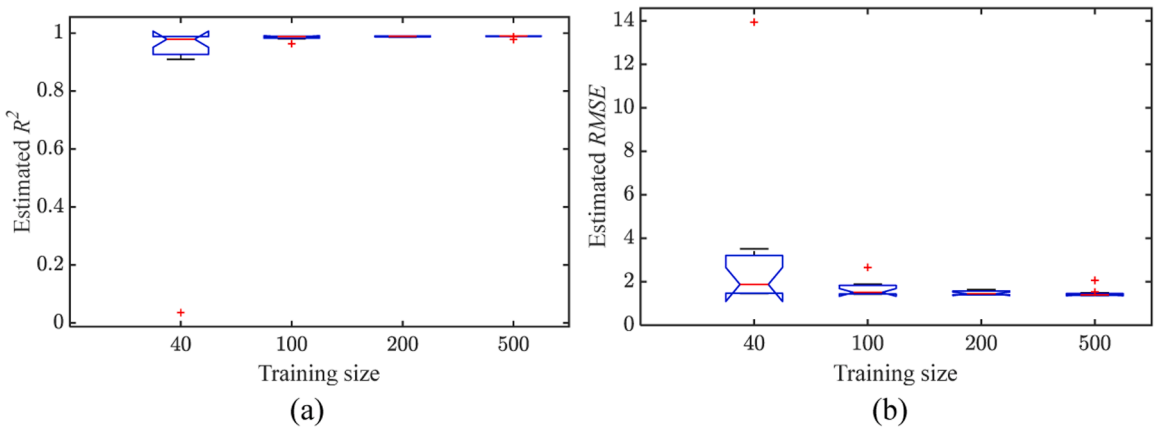


Fig. 8. The boxplots of the (a) estimated R^2 and (b) *RMSE* for $P_{cr,1}$ by CX-SVR.

model for $P_{cr,1}$. More specifically, when the size of the training samples reaches 200, the convergence trends can be captured for both estimation metrics. Similarly, 200 training samples are also determined to construct the surrogate model for $P_{cr,2}$. Then, the PDFs, CDFs and *RE*s of CDFs of $P_{cr,1}$ and $P_{cr,2}$ are estimated based on the established CX-SVR models, and the computational results are presented in Fig. 9. The brute MCS results with 1e3 iterations are considered as the benchmark.

From Fig. 9, it can be demonstrated that the proposed CX-SVR technique is capable of estimating PDFs and CDFs of both $P_{cr,1}$ and $P_{cr,2}$ with nearly overlapped estimations in reference to the brute MCS results. The same conclusions can be drawn from the scatter subplots, R^2 and *RMSE*-values annotated in Fig. 9. By further referring to *RE*s of the CDFs of $P_{cr,1}$ and $P_{cr,2}$, the maximum magnitude of *RE* is kept lower than 0.4% through the proposed CX-SVR technique. It is worth mentioning that when the CDF goes to 0, the reference

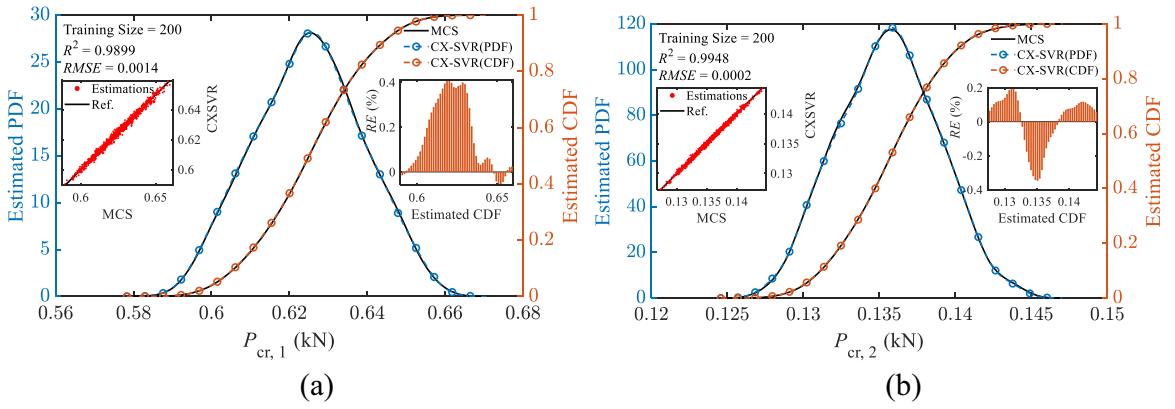


Fig. 9. Estimated PDFs and CDFs of (a) $P_{cr,1}$ and (b) $P_{cr,2}$ by CX-SVR.

and estimation are both extremely small, in such cases, *RE*-value may lose the effects to reflect the performance of the model. Thus, we calculate the *RE* on the (1-CDF) instead of CDF directly when the CDF is lower than 0.5. Such calculation enhances the interpretability of the statistics. In addition to the whole PDFs and CDFs of the concerned structural responses, the local probabilistic information can be obtained through the proposed approach without additional computational efforts. The estimated probabilistic information at specific thresholds is calculated and summarized in Table 2.

From Table 2, the specific statistical information of $P_{cr,1}$ and $P_{cr,2}$ by considering the system material and geometrical uncertainties can be effectively estimated through the proposed strategy. In reference to the brute MCS results in 1e3 iterations, the proposed approach estimates the statistical information of $P_{cr,1}$ and $P_{cr,2}$ with high accuracy.

In addition to computational effectiveness, supreme computational efficiency should be highlighted. All computation is implemented in the workstation equipped with Intel(R) Xeon(R) CPU E5-2667 v4 @ 3.20GHz. The brute MCS costs 29.82 days to complete 1e3 iterations in total. However, the proposed approach only uses 9.30 min to construct the surrogate model from 200 training datasets, which requires approximately 5.96 days to generate, and subsequent predictions for 1e3 samples on the established surrogate model only cost several milliseconds. Thus, it can be concluded that the main computational costs of the proposed approach are occupied by the generation of training datasets.

