OXFORD CDE

Journal of Computational Design and Engineering, 2023, **10**, 1547–1560

DOI: 10.1093/jcde/qwad071

Advance access publication date: 11 July 2023

Research Article

Adaptive neural network ensemble using prediction frequency

Ungki Lee[1] and Namwoo Kang[2,*]

[1]Ground Technology Research Institute, Agency for Defense Development, Bugyuseong-daero 488-160, Yuseong-gu, Daejeon 34060, Republic of Korea
[2]Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology, Daejeon 34051, Republic of Korea
*Correspondence: nwkang@kaist.ac.kr

Abstract

Neural network (NN) ensembles can reduce large prediction variance of NN and improve prediction accuracy. For highly non-linear problems with insufficient data set, the prediction accuracy of NN models becomes unstable, resulting in a decrease in the accuracy of ensembles. Therefore, this study proposes a prediction frequency-based ensemble that identifies core prediction values, which are core prediction members to be used in the ensemble and are expected to be concentrated near the true response. The prediction frequency-based ensemble classifies core prediction values supported by multiple NN models by conducting statistical analysis with a frequency distribution, which is a collection of prediction values obtained from various NN models for a given prediction point. The prediction frequency-based ensemble searches for a range of prediction values that contains prediction values above a certain frequency, and thus the predictive performance can be improved by excluding prediction values with low accuracy and coping with the uncertainty of the most frequent value. An adaptive sampling strategy that sequentially adds samples based on the core prediction variance calculated as the variance of the core prediction values is proposed to improve the predictive performance of the prediction frequency-based ensemble efficiently. Results of various case studies show that the prediction accuracy of the prediction frequency-based ensemble is higher than that of Kriging and other existing ensemble methods. In addition, the proposed adaptive sampling strategy effectively improves the predictive performance of the prediction frequency-based ensemble compared with the previously developed space-filling and prediction variance-based strategies.

Keywords: surrogate modelling, neural network, neural network ensemble, prediction, adaptive sampling

1. Introduction

A surrogate model is built to represent the true model with data obtained from a limited number of simulations or experiments; thus, predictions for computationally expensive models can be approximated (Kang *et al.*, 2019). Surrogate modelling focuses on input–output behaviour to find a model that approximates the relationship between input and output as accurately as possible. Neural network (NN) is one of the promising surrogate models and is known to be a universal function approximator with a good prediction accuracy (Pan *et al.*, 2014). Once a given data set is split into training and validation data set, the training data set is used to train the NN model and the validation data set is used to estimate the accuracy of the trained model. Applying NN as a surrogate model in various engineering domains, such as mechanical engineering (Chen *et al.*, 2020; Lin *et al.*, 2020; Ktari *et al.*, 2021; Schrader & Schauer, 2021), structural engineering (Gomes *et al.*, 2011; de Santana Gomes, 2019; Freitag et al., 2020; Yılmaz *et al.*, 2021), aerospace engineering (Bouhlel *et al.*, 2020; Du *et al.*, 2021; Zhang et al., 2021), biomedical engineering (Lu *et al.*, 2013; Eskinazi & Fregly, 2015), chemical engineering (Eason & Cremaschi, 2014; Moreno-Pérez *et al.*, 2018), civil engineering (García-Segura *et al.*, 2017; Shaw *et al.*, 2017; Thrampoulidis *et al.*, 2021), and composite structures (Papadopoulos *et al.*, 2018; Yan *et al.*, 2020), has been recently attempted.

However, NN has a problem in that different models are created in accordance with the selection of training data sets, initial parameters, and training algorithms (Zhang, 2007). This condition results in various prediction values; thus, NNs are referred to have high prediction variance. Ensembles that combine prediction values obtained from multiple component models have been developed to reduce the prediction variance of NNs (Sollich & Krogh, 1996). The basic assumption of ensembles is that the errors of a single model are compensated by the other models (Sagi & Rokach, 2018). Ensembles are generally constructed through two steps: training multiple component models and combining prediction values derived from the multiple component models. Sufficiently diverse component models with high predictive performance are required for the ensemble to have high accuracy (Deng *et al.*, 2013). Various component models can be created depending on the combination of training and validation data sets and the selection of initial parameters. If the validation error of the model is small, then the predictive performance can be estimated to be accurate. The accuracy of the ensemble is highly dependent on the combination of prediction values. Thus, various methods, such as average ensemble (Opitz & Shavlik, 1996), weighted ensemble (Bishop, 1995), and mode ensemble (Kourentzes et al., 2014), which combine prediction values, have been developed. Many studies have shown that the prediction accuracy of ensembles combining

Received: March 13, 2023. **Revised:** June 26, 2023. **Accepted:** June 26, 2023

multiple component models is higher than that of a single component model (Wolpert, 1992; Goodfellow *et al.*, 2016).

In the case of real engineering applications, the size of the data set for creating surrogate models is not usually large enough because simulations or experiments (e.g., finite element analysis and collision test) are time consuming and expensive (Gaspar *et al.*, 2017; Lee *et al.*, 2020). For highly non-linear problems, the prediction accuracy of NN decreases, and the situation worsens when the size of the data set is small. The deviation of prediction values increases as the prediction accuracy of component models becomes unstable, resulting in a decrease in the accuracy of ensembles. The average ensemble, which derives the final prediction by averaging predictions of component models, can be vulnerable to outliers. The weighted ensemble assigns different weights according to the training and validation errors of the component model. When inappropriate weights are assigned due to biased training and validation data set, the prediction accuracy decreases. The mode ensemble that can cope well with outliers is suitable when the deviation of prediction values is large. In the mode ensemble, kernel density estimation (KDE) is performed on the prediction values, and the value corresponding to the maximum density is identified as a mode and used as the final prediction value (Kourentzes et al., 2014). The mode is determined on the basis of the most frequent value; thus, the mode ensemble is insensitive to outliers compared with average and weighted ensembles. However, the accuracy of the prediction can be decreased in the presence of multiple high-frequency values. In addition to the most frequent value, keeping the possibility open to various prediction values is also requiredto resolve the problem of unstable prediction accuracy.

Therefore, this study proposes a prediction frequency-based ensemble that explores the range of core prediction values, which are core prediction members to be used in the ensemble and are expected to be concentrated near the true response. For a given prediction point, various prediction values obtained from the component models constitute a frequency distribution, and the prediction frequency-based ensemble performs statistical analysis with the frequency distribution to identify the core prediction values. To this end, a minimum frequency criterion that the frequency of core prediction values should satisfy needs to be determined. Prediction values belonging to the range supported by multiple prediction values can be classified into core prediction values, and the average of the core prediction values is determined as the final prediction value. Since the proposed ensemble considers the prediction values around the mode as well as the mode itself, the decrease in the prediction accuracy caused by the uncertainty of the mode can be alleviated. An adaptive sampling strategy, which efficiently improves the accuracy of the prediction frequency-based ensemble according to the variance of core prediction values, is also proposed. The proposed strategy, that is, the core prediction variance-based strategy, quantifies the uncertainty of next sample points by excluding outliers to accurately derive the actual uncertainty, and samples are added to the points with high uncertainty. Various examples are used to verify the proposed prediction frequency-based ensemble and adaptive sampling strategy.

The rest of this paper is organized as follows. Section 2 reviews previous works related to NNs, ensembles, and adaptive sampling strategies. Section 3 presents the proposed prediction frequency-based ensemble and the adaptive sampling strategy, which efficiently improves the predictive performance of the proposed prediction frequency-based ensemble. Section 4 demonstrates the predictive accuracy of ensemble methods and compares the proposed adaptive sampling strategy with other existing sampling strategies through case studies. Section 5 finally provides the conclusion and future directions.

2. Related Works

2.1. Application of NNs

NNs inspired by the human brain mimic the behaviour of biological neurons and have been successfully used in clustering, classification, pattern recognition, and regression (Abiodun *et al.*, 2018). The advantages of NNs are high accuracy and a fast process using parallel computing (Izeboudjen *et al.*, 2014). NNs can be used as surrogate models for analysis and prediction involving time-consuming simulations or expensive experiments. Eskinazi & Fregly (2015) proposed a novel surrogate contact modelling method based on NN to replace deformable joint contact models involving computationally expensive contact evaluations. The proposed method solves the computational bottleneck of musculoskeletal simulations or optimizations that require deformable joint contact models. Kumar *et al.* (2018) adopted NN to predict the wear of Al6061 alloy reinforced with aluminum oxide particulates. Wear height reduction according to applied load, sliding distance, and weight percentage of reinforcement is approximated through NN, and the results of experiments and the NN model predictions are in good agreement. García-Alba *et al.* (2019) used NN to reduce the computational cost required for the analysis of fecal indicator organism concentrations in estuaries. Maleki *et al.* (2019) predicted the spatial–temporal profile of pollutant concentrations and air quality indexes using NN. Liu *et al.* (2021) implemented NN to model chloride diffusion coefficient associated with chloride ingression, which is a major cause of corrosion of reinforced concrete structures. This study verifies that the predicted diffusion coefficient results are well matched with the experimental results. However, when a single NN model is used for surrogate modelling as in the studies above, different prediction values can be obtained depending on how the parameters are trained, making the prediction unstable. In addition, the prediction accuracy decreases compared to using ensembles that integrate multiple component models.

