IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 November 2024, accepted 10 January 2025, date of publication 16 January 2025, date of current version 27 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3530417

RESEARCH ARTICLE

A Novel Multi-Fidelity Support Vector Classification Method for Boundary Prediction in Engineering Applications

JINLIANG LUO, LINGZHI LIU, YOUWEI HE, AND KUAN TAN
School of Mechanical Engineering, University of South China, Hengyang 421001, China

Corresponding author: Youwei He (youwei.he@usc.edu.cn)

This work was supported in part by the Natural Science Foundation of Hunan Province under Grant 2023JJ40545, in part by the Research Foundation of Education Bureau of Hunan Province under Grant 23A0344, and in part by the Research Program of University of South China under Grant 220XQD064.

ABSTRACT The accurate prediction of failure boundaries in engineering applications is essential for ensuring safety and reliability. Traditional methods often rely heavily on high-fidelity physical experiments or numerical simulations, which are prohibitively expensive and time-consuming. In response to this challenge, our research proposes an innovative multi-fidelity support vector classification approach that leverages an abundant supply of low-fidelity data alongside a limited amount of high-fidelity data. This combination significantly reduces modeling costs while maintaining or even enhancing predictive accuracy. The key points of the proposed method include the design of a reasonable kernel function to effectively describe the relationship between the input and output of multiple fidelities, and the determination of the optimal hyperparameters. In addition, in practical engineering problems, real data often exhibit data imbalance, leading to poor performance of the trained models. Our novel method addresses this limitation by integrating a strategy for managing the data imbalance. By effectively treating data imbalance, our approach significantly improves the classification and boundary prediction capabilities of the model. To validate our method, we applied it to three distinct engineering problems: predicting the failure boundary of a zero Poisson ratio structure, analyzing surge and choke boundaries in an axial flow compressor rotor, and tackling a 31-dimensional simulation failure boundary prediction problem within the computational fluid dynamics context of the same rotor. The results demonstrate that our multi-fidelity support vector classification method not only effectively predicts boundaries in these practical scenarios but also outperforms alternative methods, showing its potential as a powerful tool for engineers.

INDEX TERMS Boundary prediction, classification, imbalance data, compressor operating boundary.

I. INTRODUCTION

In engineering design, predicting the failure boundary is an inevitable process aimed at exploring the boundary between failure and safety limit states. To predict the failure boundary, engineering design typically requires computation-intensive and real-life experiments, which often consume a significant amount of time and financial costs, such as exploring the surge and choke boundary of compressors in the aircraft engine design process [1], [2], [3], [4]. However, surrogate models based on limited simulation or experimental data can effectively overcome this drawback and have been widely used in multiple engineering applications [5], [6], [7]. Several common surrogate models are suitable for failure boundary prediction, such as Gaussian process (GP) model [8], response surface methodology (RSM) model [9], support vector machine (SVM) model [10], and artificial neural network (ANN) etc. [11].

Among the many surrogate models, SVM has a strong capability to identify subtle patterns in complex datasets and possesses good robustness and model generalization ability, making it suitable for addressing classification problems

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

with small to medium sample sizes, nonlinearity, and high dimensions [12]. Therefore, many scholars have conducted research on the SVM models. Cervantes [13] discussed various applications of SVM models in multiple domains and examined the prospects and limitations of SVM models. Liu et al. [14] proposed an iterative l_2-SVM model that achieved dual-regularized SVM on high-dimensional datasets. This method significantly reduces the computational complexity of the model. Abe [15] proposed a fuzzy SVM model for multilabel classification, defining a region with a relevant membership function for each multilabel classification. Yan [16] proposed a dual-bounded SVM model for binary classification to solve a pair of quadratic programming problems. This method reduces the computational cost of SVM models by utilizing the L1-norm distance measurement in classification. Pradhan and Sameen [17] proposed an SVM model with a rectified linear unit kernel to evaluate deep-learning processes. Wu [18] developed an SVM model with a deterministic and scalable histogram intersection kernel to increase the training speed of SVM models.

However, the above-mentioned studies on SVM models all focused on single-fidelity models trained only with single-fidelity information sources. When training surrogate models with low-fidelity (LF) information sources, the surrogate models are inaccurate although the modeling cost is lower. In contrast, training surrogate models with high-fidelity (HF) information sources can improve the accuracy of surrogate models but at a higher cost. Therefore, it is necessary to study methods that can fully utilize the advantages of HF and LF data to achieve an optimal structure at the same or lower cost. To address this issue, a multi-fidelity surrogate (MFS) model based on HF and LF sample points was developed [19]. The MFS model utilizes more LF sample points than HF sample points to capture the overall trend of the engineering system and correct it with expensive HF data, thereby improving the accuracy of the surrogate model while maintaining the same or lower cost of sample data acquisition. Owing to its advantages, the MFS model has attracted extensive research interest. Forrester [20] extended the Kriging model to a multi-fidelity kriging model by constructing a correlation matrix between the HF and LF data. Liu and Wang [21] proposed an MFS model based on an artificial NN and incorporated physics constraints to reduce the required training data and enhance the prediction accuracy. Song et al. [22] proposed an MFS model based on polynomial response surface regression to improve its prediction accuracy. Shi et al. [23] proposed an MFS model based on support vector regression, describing the correlation between LF and HF models in the mapped high-dimensional space. Song et al. [24] introduced a radial basis function-based MFS model to improve the accuracy and robustness of the model with reduced sensitivity to the correlation between the LF and HF models. Aydin et al. [25] proposed an MFS model based on an ANN to reduce the computational cost of the model.

Although many models are based on the MFS framework, regression models are generally the most commonly used. However, when it comes to predicting failure boundaries, it becomes a binary classification problem, and regression models are often unsuitable for such tasks. Currently, the multi-fidelity Gaussian process classification (MFGPC) model is the only multi-fidelity classification model that offers interpretability [26]. However, the MFGPC model mainly uses an approximate inference method called the Markov chain Monte Carlo method to predict classification results, which is often time-consuming. No research has been conducted on the multi-fidelity support vector classification (MFSVC) model. To fill this gap, this study proposed an MFSVC model. The model maps the HF and LF samples to a high-dimensional feature space through a correlation kernel function and then uses a linear model to construct the relationship function between the input and output, thereby finding the optimal decision boundary in the high-dimensional feature space.

Meanwhile, in engineering problems, the absence or difficulty of obtaining sample points may lead to imbalanced training sets, making it difficult for most surrogate models to accurately capture the inherent characteristics of the engineering system and effectively detect the corresponding minority class samples. To address imbalanced data, many scholars have proposed solutions for imbalanced data processing. Onan [27] proposed an undersampling technique based on consensus clustering to improve the accuracy of surrogate models for imbalanced data. Camacho et al. [28] applied a geometric oversampling algorithm to a regression model and verified the effectiveness of balanced data processing methods on the accuracy of convolutional neural network models. Radwan [29] combined oversampling techniques with threshold computation to balance the training set and select the optimal classification boundary. Liu et al. [30] combined oversampling and undersampling techniques with SVM models to improve their predictive performance. Tsai et al. [31] proposed an undersampling method that combines cluster analysis and instance selection to improve the classification effectiveness of classification models on imbalanced data. Among many imbalanced data processing methods, the synthetic minority oversampling technique with the edited nearest neighbors (SMOTEENN) algorithm, which combines oversampling and undersampling techniques, effectively generates minority class samples without losing useful information in engineering systems [32]. Furthermore, the SMOTEENN algorithm is now systematic and can be quickly applied to engineering systems without relying on the user experience. Therefore, this study applied the SMOTEENN algorithm to preprocess imbalanced datasets to improve the accuracy of the MFSVC model under imbalanced data conditions.

To this end, this study first developed an MFSVC model. Then, the SMOTEENN algorithm was incorporated into the MFSVC method to improve the predictive performance of

the MFSVC model with imbalanced data. Subsequently, the performance of the proposed method was examined using numerical problems and three practical engineering problems. The advantages of this method are analyzed and summarized in terms of accuracy compared with existing methods.

The remainder of this paper is organized as follows. Section II provides a detailed introduction of the proposed method. In Section III, the effectiveness of the established methods is validated using multiple numerical examples, and the influence of the number of high-precision sample points on the predictive performance of the MFSVC model is studied. Section IV presents the experimental results of the proposed method on three engineering problems and compares them with those of alternative methods. Finally, the conclusions are presented in Section V.

II. PROPOSED METHODS

This section introduces an MFSVC model based on the SVC model to address imbalanced datasets. This section introduces the principles of the SVC model to better understand the MFSVC model. Then, the establishment and strategy to cope with the imbalanced data of the MFSVC model are introduced. Finally, an implementation of the proposed method is presented.

A. SUPPORT VECTOR CLASSIFICATION

The SVC model is a classification model based on the Vapnik-Chervonenkis dimension theory and principle of structural risk minimization [33]. It aims to determine the best separating hyperplane with the maximum margin between classes by mapping the data to a high-dimensional feature space. To predict the failure boundary and classify it into failure and safe domains, the classification model is configured as a binary classification, which means that the model outputs the labels $y = \{-1, 1\}$. Let a dataset $D = \left\{\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right) | i = 1, 2, \ldots, n\right\}$ consist of s features for n sample points $\boldsymbol{x}_i$, each containing a binary classification label $y = \pm 1$. The SVC classification model is expressed as follows:

$$f(\boldsymbol{x}) = \boldsymbol{\omega}^T \cdot \phi(\boldsymbol{x}) + b \tag{1}$$

where φ denotes the feature map, ω is the weight vector, and b is a bias term. To obtain the optimal hyperplane, the following quadratic programming (QP) problem is solved [34]:

$$\Phi(\boldsymbol{\omega}, \boldsymbol{\xi}) = \min\left\{\boldsymbol{\omega}^T \cdot \boldsymbol{\omega}/2 + C\sum\nolimits_{i=1}^{n} \xi_i\right\} \tag{2}$$

subject to the constraint that all training samples are correctly classified; that is, all training samples are placed on the margin or outside the margin.

$$y_i\left(\boldsymbol{\omega}^T \cdot \phi(\boldsymbol{x}_i) + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, n \tag{3}$$

where $\xi_i, i = 1, \ldots, n$ is a nonnegative slack variable. The first part of (2) represents weight decay, which is used to regulate the size of weights and penalize overly large weights. The second part represents the classification error of all the training points. By minimizing (2), both the computational complexity of SVC and the number of training errors can be reduced. Parameter C is the regularization parameter defined as the relative weight between the two terms. The constrained QP problem defined in (2) and (3) is solved by the introducing the Lagrange multipliers $\alpha_i \geq 0, i = 1, \ldots, n$ and the Lagrange function:

$$\begin{aligned} L(\boldsymbol{\omega}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \|\boldsymbol{\omega}\|^2/2 + C\sum\nolimits_{i=1}^{n} \xi_i \\ - \sum\nolimits_{i=1}^{n} \alpha_i\left\{y_i\left[\boldsymbol{\omega}^T \cdot \phi(\boldsymbol{x}_i) + b\right] - 1 + \xi_i\right\} \end{aligned} \tag{4}$$

According to QP optimization theory, (4) can be solved by introducing the dual form of the problem:

$$\max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}}\left\{\min_{\boldsymbol{\omega}, b, \boldsymbol{\xi}} L(\boldsymbol{\omega}, b, \boldsymbol{\xi}, \boldsymbol{\alpha})\right\} \tag{5}$$

where $\boldsymbol{\alpha}$ denotes a Lagrange multiplier. The optimal solution of (5) is obtained by minimizing the Lagrange function with respect to $\boldsymbol{\omega}, b, \boldsymbol{\xi}$ and then maximizing the Lagrange function with respect to α_i. The optimal solution can be obtained by setting the first derivative of (5) to zero:

$$\begin{aligned} &\frac{\partial L}{\partial \boldsymbol{\omega}} = 0 \rightarrow \boldsymbol{\omega} = \sum_{i=1}^{n} \alpha_i y_i \phi(\boldsymbol{x}_i) \\ &\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \\ &\frac{\partial L}{\partial \boldsymbol{\xi}} = 0 \rightarrow \sum_{i=1}^{n} \alpha_i = C \\ &\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow -y_i\left[\boldsymbol{\omega}^T \cdot \phi(\boldsymbol{x}_i) + b\right] + 1 - \xi_i = 0, \\ &\qquad i = 1, \ldots, n \end{aligned} \tag{6}$$

Subsequently, by substituting (6) into (5), the problem shown in (5) can be transformed into the following dual problem:

$$\max_{\alpha}\left\{\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)\right\} \tag{7}$$

and it should be maximized under the constraints:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C \text{ for } i = 1, \ldots, n. \tag{8}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denotes the Radial Basis Function (RBF) kernel function $\boldsymbol{K}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \exp(-\sum_{k=1}^{s} \gamma \left\|\boldsymbol{x}_i^k - \boldsymbol{x}_j^k\right\|^2)$. This kernel function is commonly used in SVC, demonstrating good performance for both large and small sample sizes, and it has fewer parameters, which is why it was selected. According to the result of the last term of (7), the training vectors of the nonzero Lagrange multipliers, namely support vectors (SV), can be obtained to describe the hyperplane. To solve the QP problem mentioned above, a sequential minimal optimization (SMO) algorithm was utilized [35].

For easy understanding of the derivation for solving the α and b values of the subsequent MFSVC model, a detailed inference process is provided here. The SMO algorithm is a heuristic algorithm that iteratively optimizes the solution of a problem by selecting two α for optimization each time. Because the constraint condition of α determines the accumulation of its product with y equal to be zero, two optimizations are needed at one time to preserve the constraint condition. Let α_1 and α_2 be two initial feasible solutions. Using (6) and (7), we can obtain the following:

$$\begin{aligned}\max_{\alpha_1,\alpha_2} W(\alpha_1,\alpha_2) = \max_{\alpha_1,\alpha_2} (\ & \alpha_1+\alpha_2-\frac{1}{2}\boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_1)\alpha_1^2 \\ & -\frac{1}{2}\boldsymbol{K}(\boldsymbol{x}_2,\boldsymbol{x}_2)\alpha_2^2-\alpha_1\alpha_2 y_1 y_2 \boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_2) \\ & -\alpha_1 y_1\sum_{j=3}^{n}\alpha_j y_j \boldsymbol{K}(\boldsymbol{x}_j,\boldsymbol{x}_1) \\ & -\alpha_2 y_2\sum_{j=3}^{n}\alpha_j y_j \boldsymbol{K}(\boldsymbol{x}_j,\boldsymbol{x}_2)+c\ ) \end{aligned} \tag{9}$$

$$\text{subject to: } \begin{cases}\alpha_1 y_1+\alpha_2 y_2 = -\sum_{j=3}^{n}\alpha_j y_j = \zeta \\ 0\le\alpha_i\le C,\ i=1,2\end{cases} \tag{10}$$

where parameter c is the part that is independent of α_1 and α_2 and is treated as a constant term in this optimization. The parameter ζ is a constant. Using (1) and (10), (9) can be transformed into a unary problem, and the partial derivatives can be obtained as:

$$\begin{aligned}\frac{\partial\phi(\alpha_2)}{\partial\alpha_2} = (&-y_2y_1+1+\boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_1)\zeta y_2-\boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_1)\alpha_2 \\ &-\boldsymbol{K}(\boldsymbol{x}_2,\boldsymbol{x}_2)\alpha_2-y_2\boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_2)\zeta \\ &+2\boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_2)\alpha_2+y_2v_1-y_2v_2)\end{aligned} \tag{11}$$

where $v_i = \sum_{j=3}^{n}\alpha_j y_j \boldsymbol{K}(\boldsymbol{x}_j,\boldsymbol{x}_i)$. Let the partial derivative be equal to 0 and be simplified to obtain the update equation of parameter α:

$$\alpha_2' = \alpha_2 + \frac{y_2\{[f(\boldsymbol{x}_1)-y_1]-[f(\boldsymbol{x}_2)-y_2]\}}{\boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_1)+\boldsymbol{K}(\boldsymbol{x}_2,\boldsymbol{x}_2)-2\boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_2)} \tag{12}$$

where parameter α_i' is an unconstrained updated Lagrange multiplier. Parameter α_i is an old Lagrange multiplier. According to (11), by constraining the update equation, we obtain

$$\begin{aligned}\alpha_2'' &= \begin{cases}H & \text{if } \alpha_2' > H \\ \alpha_2' & \text{if } L\le\alpha_2'\le H \\ L & \text{if } \alpha_2' < L\end{cases} \\ \alpha_1'' &= \alpha_1+y_1y_2(\alpha_2-\alpha_2'')\end{aligned} \tag{13}$$

with

$$\begin{aligned}L &= \begin{cases}\max(0,\alpha_2-\alpha_1) & \text{if } y_1\ne y_2 \\ \max(0,\alpha_2+\alpha_1-C) & \text{if } y_1=y_2\end{cases} \\ H &= \begin{cases}\min(C,C+\alpha_2-\alpha_1) & \text{if } y_1\ne y_2 \\ \min(C,\alpha_2+\alpha_1) & \text{if } y_1=y_2\end{cases}\end{aligned} \tag{14}$$

where parameter α_i'' is a constrained updated Lagrange multiplier. The updated formula for the threshold b can be obtained as follows:

$$\begin{aligned}b_1' &= -E_1-y_1\boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_1)(\alpha_1''-\alpha_1) \\ &\quad -y_2\boldsymbol{K}(\boldsymbol{x}_2,\boldsymbol{x}_1)(\alpha_2''-\alpha_2)+b \\ b_2' &= -E_2-y_1\boldsymbol{K}(\boldsymbol{x}_1,\boldsymbol{x}_2)(\alpha_1''-\alpha_1) \\ &\quad -y_2\boldsymbol{K}(\boldsymbol{x}_2,\boldsymbol{x}_2)(\alpha_2''-\alpha_2)+b \\ b'' &= \frac{b_1'+b_2'}{2}\end{aligned} \tag{15}$$

where parameter b_i' is the updated threshold. b represents the initial set of the thresholds. b'' is a compromise between two updated thresholds. Parameter $E_i = f(\boldsymbol{x}_i)-y_i$ is the model prediction error. The above process is iterated until the growth rate of the model objective function $W(\alpha)$ is less than threshold e:

$$\frac{W(\boldsymbol{\alpha}^{t+1})-W(\boldsymbol{\alpha}^{t})}{W(\boldsymbol{\alpha}^{t})} < e \tag{16}$$

where $W(\alpha^t)$ and $W(\alpha^{t+1})$ are the objective function results for iterations t and $t+1$, respectively. After obtaining the optimal solutions α_i^{best} and b^{best}, the decision boundary $f(\boldsymbol{x})$ is determined as follows:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}\alpha_i^{best}y_i\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)+b^{best} \tag{17}$$

The kernel parameters γ and the regularization parameter C are an input parameters for training the SVM models and must be adjusted to achieve sufficient classification performance. To separate the classification results into two classes, the optimal decision function is obtained based on (10) for the SVC model:

$$g(\boldsymbol{x}) = sign(f(\boldsymbol{x})) = sign(\sum_{i=1}^{n}a_i^{best}y_i\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)+b^{best}) \tag{18}$$

where *sign* is a signed function aimed at converting the output of the function into $y=\pm1$.

B. MULTI-FIDELITY SUPPORT VECTOR CLASSIFICATION

The improved MFSVC model, based on the SVC model, is a classification model that utilizes multiple information sources to analyze and combine data. It assumes the existence of two information sources, each with a different evaluation cost and fidelity. Label L represents a simplified and cheaper information source, whereas the label H represents a more accurate and expensive information source. The LF and HF datasets are denoted as $D_L = \{(\boldsymbol{x}_i,\boldsymbol{y}_i)\,|i=1,2,\ldots,n_L\}$ and $D_H = \{(\boldsymbol{x}_i,\boldsymbol{y}_i)\,|i=1,2,\ldots,n_H\}$, respectively. It is worth noting that both datasets have s number of features. The MFSVC model aims to fuse the two datasets to provide a better prediction of the separation boundary at the HF level compared with single-fidelity models. To this end, the auto-regressive model is adopted to establish the relationship between the LF and HF functions [36]:

$$g_L(\boldsymbol{x}) = g_L(\boldsymbol{x})$$

$$g_H(\boldsymbol{x}) = \rho g_L(\boldsymbol{x}) + \delta(\boldsymbol{x}) \tag{19}$$

where the function $g_L(\boldsymbol{x})$ is the latent function of the input and output of LF dataset D_L. the function $g_H(\boldsymbol{x})$ is the latent function of the input and output of the HF dataset D_H. The parameter ρ is a scalar that needs to be inferred and represents the linear correlation between HF and LF function. And the function $\delta(\boldsymbol{x})$ is designed to capture the prediction bias of LF models at the HF sample sites. According to the basic principles of the SVC model, the basic form of the MFSVC model is defined as follows:

$$g(\boldsymbol{X}) = sign\left(\sum_{i=1}^{n_L+n_H} \alpha_i Y_i \boldsymbol{K}_{MF}(\boldsymbol{x}_L, \boldsymbol{x}_H) + b\right) \tag{20}$$

where Y_i is the response value of the support vector determined during the construction of the MFSVC. By comparing (17) and (20), it can be concluded that the key of the MFSVC model is to design a reasonable kernel function to effectively describe the relationship between multi-fidelity input data and output results. An isotropic covariance matrix is established as follows:

$$\boldsymbol{K}_{MF}(\boldsymbol{x}_L, \boldsymbol{x}_H) = \begin{bmatrix} \boldsymbol{K}_{LL} & \boldsymbol{K}_{LH} \\ \boldsymbol{K}_{HL} & \boldsymbol{K}_{HH} \end{bmatrix} \tag{21}$$

where

$$\begin{aligned}
\boldsymbol{K}_{LL} &= \sigma_L e^{-\sum_{k=1}^{s} \gamma_L \left\| x_L^{k,i} - x_L^{k,j} \right\|^2} \\
&\quad (i, j = 1, \ldots, n_L) \\
\boldsymbol{K}_{LH} &= \rho\sigma_L e^{-\sum_{k=1}^{s} \gamma_L \left\| x_L^{k,i} - x_H^{k,j} \right\|^2} \\
&\quad (i = 1, \ldots, n_L; j = 1, \ldots, n_H) \\
\boldsymbol{K}_{HL} &= \rho\sigma_L e^{-\sum_{k=1}^{s} \gamma_L \left\| x_H^{k,i} - x_L^{k,j} \right\|^2} \\
&\quad (i = 1, \ldots, n_H; j = 1, \ldots, n_L) \\
\boldsymbol{K}_{HH} &= \rho^2\sigma_L e^{-\sum_{k=1}^{s} \gamma_L \left\| x_H^{k,i} - x_H^{k,j} \right\|^2} + \sigma_H e^{-\sum_{k=1}^{s} \gamma_H \left\| x_H^{k,i} - x_H^{k,j} \right\|^2} \\
&\quad (i, j = 1, \ldots, n_H)
\end{aligned} \tag{22}$$

where parameter s is the dimension of the input data. The $\boldsymbol{K}_{\mathrm{LL}}$ and $\boldsymbol{K}_{\mathrm{HH}}$ models are designed to model each fidelity dataset separately, whereas the $\boldsymbol{K}_{\mathrm{LH}}$ and $\boldsymbol{K}_{\mathrm{HL}}$ models account for the intercorrelation between the two fidelity datasets. ρ, σ_L, σ_H, γ_L, and γ_H are hyperparameters that must be optimized using an optimization algorithm to find the best separating hyperplane. In the MFSVC model, the SMO algorithm is also used to solve the QP problem. The SMO algorithm of the MFSVC model is similar to that of the SMO algorithm used in the SVC model mentioned above. Specifically, the optimal α_i^{best} and b^{best} of the MFSVC model are determined iteratively, and the iterative equations are presented below:

$$\begin{aligned}
\alpha_2' &= \alpha_2 + \frac{Y_2\left\{[f_{MF}(\boldsymbol{X}_1) - Y_1] - [f_{MF}(\boldsymbol{X}_2) - Y_2]\right\}}{K_{MF}(\boldsymbol{X}_1, \boldsymbol{X}_1) + K_{MF}(\boldsymbol{X}_2, \boldsymbol{X}_2) - 2K_{MF}(\boldsymbol{X}_1, \boldsymbol{X}_2)} \\
\alpha_2'' &= \begin{cases} H, & \alpha_2' > H \\ \alpha_2', & L \le \alpha_2' \le H \\ L, & \alpha_2' < L \end{cases} \\
\alpha_1'' &= \alpha_1 + Y_1 Y_2(\alpha_2 - \alpha_2'') \\
b_1' &= -E_1 - Y_1 K_{MF}(\boldsymbol{X}_1, \boldsymbol{X}_1)(\alpha_1'' - \alpha_1) \\
&\quad - Y_2 K_{MF}(\boldsymbol{X}_2, \boldsymbol{X}_1)(\alpha_2'' - \alpha_2) + b \\
b_2' &= -E_2 - Y_1 K_{MF}(\boldsymbol{X}_1, \boldsymbol{X}_2)(\alpha_1'' - \alpha_1) \\
&\quad - Y_2 K_{MF}(\boldsymbol{X}_2, \boldsymbol{X}_2)(\alpha_2'' - \alpha_2) + b \\
b'' &= \frac{b_1' + b_2'}{2}
\end{aligned} \tag{23}$$

where the parameters $\{H, L\}$ in (23) are the same as those in (14), but $y_i = Y_i$. After obtaining the optimal parameters, the prediction for a new HF sample point can be obtained by

$$\hat{\boldsymbol{y}}_{new} = sign\left(\sum_{i=1}^{n_L+n_H} \alpha_i^{best} Y_i \boldsymbol{K}(\boldsymbol{X}_i, \boldsymbol{x}_{new}) + b^{best}\right) \tag{24}$$

where $K(\boldsymbol{X}_i, \boldsymbol{x}_{new})$ is defined as follows:

$$\begin{aligned}
&\boldsymbol{K}(\boldsymbol{X}_i, \boldsymbol{x}_{new}) \\
&= \begin{cases} \rho\sigma_L e^{-\sum_{k=1}^{s} \gamma_L \left\| x_L^{k,i} - x_{new}^{k} \right\|^2} \\ \rho^2\sigma_L e^{-\sum_{k=1}^{s} \gamma_L \left\| x_H^{k,j} - x_{new}^{k} \right\|^2} + \sigma_H e^{-\sum_{k=1}^{s} \gamma_H \left\| x_H^{k,j} - x_{new}^{k} \right\|^2} \end{cases} \\
&\qquad (i = 1, 2, \ldots, N_L; j = 1, 2, \ldots, N_H)
\end{aligned} \tag{25}$$

where N_{L} and N_{H} are the number of LF and HF samples in SV.

In the MFSVC, $\boldsymbol{\theta} = \{C, \rho, \sigma_L, \sigma_H, \gamma_L, \gamma_H\}$ is a hyperparameter that must be determined during the model construction step. To obtain the optimal classification hyperplane, it is necessary to choose a classification metric as the training error indicator to optimize the various hyperparameters. Choosing a training error metric to optimize the hyperparameters of the model is crucial because it directly affects the predictive performance of the model. A good training error metric can improve the accuracy of the model classification and analyze the input data as a whole. Here, the classification model metric, accuracy (ACC), was selected as the training error metric. The effect of the error metric on the performance of the MFSVC method is presented in the next section. Therefore, the optimization objective function of the MFSVC model to determine the hyperparameters can be expressed as

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} ACC(\boldsymbol{y}_H, \hat{\boldsymbol{y}}_H) \tag{26}$$

where $\hat{\boldsymbol{\theta}}$ is the hyperparameter result obtained by optimizing the objective function. The function ACC(*) is the classification performance metric function that is consists of the training input value $\boldsymbol{y}_{\mathrm{H}}$ and predicted output value $\hat{\boldsymbol{y}}_H$. For the MFSVC, the ACC is calculated by

$$\begin{aligned}
\mathrm{ACC}\left(\boldsymbol{y}_H, \hat{\boldsymbol{y}}_H\right) &= \frac{TP + TN}{TP + TN + FP + FN} \\
TP &= card(A), A = \left\{y = 1 \,\middle|\, y \in \hat{\boldsymbol{y}}_H \wedge y \in \boldsymbol{y}_H\right\}
\end{aligned}$$

$$TN = card\,(B)\,, B = \left\{ y = -1 \left| y \in \hat{\boldsymbol{y}}_H \wedge y \in \boldsymbol{y}_H \right. \right\}$$
$$FP = card\,(C)\,, C = \left\{ y = 1 \left| y \notin \hat{\boldsymbol{y}}_H \wedge y \in \boldsymbol{y}_H \right. \right\}$$
$$FN = card\,(D)\,, D = \left\{ y = -1 \left| y \notin \hat{\boldsymbol{y}}_H \wedge y \in \boldsymbol{y}_H \right. \right\} \tag{27}$$

Simultaneously, it is crucial to use an optimization algorithm to optimize various hyperparameters to obtain the best classification hyperplane. Among numerous optimization algorithms, the Genetic Algorithm (GA) is a population-based metaheuristic algorithm based on the principles of biological evolution [37], [38]. Compared with other optimization algorithms, GA can evaluate multiple solutions in the search space, making it easier to reach a global optimal solution. Additionally, the algorithm is simple, versatile, and widely applicable [39]. GA utilizes the concept of ''survival of the fittest'' by repeatedly applying genetic operators to individuals in the population to generate new populations. The number of populations (M), The representation of individuals, selection process, crossover probability (C_P), mutation probability (M_p), iterations (N), and fitness function calculation are crucial in GA. Therefore, the GA is used to optimize the hyperparameters of SVC and MFSVC to predict the data, in which the GA parameter settings are given later.

C. TREATMENT OF IMBALANCE DATA

Imbalanced training datasets often result in decreased prediction accuracy for classification models or difficulty in capturing the correct distribution of minority classes. Therefore, preprocessing of the training dataset is necessary before training the model, which involves data balancing. The SMOTEENN algorithm is a hybrid sampling algorithm that overcomes the disadvantages of both oversampling and undersampling. Firstly, it generates synthetic samples, which helps achieve a balanced ratio in the dataset. The Edited Nearest Neighbor algorithm (ENN) [40] is then used to eliminate noisy samples, ensuring that the dataset is devoid of noise and inconsistencies. By combining these two techniques, the SMOTEENN algorithm effectively balanced the accurate capture of minority class patterns and the preservation of valuable information in the dataset.

In this paper, the synthetic samples are generated in the following manner:

$$X_{new} = X_i + \left(X_i^* - X_i \right) * rand \tag{28}$$

where X_{new} represents synthetic data, and its response value y is the same as that of the minority class samples, X_i denotes a selected minority class sample, X_i^* refers to one of the k-nearest neighbors [41] from X_i, *rand* retains a random number ranging from 0 to 1. The scheme of the ENN algorithm is as follows: 1) An initial training dataset is established, denoted as $\boldsymbol{T}$, which needs to be edited in order to improve its quality; 2) In each editing iteration, the k-nearest neighbor algorithm is applied to each sample point x in the dataset $\boldsymbol{T}$ and the samples that do not conform to the k-nearest neighbor rule are discarded; (3) This process continues until no more samples can be deleted. It is worth noting that the k value in the k-nearest neighbor algorithm is a parameter that must be specified beforehand. In this study, the k value of the k-nearest neighbor algorithm was 3.