To demonstrate the applicability of the proposed approach, especially the developed CX-SVR technique in the regression based on the training datasets involving some imperfections, a small group of observations is replaced with corrupt data. Assuming the error rates of 15 and 30%, which correspond to 30 and 60 out of 200 training datasets respectively, are randomly sampled and replaced by zeros. Then, the surrogate models are re-constructed on these 'imperfect' training datasets. Within the developed CX-SVR technique, a default setting of z-score criterion is adopted with τ set to 0.25 and 'smallval' set to $1e^{-5}$, respectively. In comparison, three currently popular machine learning techniques, including traditional support vector regression (SVR), neural network (NN), and Gaussian process regression (GPR), are implemented on the same training datasets. The computational results are summarized in Table 3.

From Table 3, it can be found that the proposed CX-SVR embedded uncertainty quantification scheme possesses high robustness and accuracy in handling outliers within the training output datasets attributed to data missing. At the error rate reaches 15%, the R^2-values are well maintained by the proposed CX-SVR technique as 0.9880 and 0.9947 for $P_{cr,1}$ and $P_{cr,2}$, respectively. Then, at the error rate of 30%, the proposed CX-SVR technique still provides outstanding estimations, with the corresponding R^2-values of 0.9896 and 0.9854, respectively. The robust performance of the proposed CX-SVR technique in handling outliers in training output datasets overpasses other machine learning techniques. All other machine learning techniques share sharp decrease trends in the performance

Table 2
Estimated statistical information of $P_{cr,1}$ and $P_{cr,2}$.

Property	MCS	CX-SVR	*RE* (%)
$\mu_{P_{cr,1}}$ (kN)	0.624954	0.624877	-0.012310
$\sigma_{P_{cr,1}}$ (kN)	0.013641	0.013662	0.149378
$\Pr(\mu_{P_{cr,1}} - \sigma_{P_{cr,1}} \le P_{cr,1} < \mu_{P_{cr,1}} + \sigma_{P_{cr,1}})$	0.642684	0.641221	-0.227652
$\Pr(\mu_{P_{cr,1}} - 2\sigma_{P_{cr,1}} \le P_{cr,1} < \mu_{P_{cr,1}} + 2\sigma_{P_{cr,1}})$	0.952177	0.951720	-0.047934
$\Pr(\mu_{P_{cr,1}} - 3\sigma_{P_{cr,1}} \le P_{cr,1} < \mu_{P_{cr,1}} + 3\sigma_{P_{cr,1}})$	0.999753	0.999817	0.006417
Property	MCS	CX-SVR	*RE* (%)
$\mu_{P_{cr,2}}$ (kN)	0.135581	0.135579	-0.001414
$\sigma_{P_{cr,2}}$ (kN)	0.003233	0.003228	-0.139544
$\Pr(\mu_{P_{cr,2}} - \sigma_{P_{cr,2}} \le P_{cr,2} < \mu_{P_{cr,2}} + \sigma_{P_{cr,2}})$	0.643009	0.642526	-0.075082
$\Pr(\mu_{P_{cr,2}} - 2\sigma_{P_{cr,2}} \le P_{cr,2} < \mu_{P_{cr,2}} + 2\sigma_{P_{cr,2}})$	0.951854	0.951659	-0.020511
$\Pr(\mu_{P_{cr,2}} - 3\sigma_{P_{cr,2}} \le P_{cr,2} < \mu_{P_{cr,2}} + 3\sigma_{P_{cr,2}})$	0.999184	0.999278	0.009445

Table 3
The performance of the surrogate models when the training datasets of $P_{cr,\,1}$ and $P_{cr,\,2}$ involving outliers.

Property	Error Rate (%)	Method	R^2	*RoI* (%)	*RMSE*	*RoI* (%)
$P_{cr,\,1}$	15	CX-SVR	0.988049	–	1.496457	–
		X-SVR	0.961234	-2.7139	2.403453	60.6096
		SVR	0.957710	-3.0706	2.600261	73.7612
		NN	0.729719	-26.1455	7.606788	408.3199
		GPR	0.956936	-3.1489	2.509407	67.6899
	30	CX-SVR	0.989554	–	1.387826	–
		X-SVR	0.869798	-12.1020	3.939729	183.8777
		SVR	0.903976	-8.6481	3.602139	159.5526
		NN	0.728991	-26.3314	7.361715	430.4494
		GPR	0.843900	-14.7192	4.335969	212.4289
$P_{cr,\,2}$	15	CX-SVR	0.994697	–	0.234141	–
		X-SVR	0.976883	-1.7909	0.442432	88.9596
		SVR	0.969103	-2.5730	0.509113	117.4386
		NN	0.812464	-18.3205	1.573142	571.8781
		GPR	0.968854	-2.5981	0.509113	117.4386
	30	CX-SVR	0.985362	–	0.394155	–
		X-SVR	0.895324	-9.1376	0.883066	124.0403
		SVR	0.925949	-6.0296	0.753647	91.2057
		NN	0.739145	-24.9875	1.959176	397.0572
		GPR	0.871664	-11.5387	0.941319	138.8195

*The ratio of improvement index (*RoI*) is defined as $RoI=Q_A-Q_B/Q_B\times100\%$, where Q denotes the estimation metrics (e.g., R^2 and *RMSE*), and the subscripts A and B denote the methods of interest and the reference, respectively. Here, the CX-SVR technique is considered as the reference.

of surrogate model construction for both $P_{cr,\,1}$ and $P_{cr,\,2}$, when the error rate increases from 15% to 30%.

Generally, the computational effectiveness and efficiency of the proposed CX-SVR aided scheme, as well as the high robustness and computational stableness, in tackling probabilistic brittle fracture problems involving material, geometric randomness, and data imperfections, have been thoroughly demonstrated.

5.2. Bandgap analysis for a 3D lattice-based elastic metamaterial (EMM)

The bandgap characteristics of a 3D unit cell of EMM [66,67], shown in Fig. 10(a–c), are fully investigated by considering the systematic geometric and material uncertainty simultaneously.