Design optimization entails numerous function evaluations or simulations in objective function evaluation and sensitivity analysis, respectively; thus, many studies have adopted NN as a surrogate model. Shaw *et al.* (2017) maximized the hydropower generation of the dam using a high-fidelity hydrodynamic and water quality model approximated by NN. Bouhlel *et al.* (2020) proposed gradient-enhanced NNs, which improve the accuracy of the NN model with gradient information, to reduce the computational time required for analysis and design optimization of airfoil shape design. The proposed method yields similar results to optimization designs based on high-fidelity computational fluid dynamics. In heat source layout optimization problem, Chen *et al.* (2020) evaluated the thermal performance according to the input layout by mapping the layout and the corresponding temperature field using a feature pyramid network, which is a kind of NN. Lin *et al.* (2020) optimized the design of the automotive semi-active suspension system by employing radial basis function NN, which approximates the regulating mechanism of the hydraulic adjustable damper. Yılmaz *et al.* (2021) used NN to the continuous motorcycle protection barrier design to reduce weight and cost. Zhu *et al.* (2022) approximated thermoelectric module simulation employing NN to optimize a thermoelectric generator design to improve the power generation performance. In such optimization prob-

lems, the quality of the optimum design is significantly affected by the accuracy of the surrogate model. However, using a single NN model can lead to unstable and sub-optimal designs.

In the field of reliability-based design optimization (RBDO), which derives a highly reliable design considering the uncertainties of engineering systems, numerous function evaluations or simulations are required for the reliability calculation; thus, NN-based surrogate models have been frequently applied (Elhewy *et al.*, 2006). For reliability analysis of structures, Cheng & Li (2008) developed an artificial NN-based genetic algorithm (ANN-GA), which approximates the limit state function with an ANN model and calculates the failure probability with the GA. Papadopoulos *et al.* (2012) proposed a reliability analysis method that combines NN and subset simulation (SS). In the proposed method, NNs are trained on multiple subdomains generated from SS and used as surrogate models to increase the efficiency of SS. Lehký *et al.* (2018) presented an approach to solve RBDO based on the inverse reliability method using NN. Nezhad *et al.* (2019) introduced a group method of data handling-type NNs with general structure for the reliability analysis of structures. Lee and Lee (2022) proposed a new inverse reliability analysis method combining NN with Monte Carlo simulation. In the proposed method, NN is used to train the relationship between the small number of samples and the corresponding true percentile value. Ren *et al.* (2022) presented a local best surrogate (LBS) approach and a local weighted average surrogate (LWAS) approach that combine NNs and Kriging for structural reliability analysis. The LBS approach selects the surrogate model with minimum predicted error, and the LWAS approach assigns weights to the surrogate models according to the prediction error. NNs and Kriging are used in both approaches as candidate surrogate models. However, these studies have limitations in that a single NN model, which can cause instability of reliability analysis, is applied to RBDO, resulting in designs that do not satisfy probabilistic constraints. The aforementioned problems caused by using a single NN model for surrogate modelling can be alleviated through ensembles that integrate multiple NN models to reduce the prediction variance.

2.2. Ensembles

Ensembles that combine prediction values derived from different NN models have been widely studied to improve the prediction accuracy and robustness of NNs, and using ensembles for prediction using NNs is reasonable (Crone *et al.*, 2011). Ensembles combine multiple individual predictions in a way such as averaging or voting to show better predictive performance than any individual models (Ganaie *et al.*, 2022). The average ensemble is the simplest way to combine prediction values and derives a final prediction by averaging predictions of component models (James *et al.*, 2013). The average ensemble can be defined as

$$\tilde{y}^{Average} = \frac{1}{m}\sum_{i=1}^{m} y_i, \tag{1}$$

where $\tilde{y}^{Average}$ is the average ensemble value, and y_i indicates the prediction value of the ith component model. The average ensemble reduces the risk of overfitting and improves the predictive performance compared with a single model. However, the predictive performance of the average ensemble can deteriorate due to weak component models and skewed distributions (Ju *et al.*, 2018). The majority ensemble counts the votes of component models and derives the majority of votes as the final prediction (Kim & Upneja, 2021). Li *et al.* (2017) proposed ensemble semi-supervised deep acoustic models by integrating the majority voting ensemble with semi-supervised deep learning. Xia *et al.* (2021) proposed a weighted stacked ensemble approach that uses both label correlations and process of learning weights. For multi-label classification, the proposed approach exploits sparsity regularization to facilitate classifier selection and construction of ensemble members. Majority ensembles are applied primarily to classification problems rather than regression problems, combining several classifiers. The weighted ensemble is a method to improve the predictive performance by providing some component models with larger weights than others (Bishop, 1995). The weights are determined in accordance with the predictive accuracy because accurate component models are more reliable than inaccurate ones (Opitz & Shavlik, 1996). The weighted ensemble can be defined as

$$\tilde{y}^{Weighted} = \sum_{i=1}^{m} w_i y_i, \ \sum_{i=1}^{m} w_i = 1, \tag{2}$$

where $\tilde{y}^{Weighted}$ is the weighted ensemble value, and w_i indicates the weight of the ith component model. Finally, the weights can be calculated as (Goel *et al.*, 2007)

$$w_i = \frac{w_i^*}{\sum_{j=1}^{m} w_j^*}, \ w_i^* = (e_i + 0.05e_{avg})^{-1}, \ e_{avg} = \frac{1}{m}\sum_{j=1}^{m} e_j, \tag{3}$$

where e_i is the sum of training and validation errors of the ith component model. The component model with a small validation error is generally considered to have high predictive performance and is given a high weight (Zhang *et al.*, 2022). Pham & Olafsson (2019) proposed a new tunable weighted bagged ensemble methodology that determines where to apply weights based on out-of-bag error. Large *et al.* (2019) proposed a probabilistic weighting mechanism that combines probability outputs of effective classifiers by assigning weights based on accuracy estimates derived through cross-validation. Shahhosseini *et al.* (2022) proposed an optimization-based nested algorithm that finds both optimal weights and optimal set of hyperparameters. However, the predictive performance of weighted ensemble decreases when inappropriate weights are assigned due to biased training and validation data set. The mode defined as the most frequent value in a data set can be insensitive to outliers. Kourentzes *et al.* (2014) proposed the mode ensemble that derives a mode using KDE to perform ensemble. KDE is a nonparametric technique for estimating an unknown probability density function, in which previously known density functions of the observed data are averaged. For m prediction values, the kernel density estimator $\hat{f}_h$ can be defined as

$$\hat{f}_h(x) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right), \tag{4}$$

where K is the kernel function, and $h > 0$ is the bandwidth adjusting the smoothness of the distribution. The Gaussian kernel, which is widely used for computational reasons, is defined as

$$\phi_h(x) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{x^2}{2h^2}}. \tag{5}$$

The smoothing effect increases with the bandwidth, and providing an appropriate bandwidth for the data set is required. The optimal bandwidth h that minimizes the mean integrated squared error is given by (Silverman, 2018)

$$h = \left(\frac{4\sigma^5}{3m}\right)^{\frac{1}{5}}, \tag{6}$$

where σ indicates the standard deviation of m prediction values used for the ensemble. The value of x corresponding to the maximum density is the mode of the prediction value distribution and is used as the mode ensemble value y. If the prediction value distribution is multimodal, then a mode close to the prediction values can be selected as the mode ensemble value. However, the predictive performance of the mode ensemble may decrease when there are multiple modes.

Ensembles have been verified to improve predictive performance; thus, NN ensembles have been successfully applied to real-world applications over the past decades. Zhou *et al.* (2002) proposed an automatic pathological diagnosis procedure, which identifies lung cancer cells in the images based on NN ensembles. Jeong & Kim (2005) compared the prediction accuracy of single NN and NN ensemble for the rainfall–runoff model that predicts monthly inflows of dams. The NN ensemble used in this study uses the bootstrap method, which generates a training subset by sampling from the original training set with replacement and derives a final output by averaging individual predictions. This study reported that NN ensemble can further reduce generalization error compared with single NN. In the field of finance, Tsai & Wu (2008) used an NN ensemble that selects results with more than half votes for the problem of bankruptcy prediction and credit scoring. Zaier *et al.* (2010) implemented six NN ensembles to model lake ice thickness and compared the estimation accuracy of ensembles with single NN models. The comparison results show that stacked generalization combining high- and low-level models significantly improves the predictive capability. Linares-Rodriguez *et al.* (2013) built an NN ensemble that averages five optimized NNs to estimate daily global solar radiation over large areas. Shao *et al.* (2014) modeled NN ensemble for fault diagnosis of proton exchange membrane fuel cell systems. In this study, sub-NN models are generated to minimize the correlation among the inputs of the sub-NNs, and weight coefficients are optimized to increase the accuracy of the ensemble. Alobaidi *et al.* (2014) proposed an improved NN ensemble using median statistic for spatial and temporal solar irradiance mapping. Berkhahn *et al.* (2019) used an NN ensemble through mean predictions of NNs generated with random initial weights to forecast real-time urban flooding. Khwaja *et al.* (2020) proposed an NN ensemble method that averages forecasted loads obtained from ensembles of NNs in parallel. In the proposed method, each ensemble consists of sequentially trained NNs based on bootstrapped samples. Moreira *et al.* (2021) employed a NN ensemble that considers the ensemble weights calculated through mixture analysis to forecast medium-term photovoltaic generation. However, existing ensembles have limitations in improving the prediction performance because the stability of prediction accuracy decreases in the presence of outliers or multiple modes. The prediction frequency-based ensemble proposed in this study improves the accuracy and stability of prediction by considering the prediction values around the mode as well as the mode itself to alleviate the uncertainty of the mode. In addition, the proposed ensemble improves the agility of prediction by adaptively setting the minimum frequency criterion used to select the core prediction values to be included in the ensemble.