D. FLOWCHART AND IMPLEMENTATION

Unlike the SVC model, which can only use a single-fidelity data source for model training, this model utilizes two-fidelity data sources to form a covariance matrix, thereby establishing an MFSVC model. To address the problem of data imbalance caused by missing or difficult-to-obtain data in various practical engineering problems, this model establishes an enhanced MFSVC model (EMFSVC) combined with a data-balancing algorithm. This model introduces a genetic algorithm to optimize the various hyperparameters of the EMFSVC model to determine the optimal classification hyperplane for the dataset. To facilitate comprehension of the proposed method, its application in the field of engineering statics is demonstrated and a flowchart is produced (see Fig. 1). It can be noted that the model is mainly divided into two parts: 1) The initial LF/HF sample points should be generated in the DOE considering the design parameters of the geometric model taken into account. Subsequently, the sample points were used to establish the corresponding geometric model and mesh it. The corresponding simulation results were obtained through the application of simulation calculations, with class labels assigned in accordance with the magnitude of the simulation results. For example, when the stress experienced exceeds the allowable stress and failure occurs, the sample point can be labelled as 0, and conversely, as 1. 2) Once all the sample points have been assigned class labels, the SMOTEENN algorithm is applied to the LF and HF sample points for resampling to obtain the training set. Subsequently, the range of hyperparameter values and parameters for the GA are set, and the initial hyperparameters and covariance matrix of the MFSVC model are determined based on (20) and (21). The SMO algorithm was then utilized to optimize these parameters, leading to the identification of the optimal parameters $\{\alpha, b\}$ for the MFSVC model. The GA was applied to optimize the hyperparameters of the MFSVC model. This involves computing the training error metric of the MFSVC model, which measures the difference between the output $\boldsymbol{y}_{\mathrm{H}}$ of the training sample points and predicted output $\hat{\boldsymbol{y}}_H$. The EMFSVC model and the failure boundary were ultimately obtained by determining the best individual through GA optimization.

III. ANALYTICAL PROBLEMS

The prediction performance of the MFSVC and EMFSVC models was verified and compared with the three classification models through numerical examples in this section.

A. STEUP

To verify the effectiveness of the proposed method, analytical problems were first solved. The hardware and software environment used in this experiment was as follows: (1)

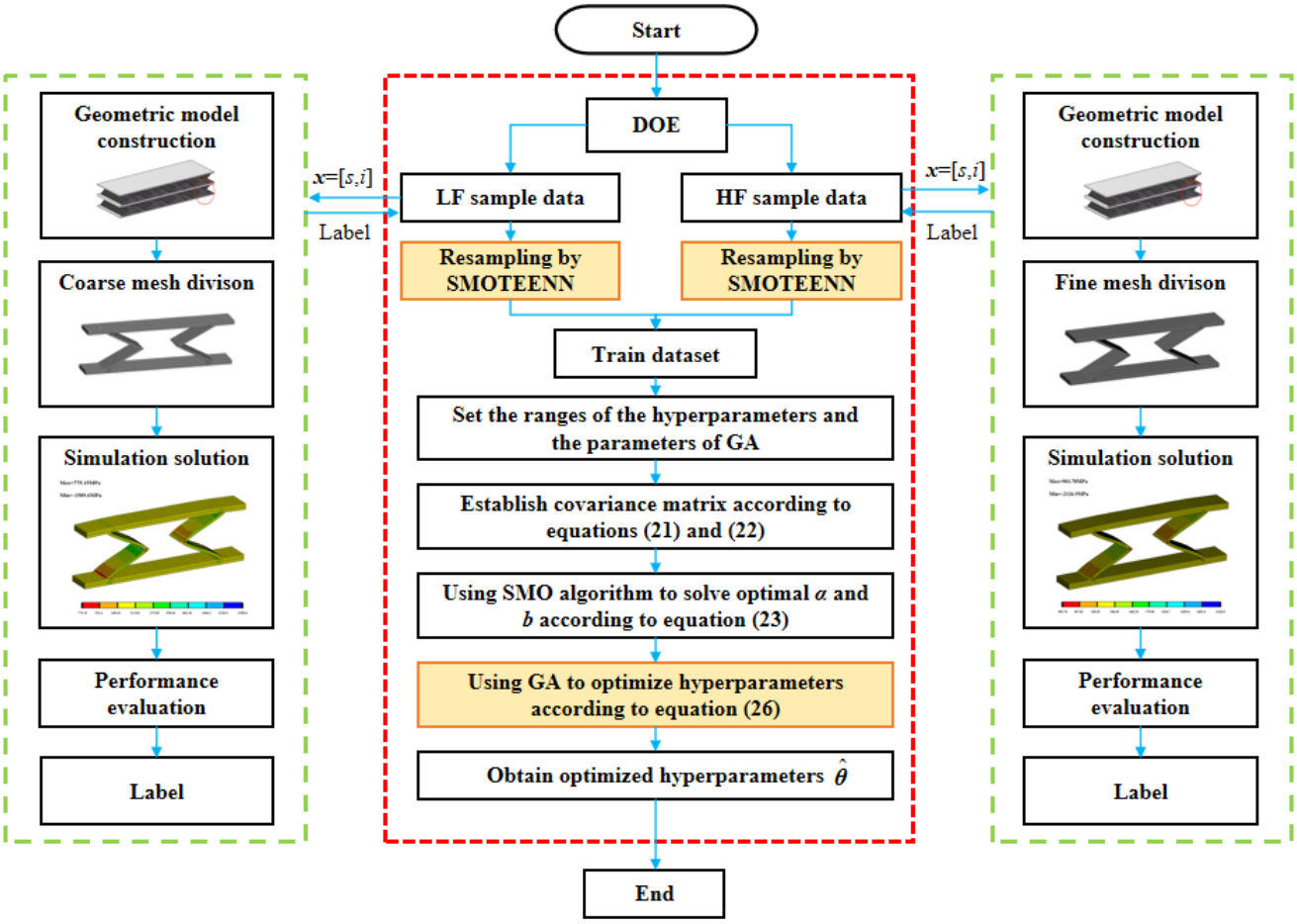


FIGURE 1. Flowchart of the EMFSVC model in the engineering problem.

Hardware environment: workstation with two Intel Xe-on E5-2680 v4 processors and 128 GB memory. (2) Software environment: Windows 10, Windows Subsystem for Linux 2, MINICONDA3, g++. (3) Third-party packages: NumPy, SciPy, Theano, pygpu, Pymc3, pyDOE, scikit-learn, mkl, and imbalanced-learn. The python language was mainly used to implement various models in this experiment, and multiple numerical examples were used for research.[1] This experiment mainly studied the classification model, so the commonly used evaluation indicators of the classification model were precision (P), recall rate (R), and F1 score (F1) [42], [43], [44]. The computational cost of the experiment was calculated using the time (T) spent on the model training.

$$\begin{aligned} \mathrm{P} &= \frac{TP}{TP+FP} \\ \mathrm{R} &= \frac{TP}{TP+FN} \\ \mathrm{F1} &= 2 \times \frac{P \times R}{P+R} \end{aligned} \tag{29}$$

where TP is true positives, which is the number of data points correctly classified from the positive class; TN is true negatives, which is the number of data points correctly predicted from the negative class; FP is false positives, which is the number of data points classified to be in the positive class but in fact belonging to the negative class; FN is false negatives, which is the number of data points classified as negative but in fact belonging to the positive class.

To verify the effectiveness and efficiency of the proposed method, the following models were compared: (1) The MFGPC model was chosen for comparison [26], as it is the only multifidelity classification method reported in the open literature. In addition, GPC was compared. For the construction of the GPC and MFGPC, a No-U-Turn sampler was utilized. It uses a chain with a target acceptance probability of 0.95 and 0.99. The first 3000 sampling points were discarded to adjust the sampler's step size, and the last 1000 sampling points were used for analysis. (2) The SVC model was included for comparison to demonstrate the efficiency of MFSVC in fusing the multi-fidelity data. SVC also utilizes GA to optimize the hyperparameter $\{C, \gamma\}$. In this numerical experiment, it was set that the cost of obtaining an expensive function evaluation was five times higher than that of a cheap

[1]The code to reproduce the experimental results can be found via https://github.com/YouweiUSC/MFSVC

function evaluation, meaning that the multifidelity model was composed of a dataset consisting of 10 expensive function evaluations and 50 cheap function evaluations. By contrast, the single-fidelity model is composed of a dataset consisting of 20 expensive function evaluations. In addition, 1000 test sample points were generated via Latin hypercube sampling (LHS) and used to test model accuracy. Following the principles of the GA, in this experiment, the hyperparameters $\{C, \rho, \sigma_L, \sigma_H, \gamma_L, \gamma_H\}$ of the MFSVC model are set as genes in the population, and the fitness function (i.e., the training error metric) is determined later. The ranges of the hyperparameters are {[0,500], [0,1], [0,1], [0,1], [0,10], [0,10]}. The parameters of the GA were set to $\{M = 40, N = 10, Cp = 0.9, Mp = 0.1\}$

First, to determine the optimal training error metric for the MFSVC model, the MFSVC model with each classification evaluation index as the training error metric was trained in numerical examples of balanced data. Subsequently, to verify the effectiveness and efficiency of the MFSVC model, the MFSVC model was compared with other algorithms using balanced data. To validate the effectiveness of the proposed method in imbalanced data, an imbalanced dataset was set up with class proportions {1:9, 2:8, 3:7, 4:6} in the numerical functions. Because precision engineering is expensive for generating high-fidelity samples, it's mostly small-sample datasets. The minority classes in a high-fidelity dataset may not exist or may not be able to satisfy the number of samples required for resampling by the SMOTENN algorithm when there is an extreme imbalance rate, so this experiment focused on medium to high imbalance level dataset. The numerical test functions used in this study are listed in Table 1.

TABLE 1. Numerical test functions.

name	function	Design Space
Branin function [45]	$y_H = \begin{cases} 1, if\,(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10(1-\frac{1}{8\pi})\cos(x_1) + 10 > 50 \\ -1, if\,(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10(1-\frac{1}{8\pi})\cos(x_1) + 10 < 50 \end{cases}$ $y_L = \begin{cases} 1, if\,1.1(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10(1-\frac{1}{8\pi})\cos(x_1) + 10 > 50 \\ -1, if\,1.1(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10(1-\frac{1}{8\pi})\cos(x_1) + 10 < 50 \end{cases}$	$x_i \in [0,1], i = 1,2$
Failure 2 function [46]	$y_H = \begin{cases} 1, & if \min\left\{ \frac{3.7 - (12x_2 - 6) + \exp[-(12x_1 - 6)^2/10] + [(12x_1 - 6)/5]^4}{11.045 - (12x_2 - 6)(12x_1 - 6)} \right\} > 0 \\ -1, & if \min\left\{ \frac{3.7 - (12x_2 - 6) + \exp[-(12x_1 - 6)^2/10] + [(12x_1 - 6)/5]^4}{11.045 - (12x_2 - 6)(12x_1 - 6)} \right\} < 0 \end{cases}$ $y_H = \begin{cases} 1, & if \min\left\{ \frac{3.8 - (12x_2 - 6) + \exp[-(12x_1 - 6)^2/10] + [(12x_1 - 6)/4.8]^4}{11.055 - (12x_2 - 6)(12x_1 - 6)} \right\} > 0 \\ -1, & if \min\left\{ \frac{3.8 - (12x_2 - 6) + \exp[-(12x_1 - 6)^2/10] + [(12x_1 - 6)/4.8]^4}{11.055 - (12x_2 - 6)(12x_1 - 6)} \right\} < 0 \end{cases}$	$x_i \in [0,1], i = 1,2$
Costabal function [26]	$y_H = \begin{cases} 1, & if\,(0.5 + \sin(2.5\pi x_1)/3 - x_2) > 0 \\ -1, & if\,(0.5 + \sin(2.5\pi x_1)/3 - x_2) > 0 \end{cases}$ $y_L = \begin{cases} 1, & if\,(0.55 + \sin(2.5\pi x_1)/2.5 - 1.2x_2) > 0 \\ -1, & if\,(0.55 + \sin(2.5\pi x_1)/2.5 - 1.2x_2) > 0 \end{cases}$	$x_i \in [0,1], i = 1,2$
Hartmann 3 function [45]	$y_H = \begin{cases} 1, & if\,-\sum_{i=1}^{4}\alpha_i \exp[-\sum_{j=1}^{3}\beta_{ij}(x_j - p_{ij})^2] > -15 \\ -1, & if\,-\sum_{i=1}^{4}\alpha_i \exp[-\sum_{j=1}^{3}\beta_{ij}(x_j - p_{ij})^2] < -15 \end{cases}$ $y_L = \begin{cases} 1, & if\,-\sum_{i=1}^{4}\alpha_i \exp[-\sum_{j=1}^{3}\beta_{ij}(x_j - 1.05p_{ij})^2] > -15 \\ -1, & if\,-\sum_{i=1}^{4}\alpha_i \exp[-\sum_{j=1}^{3}\beta_{ij}(x_j - 1.05p_{ij})^2] < -15 \end{cases}$ $\alpha = \begin{bmatrix} 1 \\ 1.2 \\ 3 \\ 3.2 \end{bmatrix}$ $\beta = \begin{bmatrix} 3.01030 \\ 0.11035 \\ 3.01030 \\ 0.11035 \end{bmatrix}$ $p = \begin{bmatrix} 0.3689 & 0.1170 & 0.2673 \\ 0.4699 & 0.4387 & 0.7470 \\ 0.1091 & 0.8732 & 0.5547 \\ 0.0381 & 0.5743 & 0.8828 \end{bmatrix}$	$x_i \in [0,1], i = 1,2,3$

B. RESULTS

1) INFLUENCE OF ERROR METRIC ON THE PERFORMANCE OF MFSVC

Before comparing with other models, determining the error measurement used in the determination of the hyperparameters of the MFSVC model is one of the steps in construction, because it directly affects the predictive performance of the model. In this study, four classification indicators were applied as error measurements to conduct experiments and verify the accuracy of error measurement on the MFSVC model. Taking four analytic problems as examples, a balanced dataset was established for training, consisting of 50 LF and 10 HF sample points, to reduce the impact of imbalanced data on model accuracy during the training process. Ten runs were performed to obtain the mean and standard deviation (SD) of the performance metrics, which are summarized in Table 2. The best results are shown in bold font. The boxplot shown in Fig. 2 is further utilized to illustrate the performance metrics over 10 independent runs on the Costabal problem.

From Table 2, it can be observed that when ACC is used as the error measurement, the predictive performance of the MFSVC model is generally better than that of the other three metrics. In the Branin function, the classification indicator

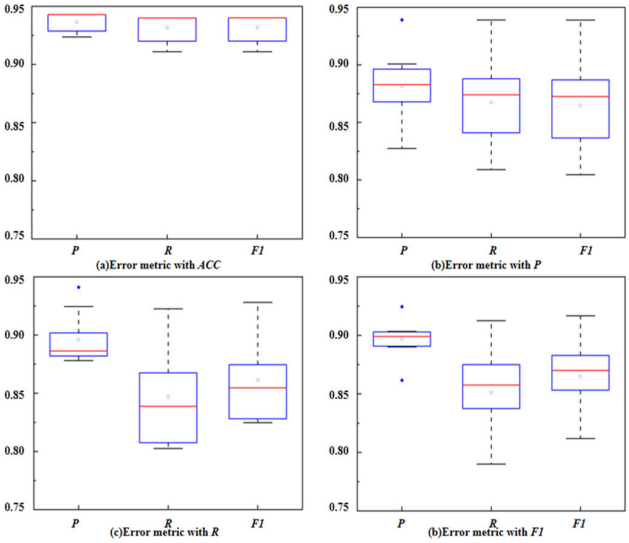


FIGURE 2. Boxplot of multiple error metrics on MFSVC in the Costabal function with balanced data.