The unit cell is characterized by four geometric parameters, including the unit cell size of 30 mm, centred sphere radius r_a, cornered sphere radius r_b, and connected cylinder radius r_c. As for the material, the unit cell is made of Nylon with the density ρ, Young's modulus E, and Poisson's ratio ν. Within the framework of FEM, the deterministic EMM model is discretised into 5666 tetrahedra elements, with a degree of freedom of 29,280. More specifically, the adopted FEM mesh is presented in Fig. 10(d). Bloch-Floquet

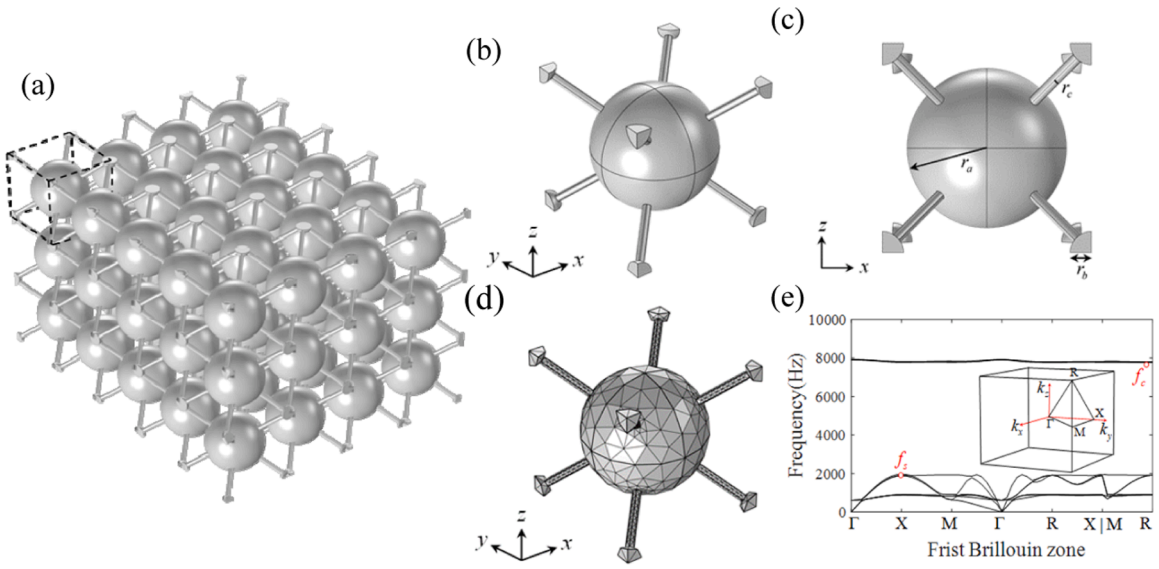


Fig. 10. Numerical model of the 3D EMM: (a) 3D view; (b) 3D view, (c) 2D view, and (d) adopted mesh of EMM unit cell; (e) a realization of band structure along the k-path in the First Brillouin Zone of the EMM.

boundary conditions [60] are applied on the boundary surfaces of the unit cells. One realization of the bandgap solution of the unit cell along the k-path of the First Brillouin zone (Γ-X-M-Γ-R-X|M-R) is depicted in Fig. 10(e).

The starting frequency f_s and cut-off frequency f_c for the first bandgap are measured. Without loss of generality, the geometric and material uncertainty are simulated in uniform distributions, with bounds of [11.4, 12.6] mm for r_a, [2.85, 3.15] mm for r_b, [0.76, 0.84] mm for r_c, [1092.5, 1207.5] kg/m^3 for ρ, [1.9, 2.1] GPa for E, and [0.38, 0.42] for ν, respectively. All inputs are generated through the LHS method.

According to the convergence study, the training size for CX-SVR model construction is set as 100. Based on the established surrogate models, PDFs, and CDFs of the concerned bandgap characteristics (i.e., f_s and f_c) are estimated and depicted in Fig. 11.

It can be demonstrated that the proposed CX-SVR technique can provide estimations on PDFs and CDFs of both f_s and f_c. The estimated PDFs and CDFs on the established CX-SVR models are overlapped with the brute MCS results. The estimated R^2-values close to 1 and relatively low *RMSE*-values can quantitively confirm the high accuracy of the established surrogate models. In addition, by further referring to the subplots in Fig. 11, the scatter plots and the estimated *RE*s of the CDFs of f_s and f_c can intuitively reflect the high accuracy of the established surrogate model, and the *RE*-values fluctuate approximately ±0.2%.

In comparison, the traditional Support Vector Regression (SVR), Neural Network (NN), and Gaussian Process Regression (GPR), are also implemented on the same 'error-free' training samples to construct the surrogate models. Their performance is summarized in Table 4.

From Table 4, it is noticed that when there is no outlier or noise embedded in the training output datasets, the proposed CX-SVR method performs at the same level as the GPR method, and slightly better than traditional SVR and NN in accuracy. Then, these machine learning methods as well as the original Extended Support Vector (X-SVR) technique are implemented for the initial bandgap frequency f_s involving some outliers. For the concerned structural response f_s, the investigation is divided into several cases by considering 5 or 10 sets of missing output datasets and different assumed outlier values (mean, maximum, or minimum value of other comprehensive training output datasets). In addition, the training inputs and testing datasets are free of outliers. The computational results are presented in Table 5.

From Table 5, it can be concluded that the proposed CX-SVR technique possesses outstanding performance in removing the effects of the outliers in training output datasets, despite the number of outliers set as 5 or 10, and the outlier values set as a mean, maximum, or minimum value of other comprehensive training output datasets. Such a robust performance of the proposed technique overpasses all other adopted machine learning techniques. More specifically, the improvement rates by using the proposed CX-SVR technique, when 5 and 10 outliers are embedded within the training output datasets, can be approximately 4% and 6% for the traditional SVR technique, 6 and 7% for NN, and 3 and 6% for GPR, respectively (without accounting the cases when the adopted method presents an extraordinarily poor performance, e.g., the estimations through the GPR when 10 outliers are considered as the minimum values). Moreover, the high robustness of the proposed CX-SVR technique is also presented on the stable performance with high accuracy in handling different valued outliers. In addition, this investigation also inspires that for surrogate model construction in engineering applications when some training datasets are missing, assigning the blanks by mean values of other comprehensive training output datasets may not yield the best training database for machine learning methods.