2.3. Adaptive sampling strategies

Adaptive sampling strategies improve the predictive performance of surrogate models by sequentially adding samples. This section reviews existing adaptive sampling strategies that can be applied to ensembles, including incremental Latin hypercube sampling (iLHS) and prediction variance-based strategies. Latin hypercube sampling (LHS) is a statistical sampling method that performs near-random sampling from multidimensional distributions (McKay *et al.*, 1979). In LHS, the range of each variable is divided into n equivalent intervals, where n is the number of sample points to be generated (Eason & Cremaschi, 2014). n coordinates are obtained for each variable by selecting one coordinate in each interval. One sample point is then constructed by randomly extracting one coordinate from each variable, and distributed n sample points can be obtained by repeating the procedure n times. Thus, the probability that a sample point belongs to each interval is equal for all variables. If a new sample point is added, then some intervals will be overfilled and the equal probability condition for intervals will be violated. Therefore, a new run of the LHS for a new number of samples is required. The iLHS is a space-filling strategy that enhances the LHS to satisfy the equal probability condition of the LHS even for an increased number of samples (Nuchitprasittichai & Cremaschi, 2013). The iLHS divides the range of the variable into equivalent intervals with a new number of samples, discards overlapping samples, and fills in new sample points to maintain the equal probability condition of LHS. Samples can be added sequentially through the iLHS to improve the prediction accuracy of ensembles.

Ensembles make predictions using various component models; thus, different prediction values are derived for a given prediction point, and the prediction variance occurs. The prediction variance represents the prediction uncertainty of the prediction point and is used to determine the next best sample point (Eason & Cremaschi, 2014). Adding samples to sample points with large prediction variance can reduce the overall prediction variance of the ensemble model. The prediction variance-based strategy sequentially adds samples based on the prediction variance. To this end, the prediction variance is calculated for all population sample points in the design space. Then, the sample point with the highest prediction variance is identified as the next best sample point. Since the prediction variance-based strategy determines the next best sample point based on prediction values derived from all component models, the prediction variance may contain errors due to component models with low accuracy, providing an inappropriate next best sample point. To solve this problem, this study proposes the core prediction variance-based strategy that can efficiently and accurately improve prediction accuracy by searching for prediction points with high uncertainty through core prediction values excluding outlier predictions. Therefore, the computational and cost burden caused by the addition of samples can be reduced compared to existing adaptive sampling strategies.

3. Proposed Method

Various ensembles that combine prediction values have been developed to reduce the large prediction variance of NN. Ensembles with high performance require inconsistent component models and their prediction values have low correlation. Component models trained on different training data sets perform prediction from different perspectives, resulting in various prediction values for a given prediction point. Therefore, determining the final prediction value by selecting high-accuracy prediction values is important. Thus, this section proposes the prediction frequency-based ensemble, which identifies the prediction values expected to be close to the true response. In addition, the adaptive sampling strategy for the proposed ensemble is introduced to enhance the predictive performance of ensemble model efficiently by sequentially adding samples.

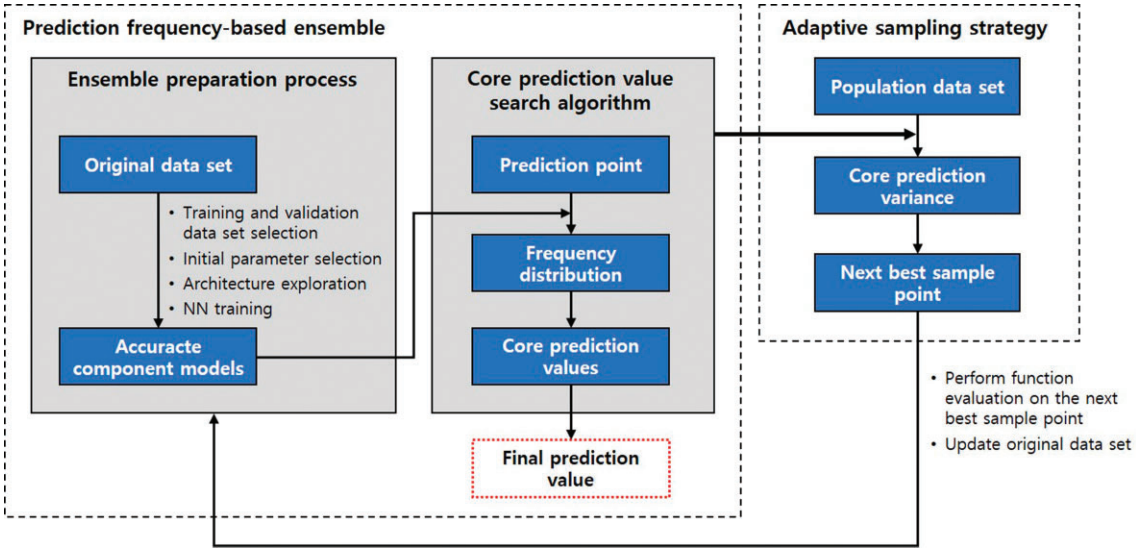

Figure 1: Information flow of the proposed method.

The basic premise of ensembles is that the computational cost of training a large number of NN models is not prohibited. In actual engineering problems, the acquisition of new data entails a heavy computational time, and thus the time required to train NN models becomes relatively insignificant. The information flow of the proposed method is shown in Fig. 1 The prediction frequency-based ensemble comprises an ensemble preparation process and a core prediction value search algorithm. For a population data set, the adaptive sampling strategy uses the core prediction search algorithm to identify the next best sample point and update the original data set. The detailed explanations for the proposed method will be provided in the following sections.

3.1. Prediction frequency-based ensemble

For NN ensemble, an ensemble preparation process that generates diverse component models with high predictive performance is required. Once sufficiently diverse and accurate component models are available, the proposed prediction frequency-based ensemble identifies core prediction values, which are expected to be concentrated near the true responses. A core prediction value search algorithm is proposed to identify core prediction values. In this study, the target of NN ensemble is regression problems, and mean square error (MSE) is used as a performance metric to measure error in training and validation processes.

3.1.1. *Ensemble preparation process*

Component models that are sufficiently diverse and have high predictive performance are required to perform ensemble. In the ensemble preparation process, various component models are created in accordance with the combination of training and validation data set and the selection of initial parameters, and models with high predictive performance are selected among them. The flowchart of the ensemble preparation process is shown in Fig. 2 Searching for the architecture of NN is required before generating component models because the architecture of NN has a considerable influence on predictive accuracy (Elsken *et al.*, 2019). In order to explore the architecture of NNs, architectures created by combinations of various numbers of neurons and layers are set as candidates. For each candidate architecture, NN is trained using various combinations of training and validation data sets, and

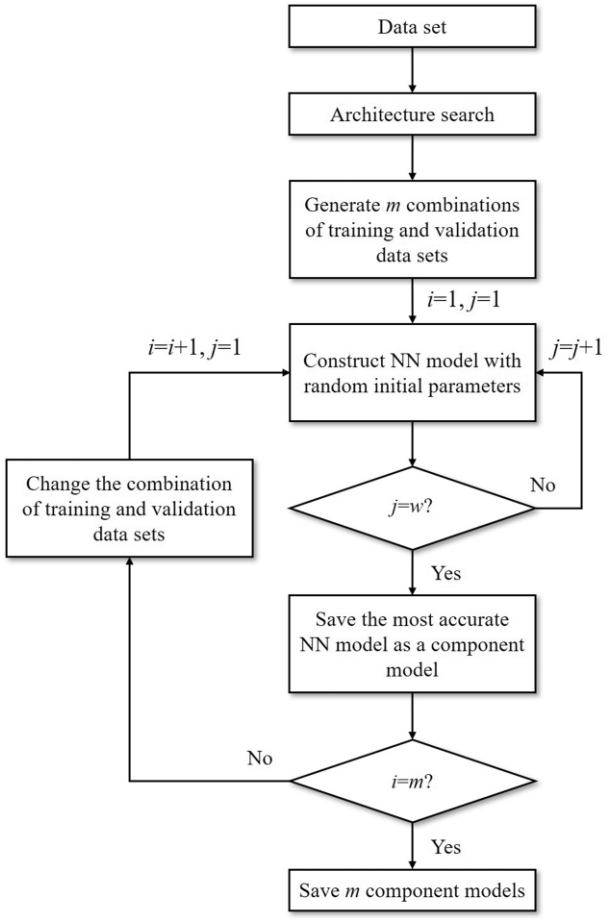

Figure 2: Flowchart of the ensemble preparation process.

then the architecture that results in the smallest sum of training and validation errors is selected. This chosen architecture can be considered to show good training and validation performance for various combinations of training and validation data sets. m combinations of training and validation data sets are generated after determining the architecture of the NN to create m compo-