{F1 = 0.987} of the MFSVC model with ACC as the training error metric was 0.065 times higher than that of the MFSVC model with R as the training error measurement

TABLE 2. Results of multiple error metrics on MFSVC in the numerical functions with balanced data.

Branin function					Failure 2 function				
Training metric		P	R	F1	Training metric		P	R	F1
ACC	Mean	**0.989**	**0.986**	**0.987**	ACC	Mean	**0.939**	**0.919**	**0.922**
	SD	**0.006**	**0.010**	**0.008**		SD	**0.017**	**0.037**	**0.033**
P	Mean	0.966	0.940	0.946	P	Mean	0.926	0.889	0.896
	SD	0.009	0.027	0.021		SD	0.020	0.044	0.041
R	Mean	0.962	0.904	0.922	R	Mean	0.926	0.897	0.901
	SD	0.012	0.072	0.053		SD	0.019	0.049	0.045
F1	Mean	0.976	0.956	0.961	F1	Mean	0.916	0.874	0.881
	SD	0.012	0.035	0.028		SD	0.018	0.048	0.044
Costabal function					Hartmann 3 function				
Training metric		P	R	F1	Training metric		P	R	F1
ACC	Mean	**0.937**	**0.932**	**0.932**	ACC	Mean	**0.947**	**0.945**	**0.945**
	SD	**0.008**	**0.011**	**0.011**		SD	**0.002**	**0.005**	**0.003**
P	Mean	0.882	0.867	0.865	P	Mean	0.939	0.923	0.927
	SD	0.031	0.039	0.041		SD	0.006	0.014	0.012
R	Mean	0.868	0.847	0.843	R	Mean	0.938	0.919	0.925
	SD	0.023	0.032	0.034		SD	0.008	0.017	0.014
F1	Mean	0.875	0.855	0.852	F1	Mean	0.939	0.921	0.926
	SD	0.025	0.037	0.040		SD	0.006	0.014	0.012

{F1 = 0.922}. In the Failure 2 function, the variances of the classification indicators {0.017, 0.037, 0.033} of the MFSVC model with ACC as the training error measurement were all smaller than the variances of the classification indicators {0.018, 0.048, 0.044} of the MFSVC model with F1 as the error measurement. In the Hartmann 3 function, the MFSVC model demonstrates a better predictive performance because it utilizes the ACC metric as the training error metric compared to other metrics.

From the boxplot above, it can be seen that in the numerical examples, the MFSVC model with ACC as the model training error measurement has significantly better predictive performance than the other three MFSVC models, and the model's robustness is also good. As shown in Figure 2, the predictive performance indicators {P, R, F1} of the MFSVC model with ACC as the error measurement were mainly distributed between 0.90 and 0.95, while the predictive performance indicators of the other three models were lower than or equal to 0.90. Based on the above investigation, it can be concluded that the MFSVC model with ACC as the training error measurement has better predictive performance and robustness. Therefore, ACC was used as the training error measurement for both the MFSVC and SVC models in the following experiments to ensure consistency in optimizing the training error indicators of the models.

2) RESULT ON NUMERICAL FUNCTION WITH BALANCE DATA

After establishing the training error measurement, the predictive performance of each model was compared by using multiple numerical examples to verify the effectiveness of the MFSVC model. The sample point setting process is described in the literature [26]. The single-fidelity model uses a dataset consisting of 20 HF samples, and the dataset with 50 LF samples and 10 HF samples was used in the construction of the multi-fidelity model. Table 3 compares the performance metrics of the compared methods with respect to the analytic problems. The best results are shown in bold font.

According to the observations in Table 3, the classification metrics {P, R, F1} of the MFSVC model were significantly better than those of the other three classification models. Although the cost of the MFSVC model increased by tens of seconds or even approximately one hundred seconds compared to that of the SVC model, it is still within an acceptable range. Comparing the SVC model and the MFSVC model, in the Hartmann 3 function, the evaluation metric {F1=0.945} of the MFSVC model improved by about 14% compared to the evaluation metric {F1=0.824} of the SVC model; in the Failure 2 function, the evaluation metric {F1=0.932} of the MFSVC model also improved by about 11% compared to the evaluation metric {F1=0.833} of the SVC model. Comparing the MFSVC model and the

TABLE 3. Results of multiple models in the numerical functions with balanced data.

Function	Evaluation metric	MFSVC	MFGPC	SVC	GPC
Branin function	P	**0.989**	0.968	0.965	0.929
	R	**0.986**	0.944	0.966	0.945
	F1	**0.987**	0.948	0.965	0.932
	T(s)	85.496	243.557	**2.567**	85.844
Failure 2 function	P	**0.944**	0.936	0.896	0.871
	R	**0.929**	0.923	0.822	0.844
	F1	**0.932**	0.926	0.833	0.851
	T(s)	146.689	336.540	**15.852**	69.383
Costabal function	P	**0.937**	0.925	0.913	0.828
	R	**0.932**	0.922	0.913	0.818
	F1	**0.932**	0.922	0.913	0.817
	T(s)	278.862	321.277	**19.604**	66.741
Hartmann 3 function	P	**0.947**	0.920	0.906	0.907
	R	**0.945**	0.922	0.795	0.796
	F1	**0.945**	0.919	0.824	0.825
	T(s)	78.108	378.700	**5.805**	60.919

MFGPC model, in the Branin function, the evaluation metric {F1=0.987} of the MFSVC model increased by 4% compared to the evaluation metric {F1=0.948} of the MFGPC model, and the cost T {85.496} of the MFSVC model also decreased by approximately 65% compared to the cost T {243.557} of the MFGPC model; in the Hartmann 3 function, the classification metric {F1=0.945} of the MFSVC model improved by 0.026 compared to the metric {F1=0.919} of the MFGPC model, and the cost T {78.108} of the MFSVC model also decreased by approximately 79% compared to the T {378.700} of the MFGPC model. To demonstrate the advantages of the predictive performance of the MFSVC model intuitively, the predicted boundaries from the four classification models in the balanced data of the Branin function are shown in Fig. 3.

From Fig. 3, it can be seen that the predicted boundary of the MFSVC model is closer to the HF boundary than that of the MFGPC model. Compared to the SVC model, the predicted boundary of the MFSVC model can also better fit the HF boundary trend. By summarizing the results of the numerical examples in balanced data, it can be concluded that in this numerical example experiment, the MFSVC model has better predictive boundary performance than the SVC and MFGPC models, the cost is also lower than that of the MFGPC model, and the cost increase compared to the SVC model is acceptable.

3) RESULT ON NUMERICAL FUNCTION WITH IMBALANCE DATA

Through the above numerical experiments with balanced data, the MFSVC model was verified to have better predictive boundary performance than the other three classification models. However, in practical situations, data are often difficult to obtain or miss, resulting in data imbalance, which leads to a decrease in predictive accuracy or difficulty in capturing the intrinsic characteristics of engineering systems using most predictive boundary models. Therefore, EMFSVC was developed in this study to solve the problem of imbalanced data. The effectiveness of EMFSVC was demonstrated by comparing it with alternative classification models on datasets with various imbalance ratios. After LHS sampling, 1000 sample points, 50 LF sample points, and 10 HF sample points were chosen as a multi-fidelity dataset based on the imbalanced ratio, and 20 HF sample points were selected as a single fidelity dataset based on the imbalanced ratio. For example, when the imbalance ratio was set to 1:9, the single-fidelity dataset consisted of two HF sample points with $\{y = 1\}$ and 18 HF sample points with $\{y = -1\}$. On the other hand, the multi-fidelity dataset consists of five LF sample points with $\{y = 1\}$ and 45 LF sample points with $\{y = -1\}$, making a total of 50 LF sample points, one HF sample point with $\{y = 1\}$, and nine HF sample points with $\{y = -1\}$, resulting in a total of 10 HF sample points. Table 5 summarizes the performance metrics of the compared methods for analytic problems on imbalanced data. The best results are bolded, and the second ones are underlined.

TABLE 4. Results of the multiple models in the Branin function with imbalanced data.

Performance metric	Method	Imbalance ratio 1:9	2:8	3:7	4:8
P	GPC	0.901	<u>0.959</u>	<u>0.968</u>	0.961
	MFGPC	**0.978**	**0.978**	0.966	0.958
	SVC	0.949	0.946	0.957	0.953
	MFSVC	0.949	0.950	0.960	<u>0.964</u>
	EMFSVC	<u>0.960</u>	<u>0.959</u>	**0.970**	**0.969**
R	GPC	<u>0.933</u>	<u>0.943</u>	<u>0.939</u>	0.905
	MFGPC	**0.969**	**0.969**	0.934	0.890
	SVC	0.836	0.806	0.896	0.869
	MFSVC	0.834	0.879	0.905	<u>0.929</u>
	EMFSVC	0.905	0.905	**0.953**	**0.939**
F1	GPC	0.913	0.914	0.941	0.908
	MFGPC	<u>0.920</u>	<u>0.920</u>	<u>0.944</u>	0.911
	SVC	0.871	0.849	0.914	0.895
	MFSVC	0.871	0.902	0.922	<u>0.939</u>
	EMFSVC	**0.921**	**0.921**	**0.958**	**0.948**

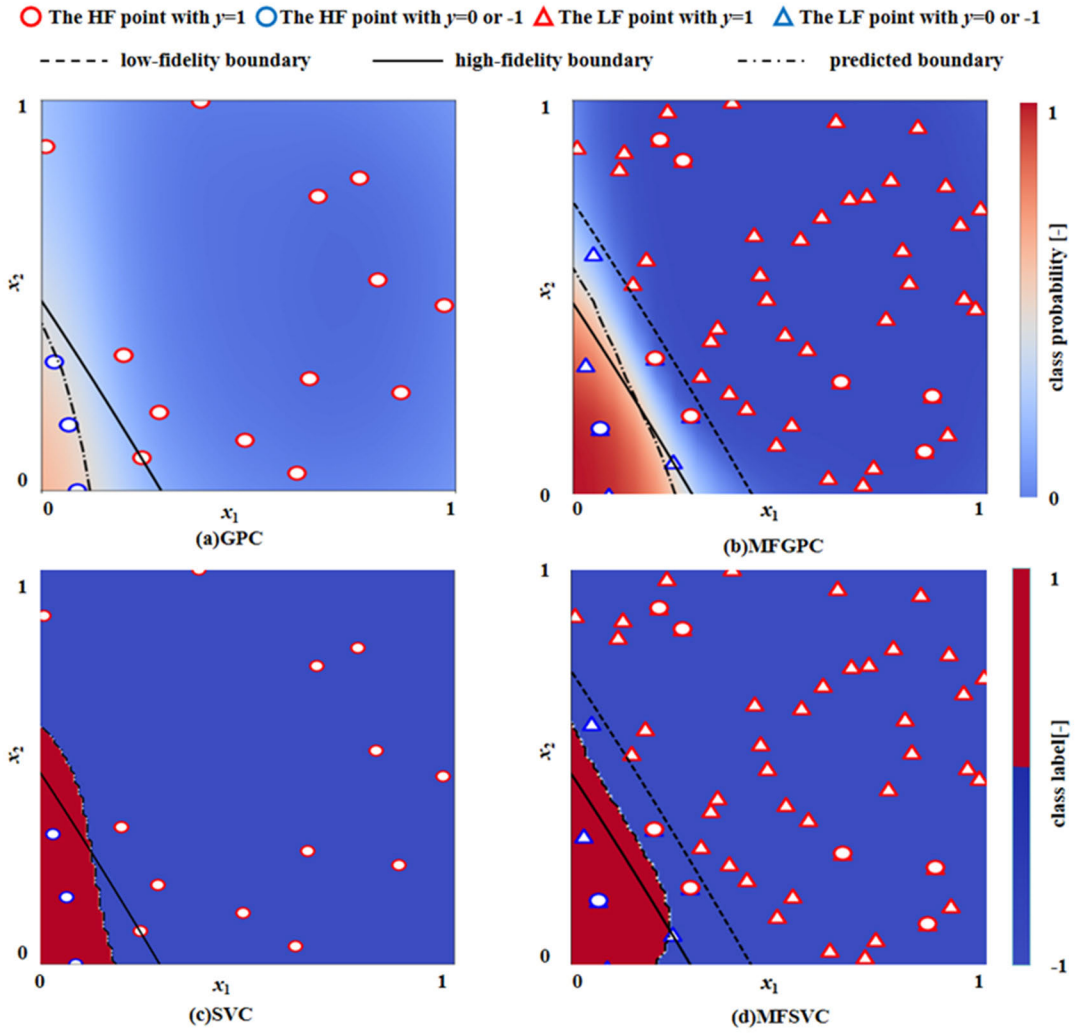


FIGURE 3. Result of multiple models in the Branin function with balanced data.

According to the above tables, the proposed method has a better overall predictive performance than other models at different imbalance ratios. In the Branin function, regardless of the imbalanced ratio, the proposed method always has better classification metrics {F1} than the other four classification models, and when the imbalanced ratio tends to be balanced, the classification metrics {P, R} of EMFSVC also become better than the other four classification models. In the Failure 2 function, EMFSVC has a better classification metric {F1} than the other models when the imbalance ratio is low. In the Costabal function, regardless of the imbalanced ratio, the EMFSVC always has a better classification metric {F1} than the other classification models, and the same is true in the Hartmann 3 function. The table also shows that in some cases, the classification performance of EMFSVC is not necessarily better than that of the other models. For example, in the failure 2 function, when the imbalanced ratio of the dataset was 4:8, the classification metrics {P=0.936, R=0.915, F1=0.919} of the MFGPC model were better than the classification metrics {P=0.933, R=0.914, F1=0.917} of the EMFSVC. The reason may be that although the SMOTEENN algorithm can effectively handle class imbalance problems to improve the predictive accuracy of the model, when the distribution of the minority class samples is uneven or overlaps, the SMOTEENN algorithm may generate noisy data or mistakenly delete data, which leads to a decrease in the predictive performance of the model. Overall, the proposed method demonstrates better predictive performance and stability compared to other models. To intuitively observe the changes in the predicted boundaries of the EMFSVC and other models in imbalanced data, the predicted boundary results of the EMFSVC and MFSVC models in the Costabal function are shown as an example in Fig. 4.

From Fig. 4, it can be observed that the predicted boundary of the proposed method in imbalanced data is closer to the HF boundary than the predicted boundary of the MFSVC model. When the imbalance ratio is {1:9, 2:8}, the peak values of the predicted boundary of the proposed method are close to the

TABLE 5. Results of the multiple models in the Failure 2 function with imbalanced data.

Evaluation metric	Model	Imbalance ratio 1:9	2:8	3:7	4:8
P	GPC	0.289	0.831	0.808	0.869
	MFGPC	**0.888**	0.888	0.917	**0.936**
	SVC	0.833	0.919	0.896	0.904
	MFSVC	0.864	0.932	0.924	0.917
	EMFSVC	0.870	**0.944**	**0.926**	0.933
R	GPC	0.272	0.464	0.507	0.813
	MFGPC	0.793	0.793	0.870	**0.915**
	SVC	0.840	0.918	0.898	0.868
	MFSVC	0.864	0.931	0.920	0.869
	EMFSVC	**0.870**	**0.940**	**0.924**	0.914
F1	GPC	0.140	0.432	0.493	0.822
	MFGPC	0.807	0.807	0.877	**0.919**
	SVC	0.830	0.918	0.894	0.875
	MFSVC	0.848	0.929	0.915	0.877
	EMFSVC	**0.855**	**0.940**	**0.921**	0.917

TABLE 6. Results of the multiple models in the Costabal function with imbalanced data.