After clarifying the outstanding performance of the proposed CX-SVR technique in surrogate model construction under training output datasets with outliers, the proposed approach is further implemented to investigate the cut-off bandgap frequency f_c involving the Gaussian noise, and to explore the constancy of the method when noise exists. In the developed CX-SVR technique, the z-score is adopted with τ set to 0.25 and 'smallval' set to $1e^{-4}$, respectively. We replaced original output datasets $\mathbf{Y}$ with $\boldsymbol{Y}+\theta\tilde{\mathbf{Y}}$, where $\theta=n_f||\mathbf{Y}||/||\tilde{\mathbf{Y}}||$ and n_f denotes a given noise factor. The value of n_f is set in {0.2, 0.3, 0.4, 0.5}. $\tilde{\mathbf{Y}}$ denotes the noise vector whose elements are standard Gaussian variables. The performance of the surrogate models through various machine-learning techniques is summarized in Table 6.

From Table 6, the adopted machine learning techniques can generate surrogate models by learning from the training samples with noisy output datasets. In comparison to other popular machine learning techniques, the proposed CX-SVR presents a superior performance in suppressing the effect of the noise within the output training datasets, by possessing the highest R^2-values and lowest *RMSE*-values, no matter how the severity of noise (n_f-value changing from 0.2 to 0.5) is. Moreover, based on the performance of the proposed CX-SVR technique, the improvement for each machine learning technique on both estimation metrics can be indicated by the *RoI*-values. Furthermore, the increasing *RoI* indexes based on the performance of each adopted machine learning technique and the proposed CX-SVR technique can be observed generally during the increase of the proportion of the noise. Thus, it can be concluded that when the noise factor n_f increases from 0.2 to 0.5, this superior performance of the proposed CX-SVR technique appears more obvious.

In addition to the effectiveness of the proposed approach in handling generalized uncertainty quantification for 3D EMM, the computational costs are summarized. All computation is implemented in the workstation equipped with Intel(R) Xeon(R) CPU E5-2667 v4 @ 3.20GHz. The brute MCS method costs 4.52 days in total to complete 1e3 iterations calculation on the FEA model. In comparison, our proposed approach only utilizes approximately 9–200 s (varying according to various kernel functions adopted) to construct the surrogate model from 100 training datasets, which costs about 10.8 h. Then, the subsequent predictions on the established surrogate model only cost several milliseconds, which can be negligible. Convincingly, such high computational efficiency, and effectiveness, as well as the robust capability in tackling data imperfection, of the proposed approach would significantly benefit the uncertainty quantification for real-world industrial applications.

6. Conclusion

Stimulated by practices, system uncertainty involving both material and geometric randomness, and data imperfections have

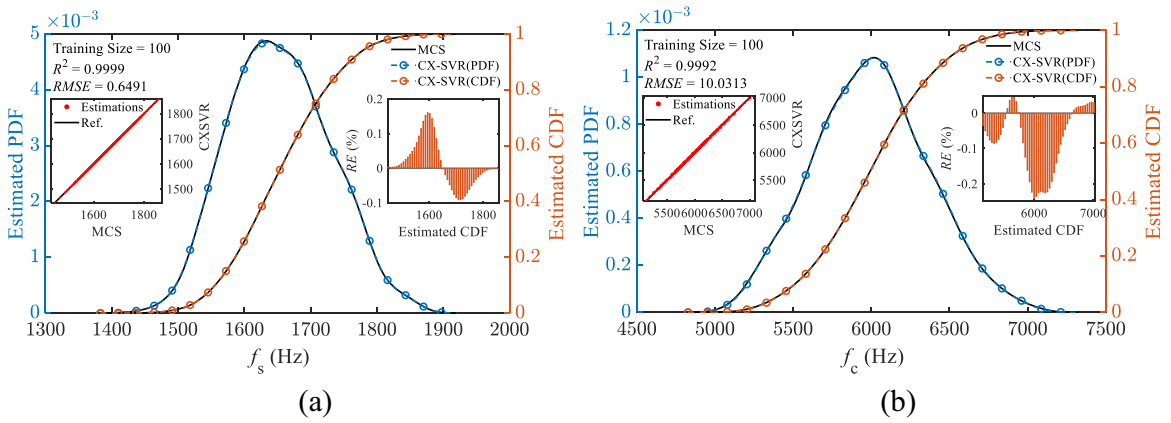


Fig. 11. Estimated PDFs and CDFs of (a) f_s and (b) f_c by CX-SVR.

Table 4
The performance of the established surrogate models under training datasets without outliers or noise.

Analysis	Method	f_s		f_c	
		R^2	*RMSE*	R^2	*RMSE*
No Outlier or Noise	CX-SVR	0.999924	0.649127	0.999227	10.031342
	SVR	0.998166	3.224060	0.995326	24.205802
	NN	0.999783	1.089254	0.998103	15.614517
	GPR	0.999928	0.629862	0.999188	10.276857

Table 5
The performance of surrogate models when the training outputs f_s involve outliers.