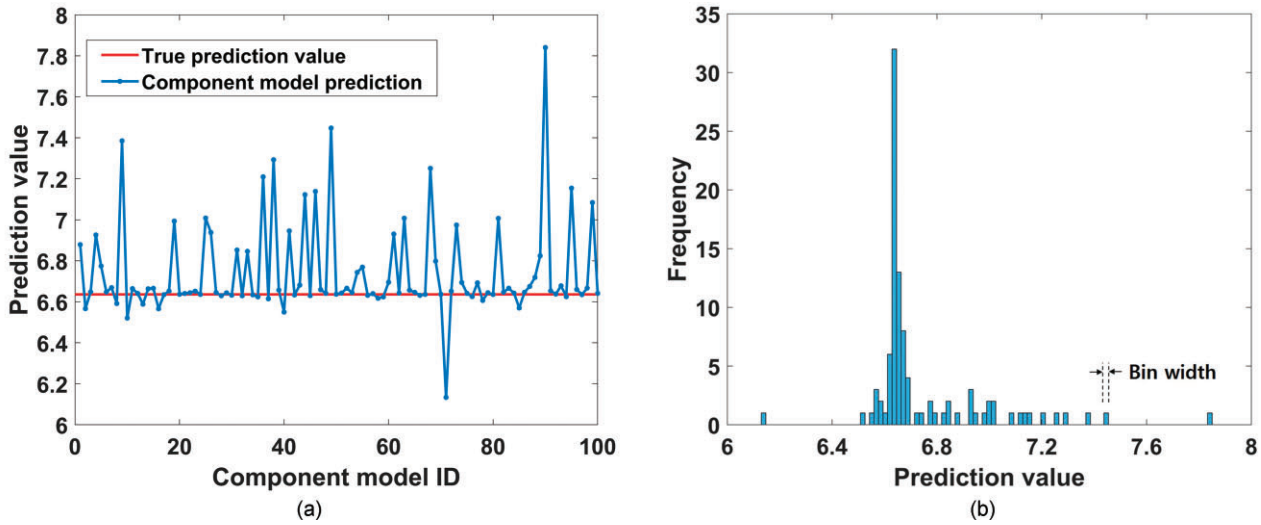

Figure 3: Example of 100 prediction values: (a) comparison between 100 prediction values and true response, (b) frequency distribution of 100 prediction values with 100 bins.

nent models. The combinations are generated on the basis of random sampling with replacement to ensure diversity and prevent bias in training and validation data sets. For each combination, generating the NN model with random initial parameters is repeated w times to generate w candidate NN models. The model with the smallest sum of training and validation errors among w NN models is then estimated to be accurate and saved as a component model for the combination. The number of candidate NN models w is empirically determined to be 100 since since it is confirmed that the NN model with high accuracy is likely to be included among 100 candidate NN models. The accuracy comparison of component models according to the number of candidate NN models w is presented in Section 4. Finally, this process is repeated for m combinations of training and validation data sets, and m accurate component models are obtained. Since the ratio of training and validation data sets is constant regardless of the size of the original data set, constructing sufficiently diverse combinations of training and validation data sets from the original data set affects the accuracy of the ensemble. Kourentzes et al. (2014) compared the accuracy of ensembles by varying the number of component models from 10 to 100, and reported that the highest accuracy can be obtained consistently when 40 or more component models are used in the ensemble. Therefore, 100 component models considered to be sufficient are assumed for ensemble in this study; thus, 100 combinations of training and validation data sets are used. From various problems using different sizes of original data sets, 100 component models are verified to be sufficient to guarantee the highest ensemble accuracy.

3.1.2. *Core prediction value search algorithm*

This section presents the proposed core prediction value search algorithm that identifies core prediction values, which are supported by multiple component models and expected to be concentrated near the true responses. The basic idea of the prediction frequency-based ensemble is to include only highly correlated core prediction values that can be identified through statistical analysis of the prediction values. The proposed ensemble enables appropriate selection of various prediction values obtained from accurate component models covering various combinations of training and validation data sets. Each component model can be regarded as a local model specialized to a specific training and validation data set used for training. Component models in which many samples near a prediction point are used for training have a high probability of deriving prediction values close to the true response. In contrast, component models trained only with samples far away from the prediction point are more likely to derive prediction values that deviate significantly from the true response. Based on this probability and prediction frequency, the proposed ensemble selects core prediction values that have a high probability of being close to the true response. The prediction values with small frequency is unlikely to be close to the true response and can be excluded from the ensemble. The proposed ensemble also classifies the prediction values supported by a certain prediction frequency as core prediction values in order to cope with the uncertainties of prediction. The frequency distribution consisting of prediction values obtained from various component models shows different types of distributions depending on the bin width. Fig. 3 demonstrates an example of 100 prediction values derived from 100 component models for a given prediction point and the frequency distribution of 100 prediction values with 100 bins. The comparison between 100 prediction values and true response is depicted in Fig. 3a and the frequency distribution of 100 prediction values with 100 bins is presented in Fig. 3b. The equation and prediction point used for the example are given as

$$g = 2x_1^2 - 1.05x_1^4 + \frac{x_1^6}{6} + x_1x_2 + x_2^2, x_i = [1.3752, -2.2960]. \quad (7)$$

The number of samples in the data set used to build the NN model is 50, and the ensemble preparation process described in Section 3.1.1 is used to generate 100 component models. In the figure, a number of prediction values are located in the vicinity of the true response. The correlation increases as the prediction values become close to each other, and prediction values with high correlation can be classified among all prediction values even if the true response is unknown. In the frequency distribution, the most frequent bin identifies highly correlated prediction values and contains additional prediction values as the bin width increases. If the bin width is too small, then the prediction values that are expected to be close to the true response cannot be sufficiently captured and thus the uncertainties of prediction cannot be considered. This insufficiency can be solved by increasing the bin width, which has the effect of increasing the number of core prediction

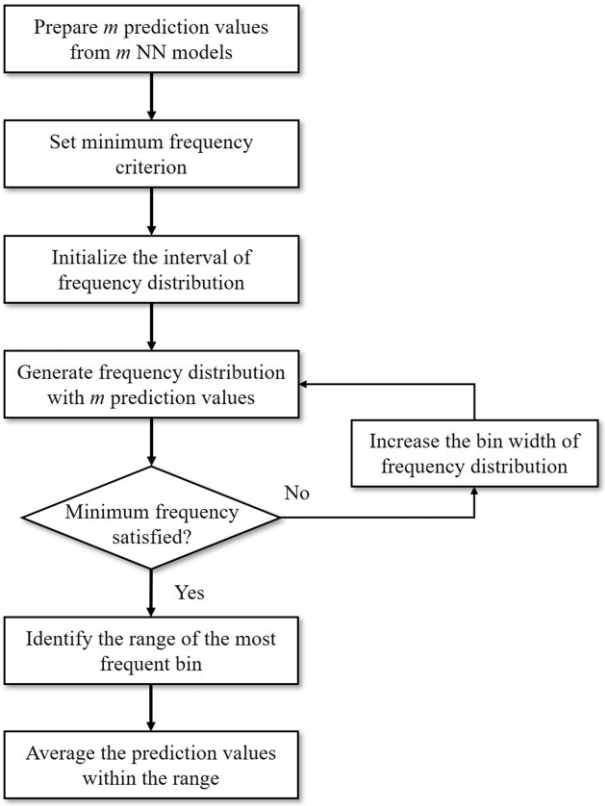

Figure 4: Flowchart of the prediction frequency-based ensemble.

values. The bin width is increased until the frequency of the most frequent bin reaches a minimum frequency criterion, which is the frequency at which the most frequent bin should satisfy. In the proposed core prediction value search algorithm, the frequency that results in the highest prediction accuracy for the original data set can be selected as the minimum frequency criterion. Although subsets of the original data set are used for training, since multiple component models generated from various training data sets are used for ensemble, the original data set can be used to the extent of searching for the optimal frequency criterion. Adaptively setting the minimum frequency criterion used to select the core prediction values helps improve the prediction agility of the prediction frequency-based ensemble. As in the example, the prediction frequency-based ensemble can cope with outliers caused by large non-linearity and insufficient number of samples, and improves the accuracy and stability of prediction by considering prediction values around the mode.

The flowchart of the prediction frequency-based ensemble is shown in Fig. 4 The following steps explain the prediction frequency-based ensemble employing the core prediction value search algorithm.

Step 1. Prepare m prediction values using m component models generated from the ensemble preparation process.
Step 2. Set a minimum frequency criterion that the frequency of the most frequent bin should satisfy.
Step 3. Initialize the interval of frequency distribution using the number of prediction values as the number of bins.
Step 4. Generate frequency distribution with m prediction values.
Step 5. If the frequency of the most frequent bin does not satisfy the minimum frequency criterion, then increase the bin width and return to Step 4. If the minimum frequency criterion is satisfied, then identify the range of the most frequent bin.
Step 6. Average of the prediction values within the range as the final prediction value.

3.2. Adaptive sampling strategy for prediction frequency-based ensemble

In this section, a new adaptive sampling strategy, that is, the core prediction variance-based strategy, is proposed to improve the predictive performance of the proposed prediction frequency-based ensemble efficiently. In the field of Kriging, various learning functions and update schemes have been developed to reduce the uncertainty of prediction by adaptively adding samples based on Kriging variance (Echard *et al.*, 2011). An individual NN model cannot quantify the prediction uncertainty; however, ensembles can derive the prediction variance for each prediction point through the prediction values obtained from component models (Yu *et al.*, 2008). The prediction variance is used to determine the next best sample point, and adding samples to the sample points with large prediction variance can reduce the overall prediction variance of the ensemble model (Eason & Cremaschi, 2014).

The prediction frequency-based ensemble makes predictions based on core prediction values. Thus, a new core prediction variance-based strategy that adds samples based on a core prediction variance indicating the variance of the core prediction values can be proposed. Prediction values supported by only a few number of component models can be considered outliers and must thus be excluded in quantifying the uncertainty of the prediction point. Therefore, the core prediction variance-based strategy quantifies the prediction uncertainty of the prediction point based on the core prediction variance and explores prediction points with high uncertainty despite the support of a certain number of component models. The proposed strategy is different from existing adaptive sampling strategies in that it treats the variance of selected core prediction values as prediction uncertainty, and the prediction accuracy of the ensemble model can be improved by adding samples to the prediction points with high core prediction variance. The core prediction variance-based strategy, which sequentially adds samples based on the core prediction variance, comprises the following procedures.

Step 1. Generate population S throughout the design space.
Step 2. Calculate the core prediction variance for all the sample points in S by using the ensemble model.
Step 3. Identify the sample point with the highest core prediction variance as the next best sample point.
Step 4. Perform function evaluation on the next best sample point and update the original data set.
Step 5. Construct the ensemble model according to the updated sample set.