Evaluation metric	Model	Imbalance ratio 1:9	2:8	3:7	4:8
P	GPC	0.456	0.785	0.771	0.807
	MFGPC	0.810	0.852	0.894	**0.907**
	SVC	0.668	0.895	0.891	0.811
	MFSVC	0.867	0.902	0.888	0.868
	EMFSVC	**0.886**	**0.912**	**0.910**	0.894
R	GPC	0.498	0.636	0.670	0.782
	MFGPC	0.689	0.792	0.869	**0.892**
	SVC	0.615	0.885	0.890	0.809
	MFSVC	0.821	0.889	0.887	0.844
	EMFSVC	**0.885**	**0.904**	**0.908**	0.889
F1	GPC	0.357	0.586	0.643	0.781
	MFGPC	0.666	0.785	0.868	0.852
	SVC	0.566	0.883	0.890	0.809
	MFSVC	0.811	0.888	0.893	0.840
	EMFSVC	**0.884**	**0.903**	**0.908**	**0.898**

peak values of the HF boundary, and the valley values of the predicted boundary of the proposed method are also close to

TABLE 7. Results of the multiple models in the Hartmann 3 function with imbalanced data.

Evaluation metric	Model	Imbalance ratio 1:9	2:8	3:7	4:8
P	GPC	0.795	0.895	0.892	0.900
	MFGPC	0.901	0.911	0.923	0.927
	SVC	0.898	0.898	0.912	0.922
	MFSVC	0.908	0.927	0.915	0.928
	EMFSVC	**0.928**	**0.940**	**0.950**	**0.937**
R	GPC	0.268	0.511	0.464	0.635
	MFGPC	0.574	0.742	0.820	0.869
	SVC	0.903	0.884	0.829	0.893
	MFSVC	**0.911**	**0.925**	0.921	0.904
	EMFSVC	0.876	0.909	**0.931**	**0.918**
F1	GPC	0.266	0.570	0.519	0.687
	MFGPC	0.636	0.782	0.846	0.884
	SVC	0.899	0.889	0.852	0.902
	MFSVC	0.902	0.914	0.915	0.911
	EMFSVC	**0.909**	**0.917**	**0.936**	**0.923**

the valley values of the HF boundary. However, the predicted boundary of the MFSVC does not exhibit the same trend as that of the HF boundary and has significant differences. When the imbalanced ratio is {3:7, 4:6}, the peak and valley values of the predicted boundary of the proposed method are closer to the peak and valley values of the HF boundary than the predicted boundary of the MFSVC model.

In conclusion, in this numerical experiment, the proposed method can effectively guarantee the predictive performance of the classification model in imbalanced datasets and even improve the predictive performance of the model. Therefore, the proposed method can effectively predict the failure boundary under the phenomenon of data imbalance and address boundary problems in engineering experiments or other application areas.

4) INFLUENCE OF THE SIZE OF HIGH-FIDELITY DATA

In numerical experiments, low-fidelity sample data are considered inexpensive and easy to obtain with negligible costs. Therefore, this experiment only investigated the impact of high-fidelity sample sizes on the proposed model in the numerical experiments. To study the effect of the high-fidelity sample size on the accuracy of the proposed model under different functions, four numerical examples were used, and the effect of the high-fidelity sample size on the accuracy of the proposed method was studied. In this experiment, the default low-fidelity sample size of 100 was set, and the impact of high-fidelity sample sizes {6, 10, 14, 18, 22, 30} on the proposed method was analyzed. The accuracy of the model was

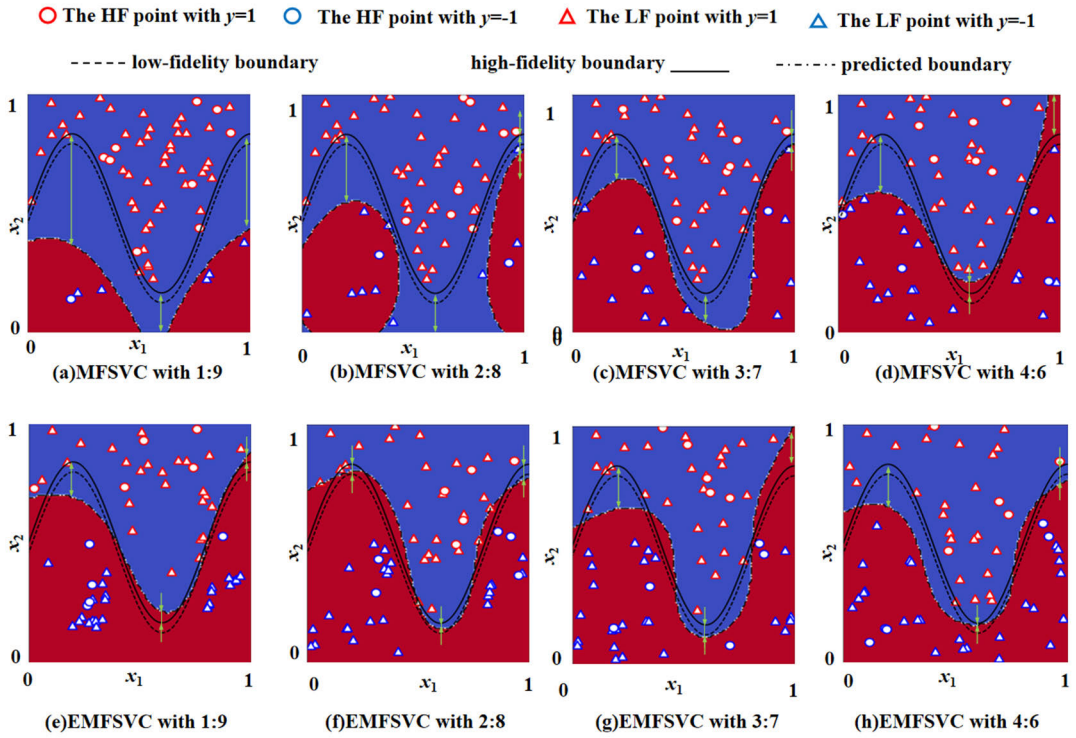


FIGURE 4. Result of imbalanced data in the Costabal function. To better clarify the key differences between the predicted boundary and the high-fidelity boundary, green solid lines have been added at the peaks and valleys of the boundary to illustrate the distance between the lines.

tested using LHS sampling with 1000 test samples. Table 9 was obtained through training on numerical examples with balanced data.

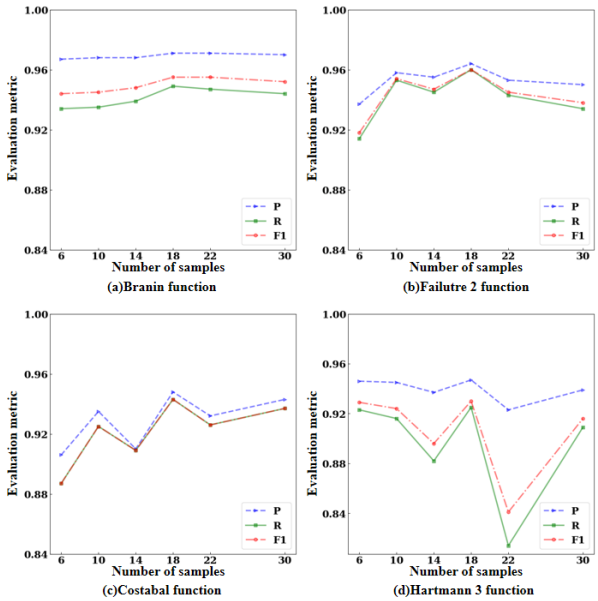


FIGURE 5. Results of the function with balanced data on the EMFSVC.

From the figures above, it can be observed that within a certain range, as the number of HF samples increases, the various classification indicators of the proposed method fluctuate. When the number of HF samples reaches a certain value and continues to increase, the various classification indicators of the proposed method decrease or fluctuate, and the upward trend may even approach zero. For example, in Fig. 5(b), when the number of HF samples reached 18, the various classification indicators of the model reached their maximum; however, when the number of HF samples continued to increase to 30, the various classification indicators of the model decreased. As shown in Fig. 5(a), when the number of HF samples increased to 18, the model's various classification measurements reached their maximum values, after which the number of HF samples continued to increase, and the model's various classification measurements started to fluctuate. To visually observe the influence of HF samples on the proposed method, experiments were conducted using the Failure 2 function, as shown in Fig. 6. As shown in Fig. 6, when the number of HF samples was less than 18, the predicted boundary of the proposed method differed significantly from that of the HF boundary. When the number of HF samples was 18, the predicted boundary of the proposed method was closest to the HF boundary, its predicted upper boundary did not intersect with the HF boundary, and its predicted lower boundary had the same trend as the HF lower boundary and was close in distance. However, when the number of HF samples continues to increase, although the proposed method's predicted boundary is close to the HF boundary, boundary crossing phenomenon

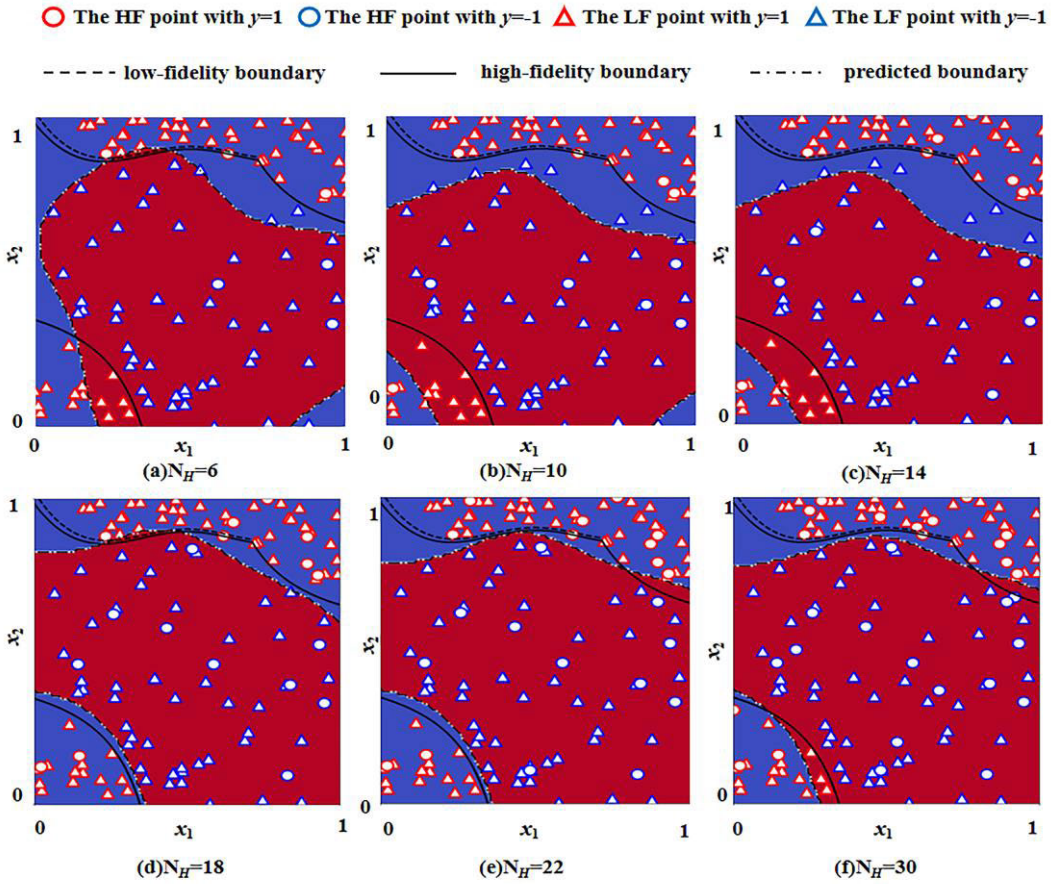


FIGURE 6. The influence of HF sample number of the proposed method in the Failure 2 function.

between the predicted boundary and the HF boundary starts to appear

In conclusion, in an imbalanced dataset, EMFSVC can better predict the boundary and has better or more stable performance in imbalanced datasets. In addition, the selection of the HF sample size directly affects the predictive boundary performance of EMFSVC. An appropriate number of HF samples can improve the classification accuracy of the prediction model, while avoiding unnecessary costs.

IV. ENGINEERING EXAMPLES

In this section, three engineering problems, including the prediction of the failure boundary of the zero Poisson ratio structure and the safe operating boundary of the axial flow compressor Rotor37 and the simulation failure boundary of modeling the isentropic efficiency of the axial compressor rotor Rotor37, are solved to verify the engineering capability of the EMFSVC.

A. THE FAILURE BOUNDARY OF ZERO POISSON RATIO STRUCTURE

EMFSVC was used to predict the failure boundaries of individual units in zero-Poisson ratio structures. A geometric model of the zero-Poisson ratio structure investigated in [45] was adopted, as shown in Fig. 7. The geometric model and design parameters of the unit cell of the zero-Poisson ratio structure are shown in Fig. 8 and Table 10, respectively. The structure was pressed using downward force. If the force is larger than the threshold value, the structure is destroyed. Therefore, predicting the safety boundary of the zero Poisson ratio structure would be beneficial for engineering applications. According to the material used in Table 9, the compressive yield strength σ_b of the zero Poisson's ratio structure was set to 250MPa in this experiment. Therefore, the maximum failure stress of the zero-Poisson's ratio structure was set to 250MPa in this experiment. Based on the data in Table 10, a static structural simulation model was constructed using the ANSYS software. A fine mesh model with 33,420 cells was used as the HF model. Accordingly, a coarse mesh model comprising 4,330 cells was used as the LF model. The HF and LF models of a unit cell with a zero Poisson ratio structure were analyzed using ANSYS software, and the stress results in the X-axis direction were obtained. The grids and simulation results of the LF and HF simulations for the unit cell of the zero-Poisson ratio structure are shown in Fig. 9 and Fig. 10, where the corresponding simulation times of the HF and LF model were 34s and 5s. Therefore, the cost ratio was 6.8, which was rounded off

to 7. The multi-fidelity model was constructed based on a dataset consisting of 10 HF and 70 LF samples. To make a fair comparison, the single-fidelity model used a dataset consisting of 20 HF samples with the identical cost of the sample data to that of the multi-fidelity model. The proportion of minority classes in the high-fidelity dataset is about 0.145, whereas in the low-fidelity dataset, it is about 0.160. To test the accuracy of the model, 1,000 samples were generated using LHS and evaluated by HF simulation with a fine mesh to obtain its associated responses.

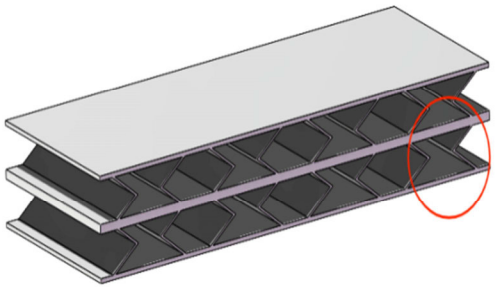


FIGURE 7. Geometric model of zero Poisson ratio structure.