Analysis	Method	R^2	*RoI* (%)	*RMSE*	*RoI* (%)
5 Outliers	CX-SVR	0.987249	–	8.511121	–
- Mean	X-SVR	0.983740	-0.3554	9.277382	9.0031
	SVR	0.929477	-5.8518	17.051192	100.3401
	NN	0.830814	-15.8455	33.419556	292.6575
	GPR	0.965557	-2.1972	12.454967	46.3376
5 Outliers	CX-SVR	0.995967	–	4.734602	–
- Max	X-SVR	0.955275	-4.0857	14.881427	214.3121
	SVR	0.961213	-3.4895	13.326893	181.4786
	NN	0.918794	-7.7485	19.737156	316.8704
	GPR	0.948249	-4.7911	15.797537	233.6614
5 Outliers	CX-SVR	0.993316	–	6.192030	–
- Min	X-SVR	0.912276	-8.1585	19.873506	220.9530
	SVR	0.964357	-2.9154	12.865055	107.7680
	NN	0.938697	-5.4987	18.195887	193.8598
	GPR	0.958408	-3.5143	14.374385	132.1433
10 Outliers	CX-SVR	0.988769	–	8.183177	–
- Mean	X-SVR	0.965762	-2.3268	12.425263	51.8391
	SVR	0.907603	-8.2088	18.884587	130.7733
	NN	0.914376	-7.5238	20.081848	145.4041
	GPR	0.931606	-5.7812	16.192177	97.8715
10 Outliers	CX-SVR	0.997691	–	3.617247	–
- Max	X-SVR	0.953829	-4.3964	16.346333	351.8998
	SVR	0.959928	-3.7850	13.653588	277.4580
	NN	0.923987	-7.3875	19.954115	451.6382
	GPR	0.937657	-6.0173	17.982538	397.1333
10 Outliers	CX-SVR	0.985808	–	9.175325	–
- Min	X-SVR	0.796336	-19.2200	27.511351	199.8406
	SVR	0.922533	-6.4186	17.934636	95.4659
	NN	0.708579	-28.1220	34.657543	277.7255
	GPR	0.548222	-44.3886	40.644364	342.9747

Table 6
The performance of surrogate models when the training outputs f_c involve noise.

n_f	Method	R^2	*RoI* (%)	*RMSE*	*RoI* (%)
0.2	CX-SVR	0.989837	–	34.252958	–
	X-SVR	0.989446	-0.0395	34.804748	1.6109
	SVR	0.945287	-4.5007	77.405613	125.9823
	NN	0.940285	-5.0061	86.794070	153.3915
	GPR	0.985721	-0.4158	41.975429	22.5454
0.3	CX-SVR	0.983322	–	43.491790	–
	X-SVR	0.977695	-0.5722	48.862857	12.3496
	SVR	0.932574	-5.1609	84.867701	95.1350
	NN	0.952627	-3.1216	78.052602	79.4651
	GPR	0.979694	-0.3690	49.347381	13.4637
0.4	CX-SVR	0.975487	–	51.848233	–
	X-SVR	0.960136	-1.5737	63.383766	22.2487
	SVR	0.904324	-7.2951	99.903500	92.6845
	NN	0.871312	-10.6793	135.772704	161.8656
	GPR	0.965352	-1.0390	63.780265	23.0134
0.5	CX-SVR	0.964426	–	62.679968	–
	X-SVR	0.937132	-2.8301	77.603043	23.8084
	SVR	0.869843	-9.8072	116.206860	85.3971
	NN	0.942320	-2.2921	86.027290	37.2485
	GPR	0.946634	-1.8448	78.460187	25.1759

substantial impacts on the estimation of structural performance. Introducing the material, geometric randomness, and data imperfections (e.g., missing data, outliers, noise) simultaneously to the engineering structure would lead to an extremely complicated and error-prone system. To provide a feasible solution, this research proposes a machine learning-aided uncertainty quantification strategy. A novel kernelized regression technique, namely the Capped Extended Support Vector Regression (CX-SVR) technique, is developed for surrogate model construction, especially when the training datasets involve data imperfections. The proposed CX-SVR technique can be solved as a quadratic programming (QP) problem with a globally optimal solution available. The embedded capped strategy enables the proposed technique to remove outliers and suppress noise effectively. Cross-validation and Bayesian hyperparameter tunning strategies are adopted to avoid over-fitting and fulfil the auto-tunning, respectively. Based on the established surrogate model, subsequent analyses, such as sampling-based methods, sensitivity analysis, and optimization programming, can be implemented with ease. Instead of running on the implicit physical model, e.g., finite element analysis (FEA), for engineering structures, through the proposed approach, the computational costs, therefore, can be greatly reduced, and the potential errors hidden in the calculation process of the FEA model can also be circumvented. Furthermore, the high robustness of the proposed approach can be summarized in four main aspects: unrestrictive selection of the system inputs and their statistical information (e.g., statistical moments, distribution types), 'perfect' or 'imperfect' system outputs, enough statistical information (including statistical moments, probability density function, i.e., PDF, and cumulative distribution function, i.e., CDF) of the system outputs, and physical problems from various engineering fields. Convincingly, the proposed framework in conjunction with the newly developed regression technique would greatly benefit the engineering applications in heterogeneous disciplines.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgment

The work presented in this paper has been supported by Australian Research Council projects IH210100048, IH200100010, DP210101353, and DP240102559.

Appendix A

Imperfections in the datasets corresponding to differences between the perfect datasets D_I and real datasets D_R can be generally divided into five kinds. To describe these different data imperfections, suppose x is an ideal data record in D_I and define R_x to be its counterpart, if any, in D_R. Also, suppose $\iota(\cdot, \cdot)$ is a given scalar-valued distance function defined on $D_I \times D_R$. Several different types of data imperfections considered in this research can be distinguished as follows,

(1) Noise or observation error: $x \in D_{\mathrm{I}},\ R_x \in D_{\mathrm{R}},\ \imath(x,\ R_x) \lessapprox \theta$;
(2) Gross errors: $x \in D_I, R_x \in D_R, i(x, R_x) \gg \theta$;
(3) Simple missing data: $x \in D_{\mathrm{I}},\ R_x \notin D_{\mathrm{R}}$;
(4) Coded missing data: $x \in D_{\mathrm{I}},\ R_x = m^* \in D_{\mathrm{R}},\ m^* = \text{special}$;
(5) Disguised missing data: $x \in D_{\mathrm{I}},\ R_x = y \in D_{\mathrm{R}},\ y = \text{arbitrary}$.

where θ represents a 'small' value used to distinguish between noise (Case 1) and gross errors (Case 2). This distinction is very essential in practice, since noise is almost always present, and all standard data analysis procedures, therefore, are designed to possess a certain tolerance to noise.