4. Case Studies

The effectiveness of the proposed prediction frequency-based ensemble and adaptive sampling strategy is verified in this section through various case studies.

4.1. Surrogate model accuracy

This section presents the results obtained from various surrogate modelling methods, including Kriging and ensembles. A total of 500 test sets are used and five surrogate modelling methods are compared to investigate the prediction accuracy of surrogate

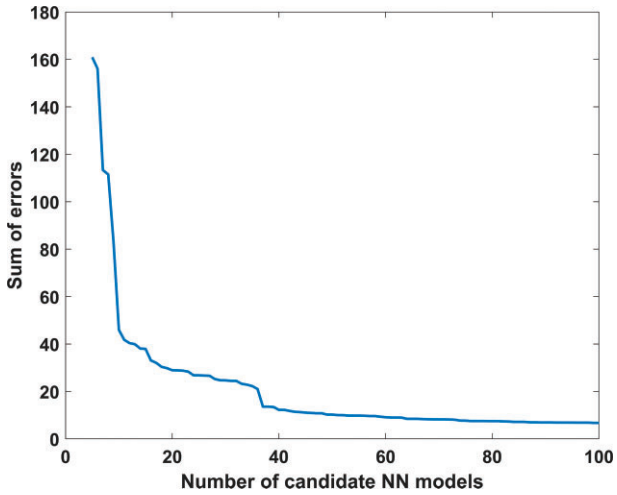

Figure 5: Comparison of the sum of errors of component models according to the number of candidate NN models.

modelling methods: (1) Method 1 is Kriging; (2) Method 2 is average ensemble; (3) Method 3 is weighted ensemble; (4) Method 4 is mode ensemble; and (5) Method 5 is the proposed prediction frequency-based ensemble. For weighted ensemble, high weight is given to a component model with a small sum of training and validation errors to prevent a poorly trained component model with a low validation error from receiving high weight. For an accuracy metric of each method, normalized root MSE (NRMSE) is calculated as

$$\mathrm{NRMSE} = \frac{1}{\bar{y}}\sqrt{\frac{\sum_{k=1}^{n_{test}} (y_k - \hat{y}_k)^2}{n_{test}}}, \tag{8}$$

where n_{test} is the number of test sets, y_k is the true response of the kth test set, $\hat{y}_k$ is the estimated value of the kth test set, and $\bar{y}$ is the mean of true responses. The ensemble preparation process presented in Section 3.1.1 is equally applied to all ensemble methods. In ensemble methods, the ratios for training and validation sets are 0.8 and 0.2, respectively. In this study, fully connected layer is applied and the hyperbolic tangent function is used as the activation function. For optimization process, the Levenberg–Marquardt (LM) algorithm is used. To prevent overfitting, early stopping and regularization that reduces the weights and biases are applied. The NN training is conducted using a standard desktop (Intel XEON CPU @ 192.0 GB GPU and 2.70 GHz).

4.1.1. *Numerical example I: 2D mathematical problem*

A 2D mathematical problem is formulated as shown below to compare the results of a highly non-linear problem with insufficient number of samples.

$$g = 2x_1^2 - 1.05x_1^4 + \frac{x_1^6}{6} + x_1x_2 + x_2^2, x_i \in [-3, 3]. \tag{9}$$

The number of samples in the data set used to build the surrogate model is 50, and the random variables x_i are uniform random variables. To verify that 100 candidate NN models are sufficient to derive an accurate component model in the ensemble preparation process, the sum of training and validation errors of component models according to the number of candidate NN models is compared in Fig. 5 The comparison result shows that 100 candidate NN models provide sufficiently accurate component models. In Method 5, the comparison results of the prediction accuracy for the original data set obtained by varying minimum frequency criteria show that the highest accuracy is achieved when the minimum frequency criterion is set to 20%. NRMSE results of test sets obtained using various minimum frequency criteria are compared and the results are shown in Table 1. The results verify that the prediction accuracy is highest when the frequency of the most frequent bin is larger than or equal to 20% of the total bins.

The NRMSE results obtained from the five methods are shown in Table 2, and the box plots of relative errors are plotted in Fig. 6 In the box plots, the median values of Methods 1, 2, 3, 4, and 5 are 1.7218, 1.4514, 0.5879, 0.1824, and 0.0682, respectively. The results verify that the prediction accuracy of the proposed prediction frequency-based ensemble is the most accurate. The prediction accuracy of Method 1 (Kriging) is lower than that of ensemble methods because the interpolation-based Kriging model becomes significantly inaccurate when the problem is highly nonlinear and the number of samples is insufficient. Among ensemble methods, the prediction accuracy of the average ensemble is low. This finding is due to the inclusion of all biased outliers of average ensemble in the prediction values. The weighted ensemble can cope with outliers better than the average ensemble because different weights are given to the prediction values according to the predictive performance of the component model. However, the prediction accuracy is lower than Methods 4 and 5 because outliers are still considered in the ensemble. Thus, Methods 4 and 5 can exclude outliers in the ensemble because these methods focus on frequent prediction values. Compared with Method 4, Method 5 further improves the prediction accuracy. Method 4 considers only the mode, whereas Method 5 considers the prediction values around the mode as well as the mode itself, thereby coping with the uncertainty of the mode.

The NRMSE results of Method 5 obtained using various numbers of samples in the data set are presented in Table 3 to further investigate the impact of the number of samples in the data set used to build the component model. If the number of samples is insufficient, then the prediction accuracy of each component model decreases and overfitting occurs, resulting in large errors in ensemble method. This indicates that the accuracy of component models needs to be secured to some extent in order for the proposed ensemble method to operate well. The results also confirmed that the prediction accuracy of Method 5 is significantly improved as the number of samples and the prediction accuracy of each component model increase.

4.1.2. *Numerical example II: modified 4D powell function*

A modified 4D Powell function with severe non-linearity is introduced in this section to compare the prediction accuracy of the five methods. The modified Powell function is formulated as

$$g = (x_{4i-3} + 10x_{4i-2})^2 + 5(x_{4i-1} - x_{4i})^2 + (x_{4i-2} - 2x_{4i-1})^6 + 10(x_{4i-3} - x_{4i})^4, x_i \in [0, 10]. \tag{10}$$

The number of samples in the data set used to build the surrogate model is 70, and the random variables x_i are uniform random variables. The NRMSE results of the modified 4D Powell function are shown in Table 4, and the box plots of relative errors are plotted in Fig. 7 In the box plots, the median values of Methods 1, 2, 3, 4, and 5 are 455.8085, 7.4849, 6.7684, 4.7140, and 3.6897, respectively. The number of samples compared with the dimension of the problem is insufficient and the non-linearity of the problem is high; thus, the prediction accuracy of Method 1 (Kriging) becomes significantly lower than that of ensemble methods. For a given prediction point, ensemble methods derive the prediction value with high accuracy by using various component mod-

Table 1: The NRMSE results obtained using various minimum frequency criteria (Method 5).

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0033	0.0032	0.0040	0.0064	0.0062	0.0087	0.0100	0.0134	0.0163	0.0167

Table 2: The NRMSE results of 2D mathematical problem.

Method 1	Method 2	Method 3	Method 4	Method 5
0.0658	0.0168	0.0109	0.0067	0.0032

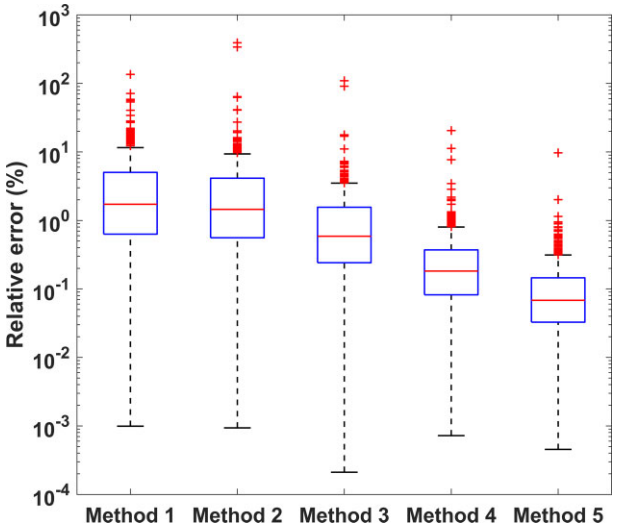

Figure 6: Box plots of relative errors obtained from 2D mathematical problem.

Table 3: The NRMSE results of Method 5 obtained using various number of samples in the data set.

10 samples	20 samples	30 samples	40 samples	50 samples
1.2148	0.4811	0.1655	0.0162	0.0032

Table 4: The NRMSE results of 4D Powell function.

Method 1	Method 2	Method 3	Method 4	Method 5
0.9028	0.1195	0.1209	0.0516	0.0466

els with high predictive performance. Methods 2 (average ensemble) and 3 (weighted ensemble) show similar accuracy. In Method 3, weights that depend only on the global accuracy of the component model are given to the prediction values obtained from a given prediction point; therefore, the prediction accuracy of component models according to the prediction point cannot be reflected. Methods 4 and 5, which select different component models according to the prediction point, show high prediction accuracy, and the proposed Method 5 shows the highest prediction accuracy.