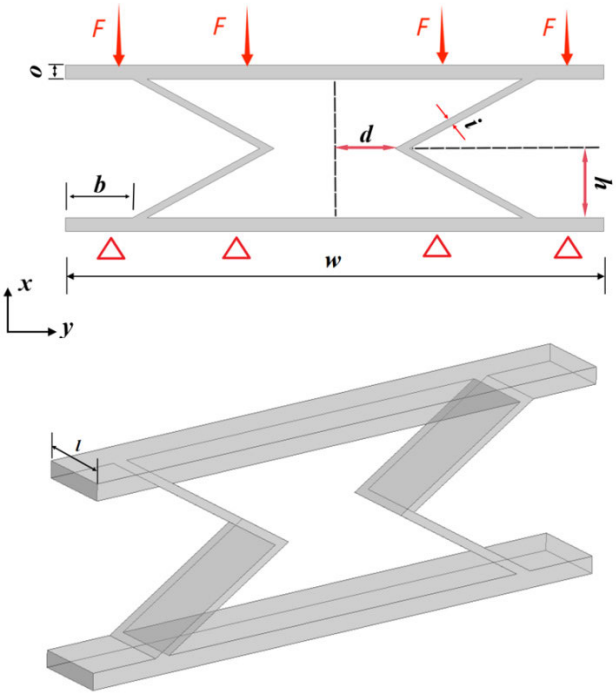


FIGURE 8. Geometric model of a unit cell of zero Poisson ratio structure.

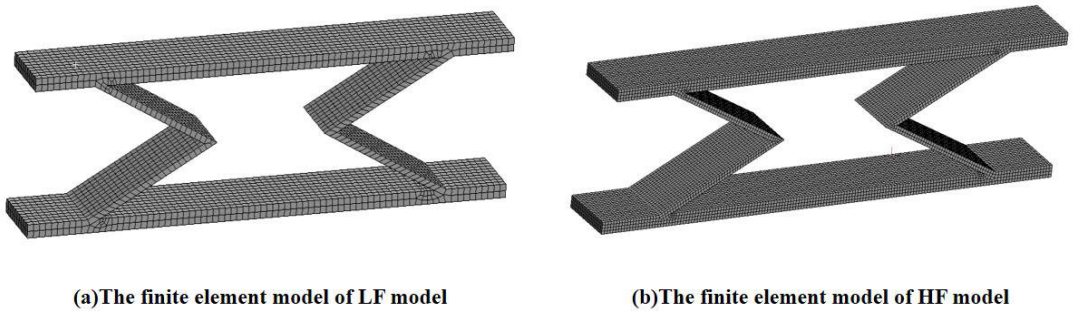
By using the simulation results to create a dataset and train multiple classification models, the prediction indicators of the multiple classification models were obtained, as shown in Table 9. From Table 9, it can be seen that in the prediction of the failure boundary of the zero-Poisson-ratio structure, EMFSVC's prediction accuracy of EMFSVC is significantly better than that of the other three classification models. By comparing the MFGPC and GPC models, the cost of building an MFGPC model increased by nearly five times. The classification indicators of the MFGPC model {P=0.946, R=0.921, F1=0.927} were all better than those of the GPC model {P=0.912, R=0.823, F1=0.842}. A similar conclusion can be drawn by comparing MFSVC and SVC models. The modeling cost is about 144.49s of a MFSVC model, which is 46.27s for an SVC model. In terms of predictive performance, the MFSVC model performed better than the SVC model. By comparing the MFSVC model and the MFGPC model, it can be found that the MFSVC modelling cost of MFSVC is 141.49s lower than that of the MFGPC model 242.20s. Meanwhile, the various classification indicators of the MFSVC model are similar to those of the MFGPC model. Furthermore, by comparing the proposed method and the MFSVC model, it can be observed that EMFSVC's cost is 301.86s, which is approximately 150 s higher than that of the MFSVC model's cost 141.49s. However, the cost was still within the acceptable range. However, EMFSVC performed better than MFSVC in terms of prediction accuracy, as the classification indicators {P=0.982, R=0.989, F1=0.980} of EMFSVC were improved by nearly 5% compared to the classification indicators of the MFSVC model {P=0.932, R=0.930, F1=0.928}. Based on the above comparisons, it can be concluded that EMFSVC has the best predictive performance for the failure boundary of the zero Poisson ratio structure and can effectively predict the failure boundaries of the problem.

TABLE 8. Design parameters of a unit cell of zero Poisson ratio structure.

Design variables	Range(values)
The distance d	3-15 mm
The thickness of the inner plate i	0.5-2 mm
The thickness of the outer plate o	2i mm
The length of the unit cell l	10 mm
The width of the unit cell w	80 mm
The distance b	10 mm
The half-height h	10 mm
Elastic modulus E	160 GPa
Poisson's ratio μ	0.2884
The applied pressure F	1 N/mm^2

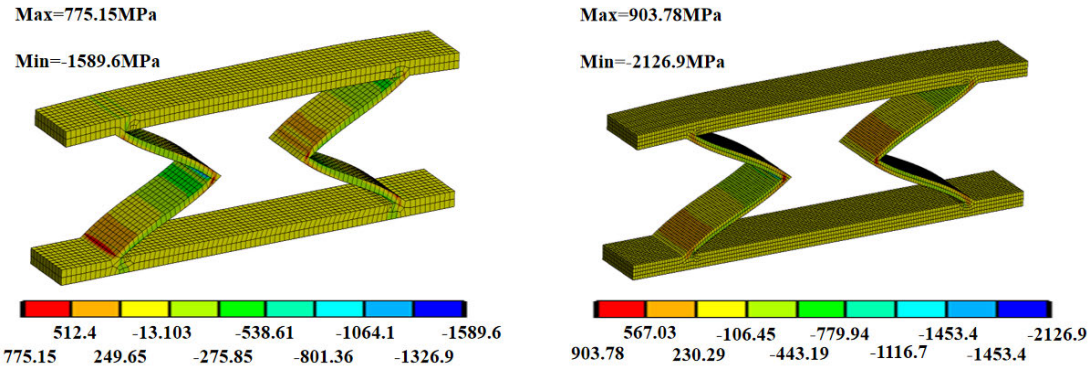
TABLE 9. The results of each model for zero Poisson ratio structure.

Model	P	R	F1	T(s)
GPC	0.731	0.855	0.788	94.200
MFGPC	0.731	0.855	0.788	793.304
SVC	0.894	0.887	0.890	**46.272**
MFSVC	0.932	0.930	0.928	144.440
EMFSVC	**0.982**	**0.989**	**0.980**	301.187



(a)The finite element model of LF model (b)The finite element model of HF model

FIGURE 9. Mesh models of a unit cell of zero Poisson ratio structure.



(a)Simulation resulit of LF model (b)Simulation resulit of HF model

FIGURE 10. Simulation result of a unit cell of zero Poisson ratio structure.

To visually understand the failure boundaries of the zero Poisson ratio structures, 10,000 high-fidelity sample points were used to approximate the true failure boundary of the zero-Poisson-ratio structure, as shown in Fig. 11. It can be noted that the area of the safe region is significantly larger than that of the failure region. The predicted boundaries of the five models are shown in Fig. 12. In MFGPC and GPC, the models were unable to effectively predict the failure boundaries, with the predicted results $\{y\}$ closely approaching 1. However, in MFSVC and SVC, MFSVC is closer to the actual boundary positions, whereas SVC is more in line with the real boundary trend. It can be observed that EMFSVC, when compared with the other models, predicts boundaries that are much closer to the true boundary, indicating a better fit. These results demonstrate that EMFSVC has good accuracy and feasibility in predicting the failure boundaries of static structures, and can compensate for the problem of decreased predictive performance in classification models caused by imbalanced real data.

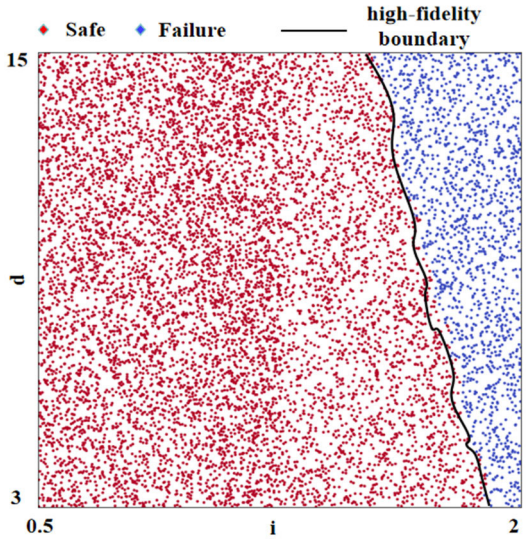


FIGURE 11. Simulation dataset of zero Poisson ratio structure.

B. THE SAFE OPERATING BOUNDARY OF AN AXIAL FLOW COMPRESSOR ROTOR

To validate the application feasibility of the proposed model in complex fluid machinery, the surge and choke boundaries of the compressor rotor Rotor37 [1] were predicted to compare the accuracy of each model. An axial flow compressor is a core component of an aeroengine. Its safe operation is

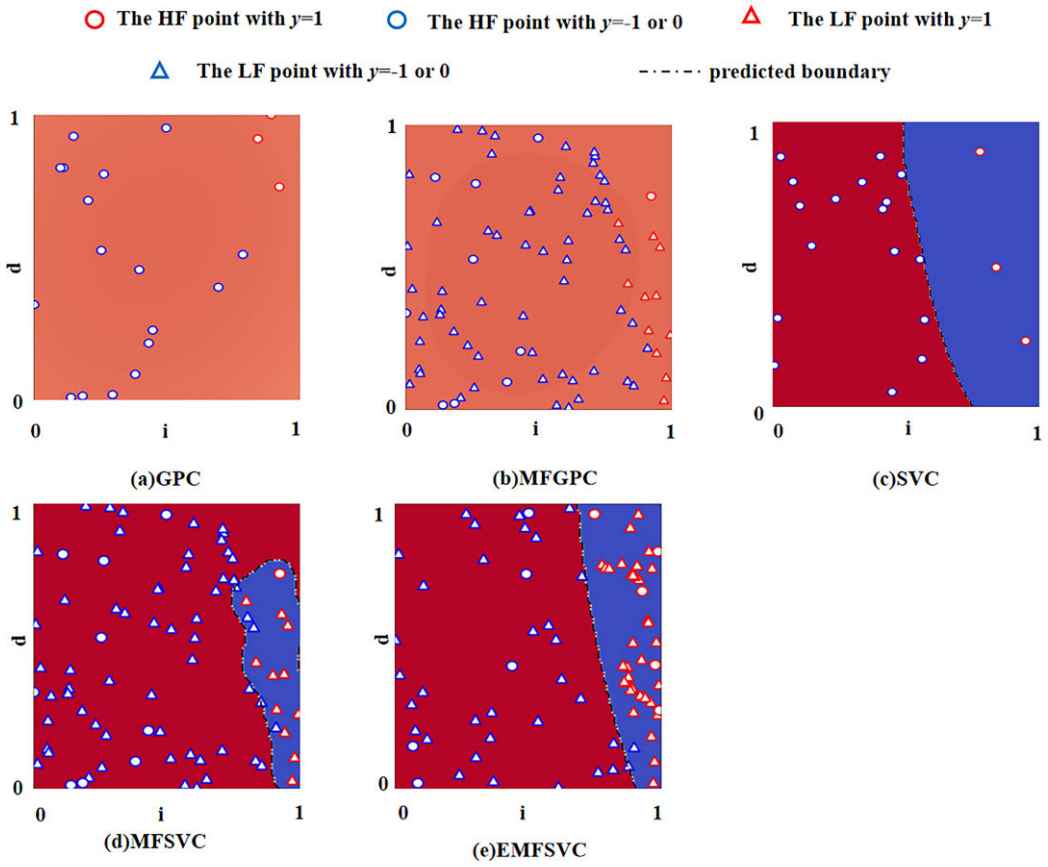


FIGURE 12. Failure boundary of zero Poisson ratio structure.

crucial for the safety of the entire engine. If the compressor in the engine experiences surging or blockage, it not only affects the performance of the engine but also leads to damage and reduces the engine's lifespan. Hence, it is essential to predict the surging and blockage boundaries of compressors. In this experiment, the NUMECA software was used to analyze the performance curve of Rotor37 at different speeds, and a dataset was established based on the speed and pressure ratio. The design parameters of the Rotor37 compressor are listed in Table 10. Fig. 14 illustrates the 3-D view. The different speeds of Rotor37 as a percentage of the initial speed were set to [100%, 90%, 80%, 70%, 60%]. A fine mesh model consisting of 843093 cells was considered as the HF model. Therefore, a coarse mesh model consisting of 364017 elements was used as the LF model. Fig. 15 presents the mesh models of the HF and LF models. The absolute pressure on the solid surface of the compressor rotor is presented in Fig. 16, where the corresponding simulation times of the HF and LF models were 9 and 33 min, respectively. Therefore, the cost ratio is 3.66, which is rounded to 4. For the multi-fidelity models, a dataset consisting of 10 HF samples and 40 LF samples was used. For a fair comparison, the single-fidelity models used a sample set with 20 HF samples. The proportion of minority classes in the high-fidelity dataset is about 0.308, whereas in the low-fidelity dataset, it is about 0.260. To test the accuracy of the model, 200 HF sample points were used as the testing sets.

TABLE 10. Design parameters of Rotor37 compressor.

Design variables	values
number of blades	36
inlet hub-to-tip diameter ratio	0.7
blade aspect ratio	1.19
tip relative inlet Mach number	1.48
hub relative inlet Mach number	1.13
tip solidity	1.29
the initial rotational speed (RPM) (r/min)	17188
total pressure ratio	2.106
polytropic efficiency (%)	88.9

Similar to the method used to verify the effectiveness of EMFSVC on zero-Poisson-ratio structures, the Rotor37 dataset was used to train various models to obtain their respective classification indicators, as shown in Table 13.

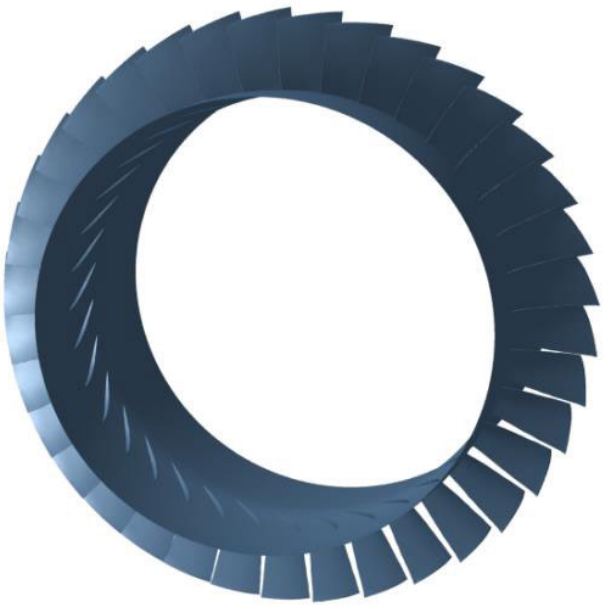


FIGURE 13. Geometric model of rotor37 compressor.

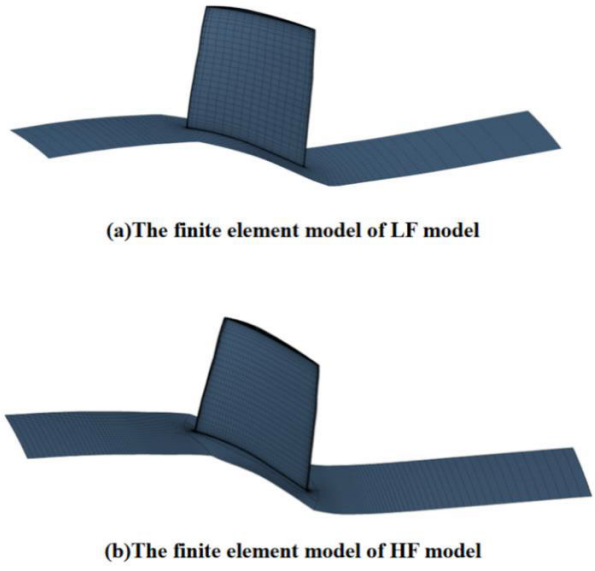


FIGURE 14. Mesh models of the Rotor37 compressor.