For missing data, mainly there are at least three different ways: the desired record x can simply be missing from the dataset (Case 3); it can be coded as a special 'missing' data value, e.g., 'NA', 'NaN', or '?' (Case 4); or it can be disguised as a valid data value with no indication that the correct value of x is either known or undefinable (Case 5). The data anomalies in Cases 3 and 4 can be easily detected at the stage of data pre-processing, and subsequently, they can be transformed into Case 5 by substituting a specific value into these blanks. Though the missing datasets have been filled, these data points appear to be inconsistent with the nominal behaviour exhibited by most of the other data points in a specified collection. Therefore, these missing datasets after filling up can approximately be considered as kind of outliers in the application.

Overall, the analysis of data imperfection is systematic, and considering other nominal behaviour models would lead to other generally more sophisticated and subtle data imperfection classes, which are not in the scope of this work.

Appendix B

The proposed CX-SVR technique supports various Mercer's kernel functions, and some commonly used kernel functions are summarized in Table 7, including the linear kernel function, polynomial kernel function, Gaussian radial basis kernel function (RBF), Laplace RBF, Sigmoid kernel function, Eponential kernel function, Matérn 3/2 kernel function [61], Matérn 5/2 kernel function [62], Generalized Chebyshev polynomial kernel function in first and second kinds (GCKT and GCKU, respectively) [63], Generalized Gegenbauer polynomial kernel function (GGK) [7,43], and Generalized Jacobi polynomial kernel function (GJK) [64,65].

Table 7
Commonly used kernel functions in the proposed CX-SVR technique.

Kernel	Expression	Hyperparameter
Linear	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$	
Polynomial	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + 1)^d$	$d \in \mathbb{N}$
RBF	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \exp(-\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2)$	$\gamma > 0$
Laplace RBF	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \exp\left(-\frac{\parallel \mathbf{x}_i - \mathbf{x}_j \parallel}{\ell}\right)$	ℓ
Sigmoid	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \tanh(\alpha \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + \beta)$	$\alpha,\ \beta$
Exponential	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{\parallel \mathbf{x}_i - \mathbf{x}_j \parallel}{\ell}\right)$	$\sigma,\ \ell$
Matérn 3/2	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \sigma^2\left(1 + \frac{\sqrt{3} \parallel \mathbf{x}_i - \mathbf{x}_j \parallel}{\ell}\right)\exp\left(-\frac{\sqrt{3} \parallel \mathbf{x}_i - \mathbf{x}_j \parallel}{\ell}\right)$	$\sigma,\ \ell$
Matérn 5/2	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \sigma^2\left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}r}{\ell}\right)$ where $r = \parallel \mathbf{x}_i - \mathbf{x}_j \parallel$	$\sigma,\ \ell$
GCKT	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \frac{\sum_{n=0}^{d} T_n(\mathbf{x}_i)^{\mathrm{T}} T_n(\mathbf{x}_j)}{\exp(\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2)}$	$d \in \mathbb{N}$ $\gamma > 0$
GCKU	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \frac{\sum_{n=0}^{d} U_n(\mathbf{x}_i)^{\mathrm{T}} U_n(\mathbf{x}_j)}{\exp(\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2)}$	$d \in \mathbb{N}$ $\gamma > 0$
GGK	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \frac{\sum_{n=0}^{d} G_n^{\alpha}(\mathbf{x}_i)^{\mathrm{T}} G_n^{\alpha}(\mathbf{x}_j)}{\exp(\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2)}$ where $\begin{cases} G_0^{\alpha}(\mathbf{x}) = 1 \\ G_1^{\alpha}(\mathbf{x}) = 2\alpha\mathbf{x} \\ G_n^{\alpha}(\mathbf{x}) = \frac{1}{n}[2(n+\alpha-1)\mathbf{x}^{\mathrm{T}} G_{n-1}^{\alpha}(\mathbf{x}) - (n+2\alpha-2) G_{n-2}^{(\alpha,\beta)}(\mathbf{x})] \end{cases}$	$d \in \mathbb{N}$ $\gamma > 0$ α
GJK	$\kappa(\mathbf{x}_i,\ \mathbf{x}_j) = \frac{\sum_{n=0}^{d} J_n^{(\alpha,\beta)}(\mathbf{x}_i)^{\mathrm{T}} J_n^{(\alpha,\beta)}(\mathbf{x}_j)}{\exp(\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2)}$ where	$d \in \mathbb{N}$ $\gamma > 0$ $\alpha,\ \beta$

(*continued on next page*)

Table 7 (*continued*)

Kernel	Expression	Hyperparameter
	$\begin{cases} P_0^{(\alpha,\beta)}(\mathbf{x}) = 1 \\ P_1^{(\alpha,\beta)}(\mathbf{x}) = \frac{1}{2}(\alpha+\beta+2)\mathbf{x} + \frac{1}{2}(\alpha-\beta) \\ P_n^{(\alpha,\beta)}(\mathbf{x}) = P_{n-1}^{(\alpha,\beta)}(\mathbf{x})^{\mathrm{T}} A_n^{(\alpha,\beta)}(\mathbf{x}) - B_n^{(\alpha,\beta)} P_{n-2}^{(\alpha,\beta)}(\mathbf{x}) \end{cases}$ and $A_n^{(\alpha,\beta)}(\mathbf{x}) = \frac{(2n+\alpha+\beta+1)[(2n+\alpha+\beta)(2n+\alpha+\beta-2)\mathbf{x} + (\alpha^2-\beta^2)]}{2n(n+\alpha+\beta)(2n+\alpha+\beta-2)}$ $B_n^{(\alpha,\beta)} = \frac{(n+\alpha-1)(n+\beta-1)(2n+\alpha+\beta)}{n(n+\alpha+\beta)(2n+\alpha+\beta-2)}$	