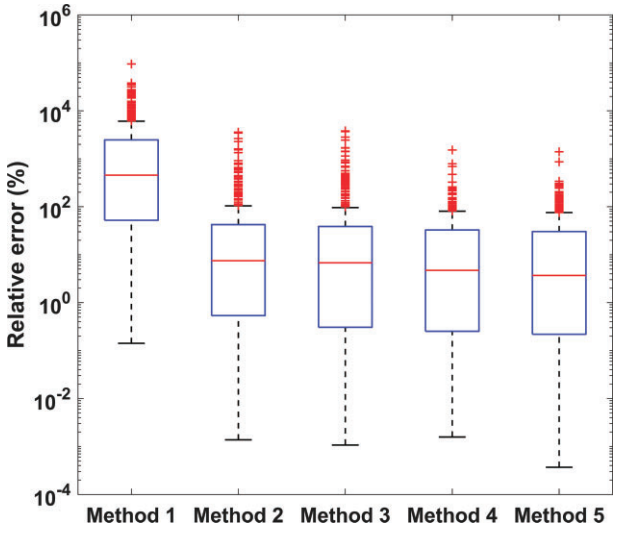

Figure 7: Box plots of relative errors obtained from 4D Powell function.

Table 5: The ranges of the input variables.

Input variables	Description	Range
r_w	Radius of borehole (m)	[0.05, 2]
R	Radius of influence (m)	[100, 50 000]
T_u	Transmissivity of upper aquifer (m^2/yr)	[63 070, 115 600]
H_u	Potentiometric head of upper aquifer (m)	[990, 1110]
T_l	Transmissivity of lower aquifer (m^2/yr)	[63.1, 116]
H_l	Potentiometric head of lower aquifer (m)	[700, 820]
L	Length of borehole (m)	[1120, 2680]
K_w	Hydraulic conductivity of borehole (m/yr)	[1500, 30 000]

4.1.3. *Numerical example III: 8D borehole model*

The borehole model used to describe the flow of water through a borehole is given as (Joseph *et al.*, 2008)

$$g = \frac{2\pi T_u (H_u - H_l)}{\ln(r/r_w)\left[1 + \frac{2LT_u}{\ln(r/r_w) r_w^2 K_w} + \frac{T_u}{T_l}\right]}, \tag{11}$$

where the response is the flow rate of water, and the ranges of the input variables, which are shown in Table 5, are appropriately chosen to increase the non-linearity of the model. The number of samples in the data set used to build the surrogate model is 120, and all the input variables are extracted from uniform distributions. The NRMSE results of the 8D borehole model are shown in Table 6, and the box plots of relative errors are plotted in Fig. 8 In the box plots, the median values of Methods 1, 2, 3, 4, and 5 are 10.8631, 0.5950, 0.4174, 0.3203, and 0.3123, respectively. The NRMSE results verify that Method 5 has higher prediction accuracy than other methods.

Table 6: The NRMSE results of 8D borehole model.

Method 1	Method 2	Method 3	Method 4	Method 5
0.2103	0.0172	0.0127	0.0099	0.0083

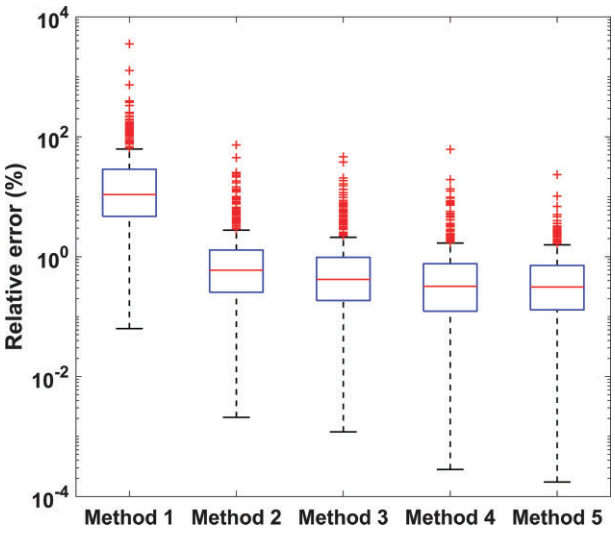

Figure 8: Box plots of relative errors obtained from 8D borehole model.

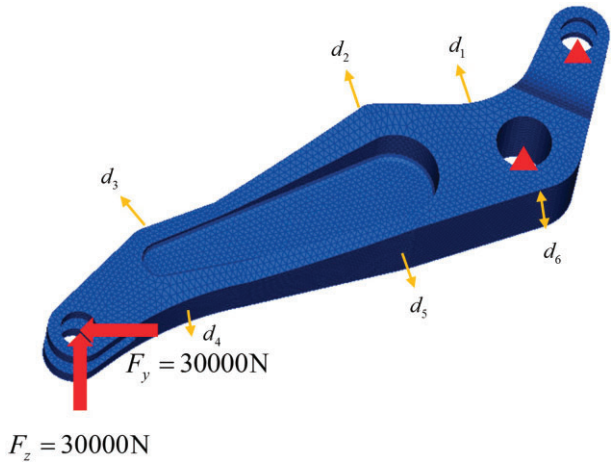

Figure 9: Boundary and loading conditions of 6D arm model.

4.1.4. *Engineering example: 6D arm model*

This section presents an engineering example predicting the maximum von Mises stress that occurs in a 6D arm model (Altair, 2017). The loading and boundary conditions of the model are presented in Fig. 9, and the ranges of input variables are shown in Table 7. The morphing parameters indicated by "dv" constitute the input variables and control the geometry of the arm. The number of samples in the data set used to build the surrogate model is 100, and all the input variables are extracted from uniform distributions. The NRMSE results of the 6D arm model are shown in Table 8, and the box plots of relative errors are plotted in Fig. 10 In the box plots, the median values of Methods 1, 2, 3, 4, and 5 are 6.4123, 0.8602, 0.8627, 0.8362, and 0.8273, respectively. The nonlinearity of this example is moderate; thus, the difference in prediction accuracy between ensemble methods is small. The NRMSE results show that Method 5 has higher prediction accuracy than other methods, verifying that the proposed Method 5 can improve the prediction accuracy compared with other ensemble methods even in engineering examples.

Table 7: Properties of input variables of 6D arm model.

Input variables	Range
dv_1	[−0.5, 4]
dv_2	[0, 3]
dv_3	[−1, 5]
dv_4	[−1, 1]
dv_5	[−1, 5]
dv_6	[−1, 4]

Table 8: The NRMSE results of 6D arm model.

Method 1	Method 2	Method 3	Method 4	Method 5
0.1249	0.0207	0.0206	0.0189	0.0182

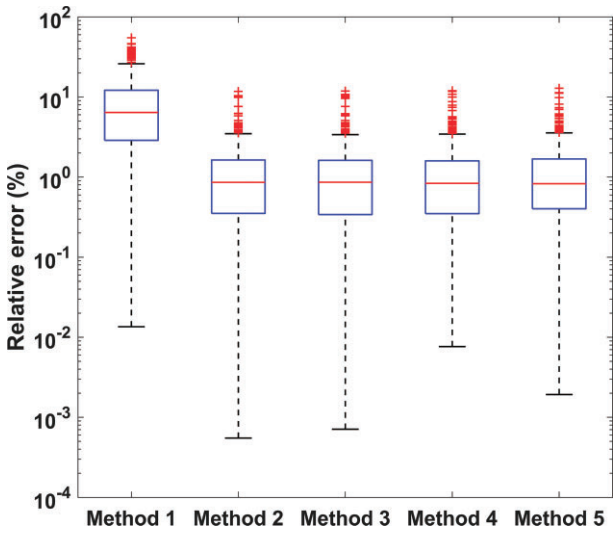

Figure 10: Box plots of relative errors obtained from the 6D arm model.

Regarding computational costs, it is difficult to directly compare Method 1 with other ensemble methods because parallel computing can be used in ensemble methods. Depending on time required for simulations, a higher computational cost is required for Method 1 to have a similar level of prediction accuracy as ensemble methods. Since the same ensemble preparation process is used for Methods 2–5, the same computational cost is required until the ensemble preparation. Then, the average CPU times required in the process of deriving the final ensemble value after ensemble preparation are 0.0036 s, 0.0126 s, 0.5447 s, and 0.797 s in Methods 2, 3, 4, and 5, respectively. These CPU times are negligible compared to the time required for simulation or ensemble preparation. Therefore, ensemble methods show little difference in computational costs.

4.2. Adaptive sampling strategy for prediction frequency-based ensemble

In this section, the proposed core prediction variance-based strategy presented in Section 3.2 is verified using several examples. For comparison, the previously developed space-filling strategy and the prediction variance-based strategy are compared. One sam-

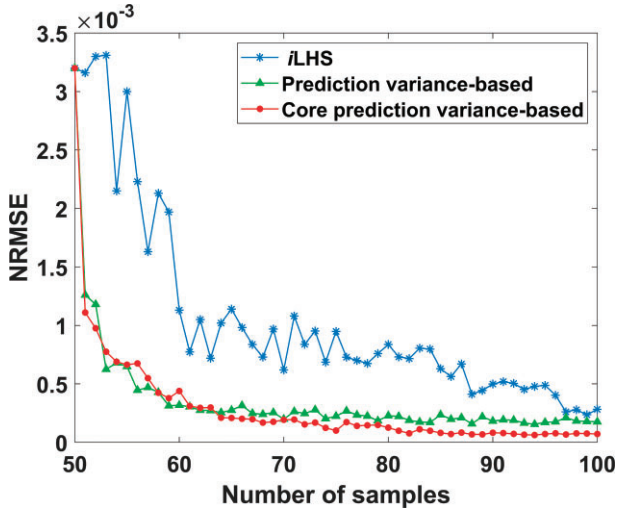

Figure 11: The NRMSE results according to the number of samples (2D mathematical problem).