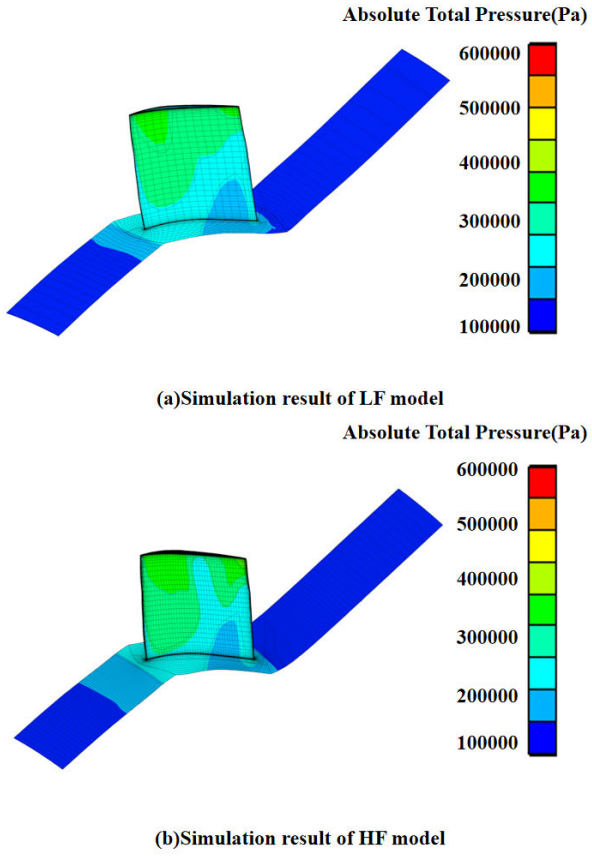


FIGURE 15. Simulation result of the Rotor37 compressor.

From Table 13, it can be seen that in the surge and choke boundary prediction dataset of the Rotor37 compressor, EMFSVC's prediction accuracy of the EMFSVC is significantly better than that of the other three classification models. Comparing the GPC with the MFGPC, the cost of the MFGPC models increased by nearly 10 times compared to the GPC, and the classification indicators of the MFGPC model {P=0.504, R=0.710, F1=0.590} were the same as those of the GPC model, indicating that the GPC and MFGPC models in this experiment had poor predictive performance on this dataset. Comparing the MFSVC model to the SVC model, the classification measurements {P=0.853 R=0.822, F1=0.796} of the MFSVC model were better than the classification measurements {P=0.849 R=0.808, F1=0.774} of the SVC model, although the cost of the MFSVC model increased by nearly twice that of the SVC model. By comparing the MFSVC model and the MFGPC model, it can be found that while the MFSVC model has a cost of {T=119.50} much lower than the MFGPC model's cost of {T=311.26}, the various classification indicators of the MFSVC model are improved by about 0.1 to 0.2 compared to the various classification indicators of the MFGPC model. By comparing the EMFSVC with the MFSVC model, it can be seen that EMFSVC's cost of EMFSVC {T=318.27} increases by approximately 198 s compared to that of the MFSVC model {T=119.50}, but the cost is still within an acceptable range. However, the various classification indicators of EMFSVC {P=0.913, R=0.910, F1=0.908} were improved by 7–14% compared with the various classification indicators of the MFSVC model {P=0.853, R=0.822, F1=0.796}. Based on the above comparisons, it can be concluded that EMFSVC has the best predictive performance for the surge and choke boundary problem of the Rotor37 compressor and can effectively predict the failure boundaries of the problem.

Similar to the previous description, to visually observe the surge and choke boundaries of the Rotor37 compressor

TABLE 11. The results of each model for the safe operating boundary of axial flow compressor.

Model	P	R	F1	T(s)
GPC	0.504	0.710	0.590	**39.703**
MFGPC	0.504	0.710	0.590	311.262
SVC	0.849	0.808	0.774	67.566
MFSVC	0.853	0.822	0.796	119.503
EMFSVC	**0.913**	**0.910**	**0.908**	318.266

and the training effect of each model, a graphical representation of the relationship between the rotational speed and pressure ratio of the compressor was obtained, and Fig. 17 was obtained. In Fig. 17, the distribution of the failure and safe samples can be observed, indicating that the surge and choke problem of Rotor37 is mainly due to the difficulty in obtaining data samples. Moreover, the MFSVC and proposed models were trained using the Rotor37 dataset, and the surge and choke prediction boundaries were obtained, as shown in Fig. 18. From Fig. 18, It can be seen that the predicted surge and choke boundaries of the Rotor37 compressor by the EMFSVC are closer to the real boundaries of the Rotor37 compressor. In the case of MFSVC and SVC, MFSVC provides a better representation of the appearance of compressor surging and choking boundaries, whereas SVC shows an overlap phenomenon in the predicted boundaries. As for MFGPC and GPC, owing to the imbalance in the data, the predicted probabilities of the models were all close to one, leading to a less satisfactory prediction performance. The experimental results obtained from predicting the surge and choke boundaries of Rotor37 indicate that the EMFSVC exhibits high accuracy and feasibility when used to predict failure boundaries in fluid dynamics simulations. This result is significant because it addresses the issue of decreased predictive performance of classification models caused by imbalanced datasets, which arise from the challenge of acquiring actual data.

C. THE SIMULATION FAILURE BOUNDARY OF MODELING THE ISENTROPIC EFFICIENCY OF Rotor37

To verify the effectiveness of this method in higher-dimensional engineering problems, it was applied to predict the simulation failure boundary of the isentropic efficiency modeling of the Rotor37. Simulation failure often results from an ill-geometry or mesh, unstable or weak convergence of the solver, etc. in computational fluid dynamics (CFD) simulations [47]. With the advancement of computational fluid dynamics (CFD) simulations and computers, optimization design combining intelligent optimization algorithms and CFD simulations has become popular in engineering fields. However, owing to the existence of the simulation failure, the iterative optimal search driven by the optimization algorithm, such as Bayesian optimization, is often halted

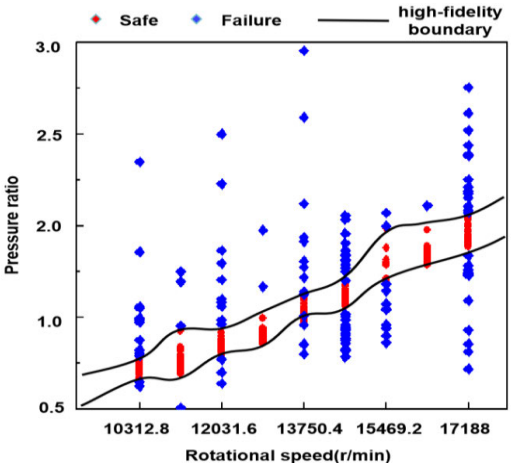


FIGURE 16. Simulation dataset of zero Poisson ratio structure.

prematurely because of the missing responses of the infill sample determined by the expected improvement criteria. Therefore, determining the boundary of the simulation failure region can be helpful for engineering optimization design.

The main design specifications and 3D view of Rotor37 are provided in the previous section and will not be introduced here. The problem mainly involves the prediction of the simulation failure boundary when simulating the isentropic efficiency of a rotor under geometric deformation. The formula for calculating the isentropic efficiency is:

$$f = \eta_c(\boldsymbol{x}) \tag{30}$$

with

$$\eta_c = \frac{h_{2s} - h_1}{h_{2r} - h_1} \tag{31}$$

where h_1 denotes the specific enthalpy of air at the inlet of the rotor. The parameters h_{2s} and h_{2r} represent the specific enthalpy of the gas at the outlet of the rotor during both isentropic and actual compression processes. The parameter $\boldsymbol{x}$ determines the shape of the blades. The parameters h_1, h_{2s}, and h_{2r} were mainly obtained from the results of CFD simulations results, and the NUMECA software was mainly used in this simulation. In this problem, the geometry definition of the Rotor37 blade is composed of three blade sections and the superposition principle, where each section is formed by adding the thickness of the inlet side and pressure side to the arc. Each section required nine parameters to determine the profile shape, as shown in Fig. 18(a). The parameters β_1 and β_2 in the figure represent the inlet and outlet blade angles, and the parameters α_{tw} and γ_{ca} represent the trailing-edge angle and sweep angle. t_{p1}, t_{p2}, t_{s1}, t_{s2}, and t_{s3} are the control points for the pressure-side and suction-side thickness distributions. In Fig. 18(b), the lines passing through the centroid of each section are the stacking lines, which allow the profiles of the middle section and the blade tip to move in the axial and circumferential directions, known as stacking line sweep and lean. Therefore, the blade shape was determined using the 31 parameters.

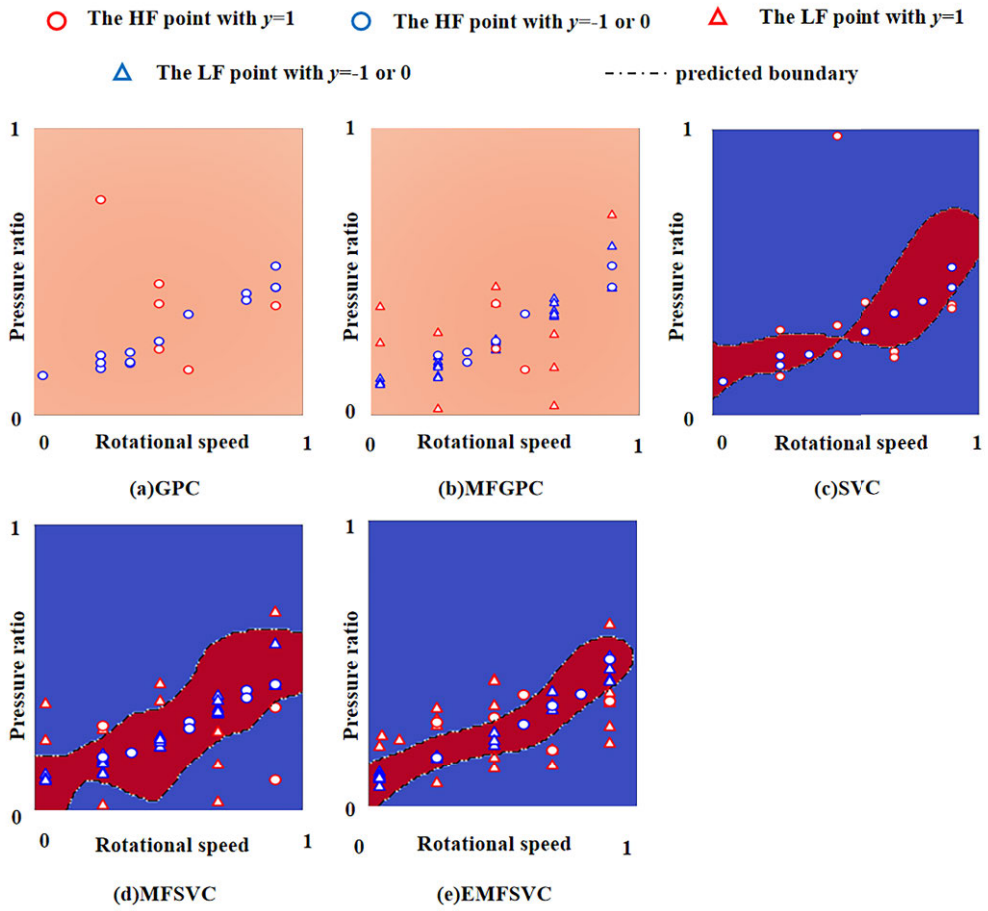


FIGURE 17. The surge and choke boundary of the Rotor37 compressor.

To generate a multi-fidelity dataset, grid resolution was used to distinguish between the HF and LF simulations, with the grid division numbers for the LF and HF simulation models being 312,077 and 799,185, respectively. In the isentropic efficiency simulation process, the fluid computational simulation was used to solve the Reynolds-averaged Navier-Stokes equations, and the Spalart-Allmaras turbulence model was utilized to resolve the turbulent flow. The LF and HF simulations were completed in about 5 and 12 min, respectively, with a cost ratio of approximately 2. Therefore, a single-fidelity dataset consisting of 400 HF sample points was used. Correspondingly, the multi-fidelity dataset consisted of 400 LF sample points and 200 HF sample points. The proportion of minority classes in the high-fidelity dataset is about 0.180, whereas in the low-fidelity dataset, it is about 0.260. To test the accuracy of the model, 10000 HF sample points were used as the testing sets.

The respective classification metrics obtained by applying the multi-fidelity and single-fidelity datasets to each classification model are listed in Table 14. From the table, it can be seen that in GPC and MFGPC, the predictive boundary performance of the MFGPC is better than that of the GPC. The classification metric {F1=0.744} of MFGPC is nearly 30% higher than that of the classification metric {F1=0.568}, but correspondingly, the cost of the model also increases. In SVC and MFSVC, the predictive boundary performance of MFSVC was also superior to that of SVC. In MFGPC and MFSVC, although the cost {T=2011.74} of MFSVC is greater than the cost {T=1286.58} of SVC, it is much smaller than the cost {T=37089.36} of MFGPC, and the predictive performance {P=0.852, R=0.827, F1=0.752} of MFSVC is better than the predictive performance {P=0.678, R=0.824, F1=0.744} of MFGPC. In comparing EMFSVC and MFSVC, although the cost {T=17262.75} of EMFSVC is higher than the cost {T=2011.74} of MFSVC, it is lower than the cost {T=37089.36} of MFGPC, and the classification metric {F1=0.881} of EMFSVC is nearly 17% higher than that of metric {F1=0.752}. From the above comparison, it can be concluded that in the simulation failure boundary of modeling the isentropic efficiency problem, the predictive performance of EMFSVC is superior to that of other models, and the cost of the model is lower than that of MFGPC. Owing to the 31-dimension of the design space, the approximated boundary from the 10000 HF simulations and the

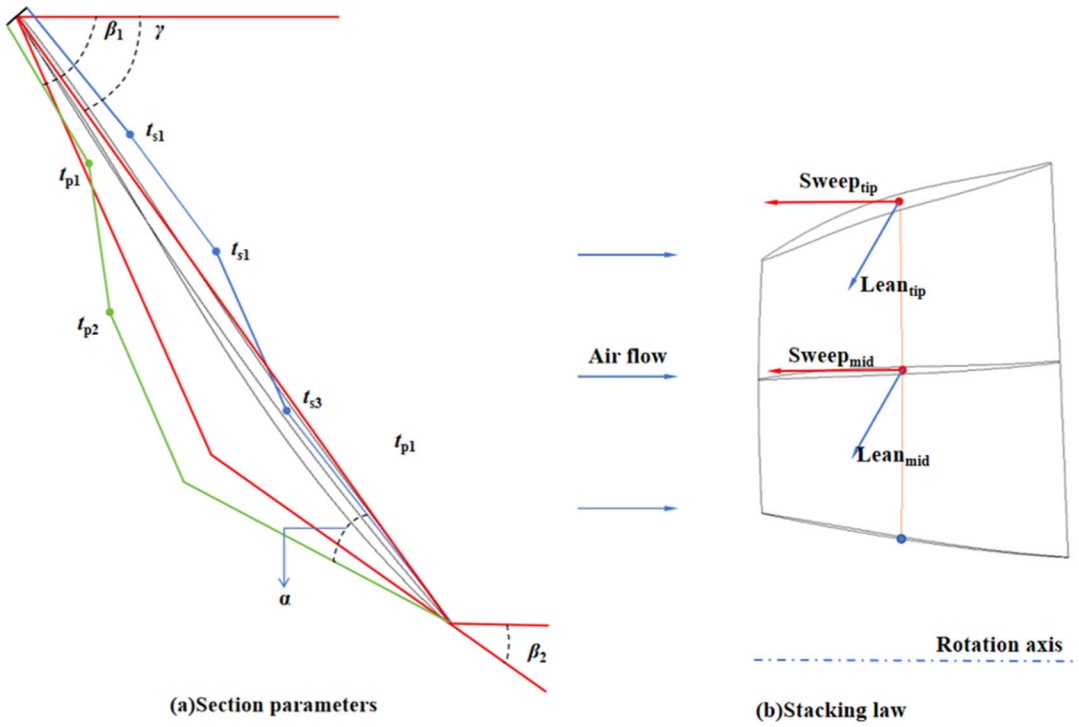


FIGURE 18. Geometric parameters for the parametric representation of the blade.

predicted boundaries of the five compared methods were not illustrated.