References

[1] S. Olugbade, S. Ojo, A.L. Imoize, J. Isabona, M.O. Alaba, A review of artificial intelligence and machine learning for incident detectors in road transport systems, Math. Comput. Appl. 27 (5) (2022) 77.
[2] M.L. Tseng, T.P.T. Tran, H.M. Ha, T.D. Bui, M.K. Lim, Sustainable industrial and operation engineering trends and challenges toward industry 4.0: a data driven analysis, J. Ind. Prod. Eng. 38 (8) (2021) 581–598.
[3] P.S. Aithal, Information communication & computation technology (ICCT) as a strategic tool for industry sectors, Int. J. Appl. Eng. Manag. Lett. 3 (2) (2019) 65–80 (IJAEML).
[4] Y. Huang, C. Shao, B. Wu, J.L. Beck, H. Li, State-of-the-art review on Bayesian inference in structural system identification and damage assessment, Adv. Struct. Eng. 22 (6) (2019) 1329–1351.
[5] Q. Liu, Y. Dai, X. Wu, X. Han, H. Ouyang, Z. Li, A non-probabilistic uncertainty analysis method based on ellipsoid possibility model and its applications in multi-field coupling systems, Comput. Methods Appl. Mech. Eng. 385 (2021) 114051.
[6] D. Wu, Q. Wang, A. Liu, Y. Yu, Z. Zhang, W. Gao, Robust free vibration analysis of functionally graded structures with interval uncertainties, Compos. Part B Eng. 159 (2019) 132–145.
[7] Q. Wang, Q. Li, D. Wu, Y. Yu, F. Tin-Loi, J. Ma, W. Gao, Machine learning aided static structural reliability analysis for functionally graded frame structures, Appl. Math. Model 78 (2020) 792–815.
[8] R.K. Pearson, Mining imperfect data: with examples in R and python, Soc. Ind. Appl. Math. (2020).
[9] A.B. Sharma, L. Golubchik, R. Govindan, Sensor faults: detection methods and prevalence in real-world datasets, ACM Trans. Sens. Netw. 6 (3) (2010) 1–39 (TOSN).
[10] J.W. Graham, Missing data analysis: making it work in the real world, Annu. Rev. Psychol. 60 (2009) 549–576.
[11] H.R. Bae, R.V. Grandhi, R.A. Canfield, An approximation approach for uncertainty quantification using evidence theory, Reliab. Eng. Syst. Saf. 86 (3) (2004) 215–225.
[12] R.C. Smith, Uncertainty Quantification: Theory, Implementation, and Applications, 12, Siam, 2013.
[13] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, S. Nahavandi, A review of uncertainty quantification in deep learning: techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297.
[14] A. Haldar, S. Mahadevan, Reliability Assessment Using Stochastic Finite Element Analysis, John Wiley & Sons, 2000.
[15] R.G. Ghanem, P.D. Spanos, Stochastic Finite Elements: A Spectral Approach, Courier Corporation, 2003.
[16] B. Van den Nieuwenhof, J.P. Coyette, Modal approaches for the stochastic finite element analysis of structures with material and geometric uncertainties, Comput. Methods Appl. Mech. Eng. 192 (33–34) (2003) 3705–3729.
[17] G. Stefanou, The stochastic finite element method: past, present and future, Comput. Methods Appl. Mech. Eng. 198 (9–12) (2009) 1031–1051.
[18] G. Stefanou, Response variability of cylindrical shells with stochastic non-Gaussian material and geometric properties, Eng. Struct. 33 (9) (2011) 2621–2627.
[19] Z. Zheng, H. Dai, M. Beer, Efficient structural reliability analysis via a weak-intrusive stochastic finite element method, Probab. Eng. Mech. (2023) 103414.
[20] K. Sepahvand, Stochastic finite element method for random harmonic analysis of composite plates with uncertain modal damping parameters, J. Sound Vib. 400 (2017) 1–12.
[21] X. Chen, J. Liu, N. Xie, H. Sun, Probabilistic analysis of embankment slope stability in frozen ground regions based on random finite element method, Sci. Cold Arid Reg. 7 (4) (2015) 0354–0364.
[22] B. Sudret, Polynomial chaos expansions and stochastic finite element methods, Risk Reliab. Geotech. Eng. (2014) 265–300.
[23] C.Z. Mooney, Monte Carlo Simulation (No. 116), Sage, 1997.
[24] E. Zio, Monte Carlo Simulation: The Method, Springer London, 2013, pp. 19–58.
[25] W.K. Liu, T. Belytschko, A. Mani, Random field finite elements, Int. J. Numer. Methods Eng. 23 (10) (1986) 1831–1845.
[26] A.H. Nayfeh, Perturbation Methods, John Wiley & Sons, 2008.
[27] R.G. Ghanem, P.D. Spanos, Spectral stochastic finite-element formulation for reliability analysis, J. Eng. Mech. 117 (10) (1991) 2351–2372.
[28] R.G. Ghanem, P.D. Spanos, Stochastic Finite Elements: A Spectral Approachs, Courier Corporation, 2004.
[29] J.D. Arregui-Mena, L. Margetts, P.M. Mummery, Practical application of the stochastic finite element method, Arch. Comput. Methods Eng. 23 (2016) 171–190.
[30] B. Rong, X. Rui, L. Tao, Perturbation finite element transfer matrix method for random eigenvalue problems of uncertain structures, J. Appl. Mech. 79 (2) (2012).
[31] S. Rahman, B. Rao, A perturbation method for stochastic meshless analysis in elastostatics, Int. J. Numer. Methods Eng. 50 (8) (2001) 1969–1991.
[32] Çavdar, Ö., Bayraktar, A., Çavdar, A., & Adanur, S., 2008, Perturbation based stochastic finite element analysis of the structural systems with composite sections under earthquake forces.
[33] M. Kaminski, The Stochastic Perturbation Method for Computational Mechanics, John Wiley & Sons, 2013.
[34] D.M. Do, W. Gao, C. Song, Stochastic finite element analysis of structures in the presence of multiple imprecise random field parameters, Comput. Methods Appl. Mech. Eng. 300 (2016) 657–688.
[35] K.G. Jos, K.J. Vinoy, An efficient SSFEM-POD scheme for wideband stochastic analysis of permittivity variations, IEEE Trans. Antennas Propag. (2022).
[36] E. Pitz, K. Pochiraju, AI/ML for quantification and calibration of property uncertainty in composites. Machine Learning Applied to Composite Materials, Springer Nature Singapore, Singapore, 2022, pp. 45–76.
[37] K. Li, D. Wu, W. Gao, C. Song, Spectral stochastic isogeometric analysis of free vibration, Comput. Methods Appl. Mech. Eng. 350 (2019) 1–27.
[38] K. Li, D. Wu, W. Gao, Spectral stochastic isogeometric analysis for static response of FGM plate with material uncertainty, Thin Walled Struct. 132 (2018) 504–521.