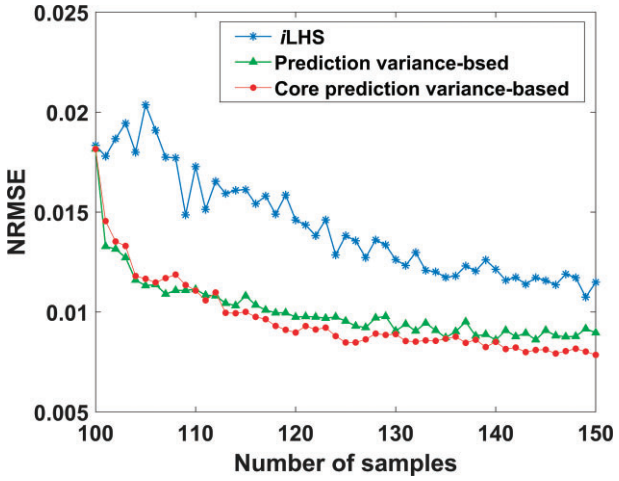

Figure 12: The NRMSE results according to the number of samples (6D arm model).

ple is sequentially added in accordance with each adaptive sampling strategy to verify the performance of each adaptive sampling strategy, and the prediction accuracy of the updated prediction frequency-based ensemble is compared through NRMSE. The iLHS, prediction variance-based strategy, and the proposed core prediction variance-based strategy are compared.

4.2.1. *Numerical example: 2D mathematical problem*

Three adaptive sampling strategies are applied for the 2D mathematical problem presented in Section 4.1.1. Initially, the number of samples in the data set used to build the prediction frequency-based ensemble is 50, and samples are sequentially added in accordance with the adaptive sampling strategy until the total number of samples reaches 100. The NRMSE results according to the number of samples are presented in Fig. 11. In all three strategies, the predictive performance of the prediction frequency-based ensemble increases as samples are added. The results show that the proposed core prediction variance-based strategy most efficiently increases the predictive performance of the prediction frequency-based ensemble. In the case of iLHS, the improvement of prediction accuracy is relatively low because the prediction uncertainty at each prediction point is disregarded and samples are added randomly. The results verify that adding samples considering the core prediction variance rather than the prediction variance improves the predictive performance of the prediction frequency-based ensemble efficiently. In the early stage of adding samples, the prediction variance-based strategy shows a slightly higher prediction accuracy; however, the core prediction variance-based strategy becomes more accurate as additional samples are included. When the number of samples is 100, the core prediction variance-based strategy shows 74.7% and 59.4% lower NRMSE than iLHS and prediction variance-based strategy, respectively. In addition, to reach the NRMSE of the core prediction variance-based strategy obtained with 100 samples, iLHS requires 287 samples and the prediction variance-based strategy requires 219 samples.

4.2.2. *Engineering example: 6D arm model*

In this section, three adaptive sampling strategies are compared for the 6D arm model presented in Section 4.1.4. Initially, 100 samples are used to build the prediction frequency-based ensemble, and 50 samples are added sequentially according to the adaptive sampling strategy. The NRMSE results according to the number of samples are presented in Fig. 12 The results verify that the proposed core prediction variance-based strategy most efficiently improves the predictive performance of the prediction frequency-based ensemble in the engineering example. As in Section 4.2.1, the NRMSE difference between the core prediction variance- and prediction variance-based strategies becomes clear as samples are added. When the number of samples is 100, the core prediction variance-based strategy shows 31.6% and 12.3% lower NRMSE than iLHS and prediction variance-based strategy, respectively. To reach the NRMSE of the core prediction variance-based strategy obtained with 150 samples, iLHS requires 196 samples and the prediction variance-based strategy requires 189 samples. In the case of engineering applications, high computational cost is involved in performing simulation. Therefore, the proposed core prediction variance-based strategy helps obtain a prediction frequency-based ensemble model with high predictive performance using relatively low computational effort.

5. Conclusions

NNs are considered promising surrogate models with good prediction accuracy, and various ensembles have been developed to reduce the large prediction variance of NN. This study proposes the prediction frequency-based ensemble that identifies core prediction values, which are expected to be concentrated near the true response, and derives the average of the core prediction values as the final prediction value. In the ensemble preparation process, various component models are generated in accordance with the combination of training and validation data sets and the selection of initial parameters, and models with high predictive performance are selected. For a given prediction point, various prediction values can be obtained from the component models, and the prediction frequency-based ensemble performs statistical analysis with the frequency distribution consisting of the various prediction values to classify the concentrated prediction values. A core prediction variance-based strategy, which sequentially adds samples based on the core prediction variance, is proposed to improve the predictive performance of the proposed prediction frequency-based ensemble efficiently. The core prediction variance is the variance of the core prediction values. After all the

core prediction variances are calculated for the generated population, the sample point with the highest core prediction variance becomes the next best sample point. The proposed strategy is different from existing adaptive sampling strategies in that the variance of selected core prediction values is treated as prediction uncertainty. This results in efficient and delicate sample recommendations. Various case studies verified that the prediction accuracy of the prediction frequency-based ensemble is higher than that of Kriging and other existing ensemble methods and the core prediction variance-based strategy efficiently improves the predictive performance of the prediction frequency-based ensemble.

The contribution of this study can be summarized as follows. First, the proposed prediction frequency-based ensemble can improve the predictive performance by excluding outlier prediction values for highly non-linear problems with insufficient data set. Secondly, the proposed prediction frequency-based ensemble can cope with the uncertainty of the most frequent value by adaptively identifying the most frequent bin. Thirdly, the proposed core prediction variance-based strategy efficiently improves the predictive performance of the prediction frequency-based ensemble by calculating the core prediction variance and sequentially adding samples to the point where the actual prediction error is expected to be large.

This study also has some limitations. The non-linearities and properties of the problems required for the prediction frequency-based ensemble to work efficiently should be investigated quantitatively. In addition, the effect of the minimum frequency criterion on the prediction accuracy should be studied. In the future work, the prediction frequency-based ensemble can be applied to increase the efficiency and accuracy of reliability analysis. A learning function that recommends the sample point near the limit state with a large core prediction variance as the next best sample point must be developed to increase the accuracy of prediction frequency-based ensemble efficiently. In addition, a hybrid sampling strategy that combines both core prediction variance- and prediction variance-based strategies can be developed to alleviate the problem of the core prediction variance-based strategy, which shows a lower accuracy improvement in the early stage. To this end, the prediction variance-based strategy can be used at the beginning and the core prediction variance-based strategy can be used after a certain number of samples are added.

Replication of results

The code for this paper is available at the website: (https://drive.google.com/file/d/1koCF_SSwRcXMs28ThIfCx4-ft0KCEvUB/view?usp=sharing).

Acknowledgments

This work was supported by the National Research Foundation of Korea (2018R1A5A7025409) and the Ministry of Science and ICT of Korea (No.2022-0-00969).

Conflict of interest statement

The authors declare that there is no conflict of interest.

References

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, **4**, e00938. https://doi.org/10.1016/j.heliyon.2018.e00938.

Alobaidi, M. H., Marpu, P. R., Ouarda, T. B. M. J., & Ghedira, H. (2014). Mapping of the solar irradiance in the UAE using advanced artificial neural network ensemble. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **7**, 3668–3680. https://doi.org/10.1109/JSTARS.2014.2331255.

Altair (2017). Altair HyperStudy tutorials. https://www.altairhyperworks.com. Accessed 24 Jun 2022.

Berkhahn, S., Fuchs, L., & Neuweiler, I. (2019). An ensemble neural network model for real-time prediction of urban floods. *Journal of Hydrology*, **575**, 743–754. https://doi.org/10.1016/j.jhydrol.2019.05.066.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, New York, USA. https://books.google.es/books?id=-aAwQO_-rXwC.

Bouhlel, M. A., He, S., & Martins, J. R. R. A. (2020). Scalable gradient–enhanced artificial neural networks for airfoil shape design in the subsonic and transonic regimes. *Structural and Multidisciplinary Optimization*, **61**, 1363–1376. https://doi.org/10.1007/s00158-020-02488-5.

Chen, X., Chen, X., Zhou, W., Zhang, J., & Yao, W. (2020). The heat source layout optimization using deep learning surrogate modelling. *Structural and Multidisciplinary Optimization*, **62**, 3127–3148. https://doi.org/10.1007/s00158-020-02659-4.

Cheng, J., & Li, Q. S. (2008). Reliability analysis of structures using artificial neural network based genetic algorithms. *Computer Methods in Applied Mechanics and Engineering*, **197**, 3742–3750. https://doi.org/10.1016/j.cma.2008.02.026.

Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting*, **27**, 635–660. https://doi.org/10.1016/j.ijforecast.2011.04.001.

Deng, H., Runger, G., Tuv, E., & Vladimir, M. (2013). A time series forest for classification and feature extraction. *Information Sciences*, **239**, 142–153. https://doi.org/10.1016/j.ins.2013.02.030.

de Santana Gomes, W. J. (2019). Structural reliability analysis using adaptive artificial neural networks. *ASCE- ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, **5**, 41004. https://doi.org/10.1115/1.4044040.

Du, X., He, P., & Martins, J. R. (2021). Rapid airfoil design optimization via neural networks-based parameterization and surrogate modelling. *Aerospace Science and Technology*, **113**, 106701. https://doi.org/10.1016/j.ast.2021.106701.

Eason, J., & Cremaschi, S. (2014). Adaptive sequential sampling for surrogate model generation with artificial neural networks. *Computers & Chemical Engineering*, **68**, 220–232. https://doi.org/10.1016/j.compchemeng.2014.05.021.

Echard, B., Gayton, N., & Lemaire, M. (2011). AK–MCS: An active learning reliability method combining Kriging and Monte Carlo simulation. *Structural Safety*, **33**, 145–154. https://doi.org/10.1016/j.strusafe.2011.01.002.