TABLE 12. The results of each model for the simulation failure boundary of modeling the isentropic efficiency.

Model	P	R	F1	T(s)
GPC	0.678	0.824	0.744	9422.097
MFGPC	0.678	0.824	0.744	26754.977
SVC	0.856	0.830	0.759	**1286.577**
MFSVC	0.859	0.837	0.762	2011.740
EMFSVC	**0.937**	**0.935**	**0.930**	17262.754

In summary, the EMFSVC can construct more accurate multi-fidelity models for failure boundary prediction in engineering problems and demonstrate the effectiveness of EMFSVC in high-dimensional engineering problems.

V. CONCLUSION

In this paper, an MFSVC model based on the SVC model is proposed, and then the method is improved to deal with imbalanced data. The influence of the optimization objective function on the determination of the hyperparameters of the MFSVC/EMFSVC model was explored. To verify the effectiveness of the MFSVC model, it was compared with GPC, SVC, and MFGPC using several numerical examples with balanced data. The numerical results demonstrate the effectiveness of the MFSVC model for balanced data and its ability to achieve higher accuracy than the other models. In addition, to study the influence of the number of HF samples on the performance of the EMFSVC, each classification model was compared in several numerical examples with balanced data. The study revealed that increasing the number of HF samples after reaching a certain quantity did not significantly improve the accuracy of the EMFSVC. Therefore, selecting an appropriate number of high-accuracy samples can effectively reduce unnecessary costs while ensuring the predictive performance of the proposed model.

Furthermore, it was compared with MFSVC, MFGPC, GPC, and SVC in four imbalanced numerical examples and three practical engineering examples. In the imbalanced numerical examples, EMFSVC showed a better overall predictive performance than the other models. In the case of extremely imbalanced datasets, EMFSVC can improve the prediction accuracy by approximately 9% compared with MFSVC. In the three engineering examples, EMFSVC outperformed other boundary prediction models, especially in the prediction of the surge and choke boundaries of the compressor rotor, where the prediction performance was improved by nearly 14% compared to MFSVC. Although EMFSVC can improve classification performance in imbalanced data problems, its computational cost will increase compared to MFSVC, and may even be the same as MFGPC. Therefore, it is still worth exploring how to effectively reduce the computational cost of EMFSVC and ensure its predictive performance in imbalanced data problems. In the future, we plan to combine cost-sensitive learning or active learning

with the proposed method and validate its predictive performance under more challenging data scenarios.

REFERENCES

[1] R. L. Davis and J. Yao, "Prediction of compressor stage performance from choke through stall," *J. Propuls. Power*, vol. 22, no. 3, pp. 550–557, May 2006, doi: 10.2514/1.15463.

[2] C. Rodgers, "Typical performance characteristics of gas turbine radial compressors," *J. Eng. Power*, vol. 86, no. 2, pp. 161–170, Apr. 1964, doi: 10.1115/1.3677568.

[3] T. Arima, T. Sonoda, M. Shirotori, A. Tamura, and K. Kikuchi, "A numerical investigation of transonic axial compressor rotor flow using a low-Reynolds-number k-ε turbulence model," *J. Turbomachinery*, vol. 121, no. 1, pp. 44–58, Jan. 1999, doi: 10.1115/1.2841233.

[4] H. Khaleghi, "Stall inception and control in a transonic fan, part A: Rotating stall inception," *Aerosp. Sci. Technol.*, vol. 41, pp. 250–258, Feb. 2015, doi: 10.1016/j.ast.2014.12.004.

[5] U. Maulik and D. Chakraborty, "Remote sensing image classification: A survey of support-vector-machine-based advanced techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 33–52, Mar. 2017, doi: 10.1109/MGRS.2016.2641240.

[6] D. C. Toledo-Pérez, J. Rodríguez-Reséndiz, R. A. Gómez-Loenzo, and J. C. Jauregui-Correa, "Support vector machine-based EMG signal classification techniques: A review," *Appl. Sci.*, vol. 9, no. 20, p. 4402, Oct. 2019, doi: 10.3390/app9204402.

[7] Y. Li, T. Liu, and Y. Xie, "Thermal fluid fields reconstruction for nanofluids convection based on physics-informed deep learning," *Sci. Rep.*, vol. 12, no. 1, p. 12567, Jul. 2022, doi: 10.1038/s41598-022-16463-1.

[8] C. E. Rasmussen and C. K. I. Williams, "Gaussian processes for machine learning," in *Adaptive Computation and Machine Learning*. Cambridge, MA, USA: MIT Press, 2008.

[9] M. A. Bezerra, R. E. Santelli, E. P. Oliveira, L. S. Villar, and L. A. Escaleira, "Response surface methodology (RSM) as a tool for optimization in analytical chemistry," *Talanta*, vol. 76, no. 5, pp. 965–977, Sep. 2008, doi: 10.1016/j.talanta.2008.05.019.

[10] J. M. Moguerza and A. Muñoz, "Support vector machines with applications," *Stat. Sci.*, vol. 21, no. 3, pp. 322–336, Aug. 2006, doi: 10.1214/088342306000000493.

[11] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, Nov. 2018, Art. no. e00938, doi: 10.1016/j.heliyon.2018.e00938.

[12] S. J. Huang, N. G. Cai, and P. P. Pacheco, "Applications of support vector machine (SVM) learning in cancer genomics," *CGP*, vol. 15, no. 1, pp. 41–51, Jan. 2018, doi: 10.21873/cgp.20063.

[13] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[14] D. Liu, H. Qian, G. Dai, and Z. Zhang, "An iterative SVM approach to feature selection and classification in high-dimensional datasets," *Pattern Recognit.*, vol. 46, no. 9, pp. 2531–2537, Sep. 2013, doi: 10.1016/j.patcog.2013.02.007.

[15] S. Abe, "Fuzzy support vector machines for multilabel classification," *Pattern Recognit.*, vol. 48, no. 6, pp. 2110–2117, Jun. 2015, doi: 10.1016/j.patcog.2015.01.009.

[16] H. Yan, Q. Ye, T. Zhang, D.-J. Yu, X. Yuan, Y. Xu, and L. Fu, "Least squares twin bounded support vector machines based on L1-norm distance metric for classification," *Pattern Recognit.*, vol. 74, pp. 434–447, Feb. 2018, doi: 10.1016/j.patcog.2017.09.035.

[17] B. Pradhan and M. I. Sameen, "Manifestation of SVM-based rectified linear unit (ReLU) kernel function in landslide modelling," in *Proc. Space Sci. Commun. Sustainability*, W. Suparta, M. Abdullah, and M. Ismail, Eds., Singapore: Springer, Oct. 2017, pp. 185–195, doi: 10.1007/978-981-10-6574-3_16.

[18] J. Wu, "Efficient HIK SVM learning for image classification," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4442–4453, Oct. 2012, doi: 10.1109/TIP.2012.2207392.

[19] C. Park, R. T. Haftka, and N. H. Kim, "Remarks on multi-fidelity surrogates," *Structural Multidisciplinary Optim.*, vol. 55, no. 3, pp. 1029–1050, Mar. 2017, doi: 10.1007/s00158-016-1550-y.

[20] A. I. J. Forrester, A. Sóbester, and A. J. Keane, "Multi-fidelity optimization via surrogate modelling," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 463, no. 2088, pp. 3251–3269, Dec. 2007, doi: 10.1098/rspa.2007.1900.

[21] D. Liu and Y. Wang, "Multi-fidelity physics-constrained neural network and its application in materials modeling," *J. Mech. Design*, vol. 141, no. 12, Dec. 2019, Art. no. 121403, doi: 10.1115/1.4044400.

[22] Y. Song, Q. S. Cheng, and S. Koziel, "Multi-fidelity local surrogate model for computationally efficient microwave component design optimization," *Sensors*, vol. 19, no. 13, p. 3023, Jul. 2019, doi: 10.3390/s19133023.

[23] M. Shi, L. Lv, W. Sun, and X. Song, "A multi-fidelity surrogate model based on support vector regression," *Structural Multidisciplinary Optim.*, vol. 61, no. 6, pp. 2363–2375, Jun. 2020, doi: 10.1007/s00158-020-02522-6.

[24] X. Song, L. Lv, W. Sun, and J. Zhang, "A radial basis function-based multi-fidelity surrogate model: Exploring correlation between high-fidelity and low-fidelity models," *Structural Multidisciplinary Optim.*, vol. 60, no. 3, pp. 965–981, Sep. 2019, doi: 10.1007/s00158-019-02248-0.

[25] R. C. Aydin, F. A. Braeu, and C. J. Cyron, "General multi-fidelity framework for training artificial neural networks with computational models," *Frontiers Mater.*, vol. 6, p. 61, Apr. 2019, doi: 10.3389/fmats.2019.00061.

[26] F. Sahli Costabal, P. Perdikaris, E. Kuhl, and D. E. Hurtado, "Multi-fidelity classification using Gaussian processes: Accelerating the prediction of large-scale computational models," *Comput. Methods Appl. Mech. Eng.*, vol. 357, Dec. 2019, Art. no. 112602, doi: 10.1016/j.cma.2019.112602.

[27] A. Onan, "Consensus clustering-based undersampling approach to imbalanced learning," *Sci. Program.*, vol. 2019, pp. 1–14, Mar. 2019, doi: 10.1155/2019/5901087.

[28] L. Camacho, G. Douzas, and F. Bacao, "Geometric SMOTE for regression," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116387, doi: 10.1016/j.eswa.2021.116387.

[29] A. M. Radwan, "Enhancing prediction on imbalance data by thresholding technique with noise filtering," in *Proc. 8th Int. Conf. Inf. Technol. (ICIT)*. Amman, Jordan, May 2017, pp. 399–404, doi: 10.1109/ICITECH.2017.8080033.

[30] Y. Liu, X. Yu, J. X. Huang, and A. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Inf. Process. Manage.*, vol. 47, no. 4, pp. 617–631, Jul. 2011, doi: 10.1016/j.ipm.2010.11.007.

[31] C.-F. Tsai, W.-C. Lin, Y.-H. Hu, and G.-T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Inf. Sci.*, vol. 477, pp. 47–54, Mar. 2019, doi: 10.1016/j.ins.2018.10.029.

[32] X. Wang, J. Ren, H. Ren, W. Song, Y. Qiao, Y. Zhao, L. Linghu, Y. Cui, Z. Zhao, L. Chen, and L. Qiu, "Diabetes mellitus early warning and factor analysis using ensemble Bayesian networks with SMOTE-ENN and Boruta," *Sci. Rep.*, vol. 13, no. 1, p. 12718, Aug. 2023, doi: 10.1038/s41598-023-40036-5.

[33] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 101–121, doi: 10.1016/B978-0-12-815739-8.00006-7.

[34] Q. Wu and D.-X. Zhou, "SVM soft margin classifiers: Linear programming versus quadratic programming," *Neural Comput.*, vol. 17, no. 5, pp. 1160–1187, May 2005, doi: 10.1162/0899766053491896.

[35] K. M. Nakanishi, K. Fujii, and S. Todo, "Sequential minimal optimization for quantum-classical hybrid algorithms," *Phys. Rev. Res.*, vol. 2, no. 4, Oct. 2020, Art. no. 043158, doi: 10.1103/PhysRevResearch.2.043158. [Online]. Available: https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/

[36] M. Kennedy, "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, vol. 87, no. 1, pp. 1–13, Mar. 2000, doi: 10.1093/biomet/87.1.1.

[37] M. Gendreau and J.-Y. Potvin, *Handbook of Metaheuristics* (International Series in Operations Research & Management Science), vol. 146. Boston, MA, USA: Springer, 2010, doi: 10.1007/978-1-4419-1665-5.

[38] Z. Tao, L. Huiling, W. Wenwen, and Y. Xia, "GA-SVM based feature selection and parameter optimization in hospitalization expense modeling," *Appl. Soft Comput.*, vol. 75, pp. 323–332, Feb. 2019, doi: 10.1016/j.asoc.2018.11.001.

[39] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: Past, present, and future," *Multimedia Tools Appl.*, vol. 80, no. 5, pp. 8091–8126, Feb. 2021, doi: 10.1007/s11042-020-10139-6.

[40] D. Guan, W. Yuan, Y.-K. Lee, and S. Lee, "Nearest neighbor editing aided by unlabeled data," *Inf. Sci.*, vol. 179, no. 13, pp. 2273–2282, Jun. 2009, doi: 10.1016/j.ins.2009.02.011.

[41] I. Tomek, "A generalization of the k-NN rule," *IEEE Trans. Syst., Man, Cybern.*, vols. SMC–6, no. 2, pp. 121–126, Feb. 1976, doi: 10.1109/TSMC.1976.5409182.
[42] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.
[43] M. Pajany, S. Venkatraman, U. Sakthi, M. Sujatha, and M. K. Ishak, "Optimal fuzzy deep neural networks-based plant disease detection and classification on UAV-based remote sensed data," *IEEE Access*, vol. 12, pp. 162131–162144, 2024, doi: 10.1109/ACCESS.2024.3488751.
[44] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique," *IEEE Access*, vol. 9, pp. 71194–71209, 2021, doi: 10.1109/ACCESS.2021.3079204.
[45] J. Liu, J. Yi, Q. Zhou, and Y. Cheng, "A sequential multi-fidelity surrogate model-assisted contour prediction method for engineering problems with expensive simulations," *Eng. Comput.*, vol. 38, no. 1, pp. 31–49, Feb. 2022, doi: 10.1007/s00366-020-01043-6.
[46] C. Ling and Z. Lu, "Support vector machine-based importance sampling for rare event estimation," *Structural Multidisciplinary Optim.*, vol. 63, no. 4, pp. 1609–1631, Apr. 2021, doi: 10.1007/s00158-020-02809-8.
[47] Y. He and J. Luo, "An efficient hierarchical Kriging modeling method for high-dimension multi-fidelity problems," 2022, *arXiv:2301.00216*.

JINLIANG LUO was born in February 1968. He received the bachelor's degree from Henan University of Science and Technology, the master's degree in power machinery and engineering from Wuhan University of Technology, and the D.Eng. degree in mechanical design and theory from Chongqing University. He is currently a Master's Supervisor and the Head of the Department of Energy and Power, University of South China. He has led several national or provincial research projects and published his research in international academic journals. His research interests include robotics, vehicle engineering, machinery design, and intelligent machinery.

LINGZHI LIU was born in August 1999. He received the bachelor's degree from Hunan University of Science and Technology. He is currently pursuing the master's degree with the University of South China. His main research interests include fluid simulation, Bayesian optimization, and classification algorithms.

YOUWEI HE was born in August 1992. He received the B.Eng. and Ph.D. degrees from the School of Energy and Power Engineering, Xi'an Jiaotong University, in 2014 and 2021, respectively. He was under joint training with the Department of Information Engineering, Ghent University, Belgium, from December 2019 to November 2020. He has taken part in several national and corporate projects. He has published papers in international journals and conferences. His research interests include Bayesian optimization algorithms, impeller mechanical design and optimization, lightweight chillers, and fluid simulation.

KUAN TAN was born July 2000. He received the bachelor's degree from Changsha University of Science and Technology. He is currently pursuing the master's degree with the University of South China. His primary research interests include data mining, fluid simulation, and impeller optimization.