[39] K.R.G. Hewawasam, K. Premaratne, M.L. Shyu, Rule mining and classification in a situation assessment application: a belief-theoretic approach for handling data imperfections, IEEE Trans. Syst. Man Cybern. B 37 (6) (2007) 1446–1459 (Cybernetics).
[40] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J.N. Chiang, Z. Wu, X. Ding, Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation, Med. Image Anal. 63 (2020) 101693.
[41] L. Sun, J.X. Wang, Physics-constrained bayesian neural network for fluid flow reconstruction with sparse and noisy data, Theor. Appl. Mech. Lett. 10 (3) (2020) 161–169.
[42] J. Feng, L. Liu, D. Wu, G. Li, M. Beer, W. Gao, Dynamic reliability analysis using the extended support vector regression (X-SVR), Mech. Syst. Signal Process. 126 (2019) 368–391.
[43] Q. Wang, D. Wu, F. Tin-Loi, W. Gao, Machine learning aided stochastic structural free vibration analysis for functionally graded bar-type structures, Thin Walled Struct. 144 (2019) 106315.
[44] C. Wang, Q. Ye, P. Luo, N. Ye, L. Fu, Robust capped L1-norm twin support vector machine, Neural Netw. 114 (2019) 47–59.
[45] Y. Li, H. Sun, W. Yan, Q. Cui, R-CTSVM+: robust capped L1-norm twin support vector machine with privileged information, Inf. Sci. 574 (2021) 12–32 (Ny).
[46] S. Chinchalkar, D.L. Taylor, Geometric uncertainties in finite element analysis, Comput. Syst. Eng. 5 (2) (1994) 159–170.
[47] Rozvany, G.I., & Lewiński, T., eds., 2014, Topology optimization in structural and continuum mechanics.
[48] Multiphysics, C.O.M.S.O.L, 2013, Comsol multiphysics reference manual. COMSOL Grenoble, France, 1084, 834.
[49] R.K. Pearson, Mining Imperfect Data: Dealing with Contamination and Incomplete Records, Society for Industrial and Applied Mathematics, 2005.
[50] G. Gordon, R. Tibshirani, Karush-kuhn-tucker conditions, Optimization 10 (725/36) (2012) 725.
[51] P.J. Rousseeuw, M. Hubert, Robust statistics for outlier detection, Wiley interdiscip. Rev. Data min. Knowl. Discov. 1 (1) (2011) 73–79.
[52] H.P. Vinutha, B. Poornima, B.M. Sagar, Detection of outliers using interquartile range technique from intrusion dataset, in: Proccedings of the Information and Decision Sciences: Proceedings of the 6th International Conference on FICTA, Springer Singapore, 2018, pp. 511–518.
[53] F. Pukelsheim, The three sigma rule, Am. Stat. 48 (2) (1994) 88–91.
[54] M.W. Browne, Cross-validation methods, J. Math. Psychol. 44 (1) (2000) 108–132.
[55] R.R. Picard, R.D. Cook, Cross-validation of regression models, J. Am. Stat. Assoc. 79 (387) (1984) 575–583.
[56] T.T. Joy, S. Rana, S. Gupta, S. Venkatesh, December, hyperparameter tuning for big data using Bayesian optimisation, in: Proceedings of the 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2574–2579.
[57] J. Wu, X.Y. Chen, H. Zhang, L.D. Xiong, H. Lei, S.H. Deng, Hyperparameter optimization for machine learning models based on Bayesian optimization, J. Electron. Sci. Technol. 17 (1) (2019) 26–40.
[58] M. Ambati, T. Gerasimov, L. De Lorenzis, A review on phase-field models of brittle fracture and a new fast hybrid formulation, Comput. Mech. 55 (2015) 383–405.
[59] C. Miehe, M. Hofacker, F. Welschinger, A phase field model for rate-independent crack propagation: robust algorithmic implementation based on operator splits, Comput. Methods Appl. Mech. Eng. 199 (45–48) (2010) 2765–2778.
[60] M. Zhang, C. Hu, C. Yin, Q.H. Qin, J. Wang, Design of elastic metamaterials with ultra-wide low-frequency stopbands via quantitative local resonance analysis, Thin Walled Struct. 165 (2021) 107969.
[61] A. Melkumyan, F. Ramos, Multi-kernel Gaussian processes, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011.
[62] N. Zhang, J. Xiong, J. Zhong, K. Leatham, Gaussian process regression method for classification for high-dimensional data with limited samples, in: Proceedings of the 8th International Conference on Information Science and Technology (ICIST), IEEE, 2018, pp. 358–363.
[63] S. Ozer, C.H. Chen, H.A. Cirpan, A set of new Chebyshev kernel functions for support vector machine pattern classification, Pattern Recognit. 44 (7) (2011) 1435–1447.
[64] Q. Wang, D. Wu, G. Li, W. Gao, A virtual model architecture for engineering structures with twin extended support vector regression (TX-SVR) method, Comput. Methods Appl. Mech. Eng. 386 (2021) 114121.
[65] Q. Wang, Y. Feng, D. Wu, C. Yang, Y. Yu, G. Li, W. Gao, Polyphase uncertainty analysis through virtual modelling technique, Mech. Syst. Signal Process 162 (2022) 108013. SHAPE * MERGEFORMAT.
[66] M. Zhang, Q. Wang, Z. Luo, W. Gao, Stochastic bandgap optimization for multiscale elastic metamaterials with manufacturing imperfections, Int. J. Mech. Sci 268 (2024) 109035.
[67] M. Zhang, Q. Wang, Z. Luo, W. Gao, Virtual model-aided reliability analysis considering material and geometrical uncertainties for elastic metamaterials, Mech. Syst. Signal Process. 211 (2024) 111199.