Elhewy, A. H., Mesbahi, E., & Pu, Y. (2006). Reliability analysis of structures using neural network method. *Probabilistic Engineering Mechanics*, **21**, 44–53. https://doi.org/10.1016/j.probengmech.2005.07.002.

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural architecture search: A survey. *Journal of Machine Learning Research*, **20**, 1–21. http://jmlr.org/papers/v20/18-598.html.

Eskinazi, I., & Fregly, B. J. (2015). Surrogate modelling of deformable joint contact using artificial neural networks. *Medical Engineering*

& *Physics*, **37**, 885–891. https://doi.org/10.1016/j.medengphy.2015.06.006.

Freitag, S., Edler, P., Kremer, K., & Meschke, G. (2020). Multilevel surrogate modelling approach for optimization problems with polymorphic uncertain parameters. *International Journal of Approximate Reasoning*, **119**, 81–91. https://doi.org/10.1016/j.ijar.2019.12.015.

Ganaie, M. A., Hu, M., Malik, A., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, **115**, 105151. https://doi.org/10.1016/j.engappai.2022.105151.

García-Alba, J., Bárcena, J. F., Ugarteburu, C., & García, A. (2019). Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries. *Water Research*, **150**, 283–295. https://doi.org/10.1016/j.watres.2018.11.063.

García-Segura, T., Yepes, V., & Frangopol, D. M. (2017). Multi-objective design of post-tensioned concrete road bridges using artificial neural networks. *Structural and Multidisciplinary Optimization*, **56**, 139–150. https://doi.org/10.1007/s00158-017-1653-0.

Gaspar, B., Teixeira, A. P., & Soares, G. C. (2017). Adaptive surrogate model with active refinement combining Kriging and a trust region method. *Reliability Engineering & System Safety*, **165**, 277–291. https://doi.org/10.1016/j.ress.2017.03.035.

Goel, T., Haftka, R. T., Wei, S., & Queipo, N. V. (2007). Ensemble of surrogates. *Structural and Multidisciplinary Optimization*, **33**, 199–216. https://doi.org/10.1007/s00158-006-0051-9.

Gomes, H. M., Awruch, A. M., & Lopes, P. A. M. (2011). Reliability based optimization of laminated composite structures using genetic algorithms and Artificial Neural Networks. *Structural Safety*, **33**, 186–195. https://doi.org/10.1016/j.strusafe.2011.03.001.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press, Massachusetts, USA. http://www.deeplearningbook.org.

Izeboudjen, N., Larbes, C., & Farah, A. (2014). A new classification approach for neural networks hardware: From standards chips to embedded systems on chip. *Artificial Intelligence Review*, **41**, 491–534. https://doi.org/10.1007/s10462-012-9321-7.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer, New York, USA. https://link.springer.com/book/10.1007/978-1-4614-7138-7.

Jeong, D. I., & Kim, Y. O. (2005). Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction. *Hydrological Processes*, **19**, 3819–3835. https://doi.org/10.1002/hyp.5983.

Joseph, V. R., Hung, Y., & Sudjianto, A. (2008). Blind kriging: A new method for developing metamodels. *Journal of Mechanical Design*, **130**, 31102. https://doi.org/10.1115/1.2829873.

Ju, C., Bibaut, A., & van der Laan, M. (2018). The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, **45**, 2800–2818. https://doi.org/10.1080/02664763.2018.1441383.

Kang, K., Qin, C., Lee, B., & Lee, I. (2019). Modified screening-based kriging method with cross validation and application to engineering design. *Applied Mathematical Modelling*, **70**, 626–642. https://doi.org/10.1016/j.apm.2019.01.030.

Khwaja, A. S., Anpalagan, A., Naeem, M., & Venkatesh, B. (2020). Joint bagged-boosted artificial neural networks: Using ensemble machine learning to improve short-term electricity load forecasting. *Electric Power Systems Research*, **179**, 106080. https://doi.org/10.1016/j.epsr.2019.106080.

Kim, S. Y., & Upneja, A. (2021). Majority voting ensemble with a decision tree for business failure prediction during economic downturns. *Journal of Innovation & Knowledge*, **6**, 112–123. https://doi.org/10.1016/j.jik.2021.01.001.

Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, **41**, 4235–4244. https://doi.org/10.1016/j.eswa.2013.12.011.

Ktari, Z., Leitão, C., Prates, P. A., & Khalfallah, A. (2021). Mechanical design of ring tensile specimen via surrogate modelling for inverse material parameter identification. *Mechanics of Materials*, **153**, 103673. https://doi.org/10.1016/j.mechmat.2020.103673.

Kumar, G. B., Pramod, R., Rao, C. S. P., & Gouda, P. S. S. (2018). Artificial neural network prediction on wear of Al6061 alloy metal matrix composites reinforced with-Al2o3. *Materials Today: Proceedings*, **5**, 11268–11276. https://doi.org/10.1016/j.matpr.2018.02.093.

Large, J., Lines, J., & Bagnall, A. (2019). A probabilistic classifier ensemble weighting scheme based on cross validated accuracy estimates. *Data Mining and Knowledge Discovery*, **33**, 1674–1709. https://doi.org/10.1007/s10618-019-00638-y.

Lee, U., & Lee, I. (2022). Efficient sampling-based inverse reliability analysis combining Monte Carlo simulation (MCS) and feedforward neural network (FNN). *Structural and Multidisciplinary Optimization*, **65**, 8. https://doi.org/10.1007/s00158-021-03144-2.

Lee, U., Park, S., & Lee, I. (2020). Robust design optimization (RDO) of thermoelectric generator system using non-dominated sorting genetic algorithm II (NSGA-II). *Energy*, **196**, 117090. https://doi.org/10.1016/j.energy.2020.117090.

Lehký, D., Slowik, O., & Novák, D. (2018). Reliability-based design: Artificial neural networks and double-loop reliability-based optimization approaches. *Advances in Engineering Software*, **117**, 123–135. https://doi.org/10.1016/j.advengsoft.2017.06.013.

Li, S., Lu, X., Sakai, S., Mimura, M., & Kawahara, T. (2017). Semi-supervised ensemble DNN acoustic model training. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. **2017**, New Orleans, LA, USA. (pp. 5270–5274). https://doi.org/10.1109/ICASSP.2017.7953162.

Lin, J., Li, H., Huang, Y., Huang, Z., & Luo, Z. (2020). Adaptive artificial neural network surrogate model of non-linear hydraulic adjustable damper for automotive semi-active suspension system. *IEEE Access*, **8**, 118673–118686. https://doi.org/10.1109/ACCESS.2020.3004886.

Linares-Rodriguez, A., Ruiz-arias, J. A., Pozo-vazquez, D., & Tovar-pescador, J. (2013). An artificial neural network ensemble model for estimating global solar radiation from Meteosat satellite images. *Energy*, **61**, 636–645. https://doi.org/10.1016/j.energy.2013.09.008.

Liu, Q. F., Iqbal, M. F., Yang, J., Lu, X. Y., Zhang, P., & Rauf, M. (2021). Prediction of chloride diffusivity in concrete using artificial neural network: Modelling and performance evaluation. *Construction and Building Materials*, **268**, 121082. https://doi.org/10.1016/j.conbuildmat.2020.121082.

Lu, Y., Pulasani, P. R., Derakhshani, R., & Guess, T. M. (2013). Application of neural networks for the prediction of cartilage stress in a musculoskeletal system. *Biomedical Signal Processing and Control*, **8**, 475–482. https://doi.org/10.1016/j.bspc.2013.04.004.

McKay, M. D., Conover, W. J., & Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–245. https://doi.org/10.2307/1268522.

Maleki, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Birgani, Y. T., & Rahmati, M. (2019). Air pollution prediction by using an artificial neural network model. *Clean Technologies and Environmental Policy*, **21**, 1341–1352. https://doi.org/10.1007/s10098-019-01709-w.

Moreira, M. O., Balestrassi, P. P., Paiva, A. P., Ribeiro, P. F., & Bonatto, B. D. (2021). Design of experiments using artificial neural network ensemble for photovoltaic generation forecasting. *Renewable and Sustainable Energy Reviews*, **135**, 110450. https://doi.org/10.1016/j.rser.2020.110450.

Moreno-Pérez, J., Bonilla-Petriciolet, A., Mendoza-Castillo, D. I., Reynel-Ávila, H. E., Verde-Gómez, Y., & Trejo-Valencia, R. (2018). Artificial neural network-based surrogate modelling of multi-component dynamic adsorption of heavy metals with a biochar. *Journal of Environmental Chemical Engineering*, **6**, 5389–5400. https://doi.org/10.1016/j.jece.2018.08.038.

Nezhad, H. B., Miri, M., & Ghasemi, M. R. (2019). New neural network-based response surface method for reliability analysis of structures. *Neural Computing and Applications*, **31**, 777–791. https://doi.org/10.1007/s00521-017-3109-2.

Nuchitprasittichai, A., & Cremaschi, S. (2013). An algorithm to determine sample sizes for optimization with artificial neural networks. *AIChE Journal*, **59**, 805–812. https://doi.org/10.1002/aic.13871.

Opitz, D. W., & Shavlik, J. W. (1996). Actively searching for an effective neural network ensemble. *Connection Science*, **8**, 337–353. https://doi.org/10.1080/095400996116802.

Pan, I., Babaei, M., Korre, A., & Durucan, S. (2014). Artificial neural network based surrogate modelling for multi-objective optimisation of geological CO_2 storage operations. *Energy Procedia*, **63**, 3483–3491. https://doi.org/10.1016/j.egypro.2014.11.377.

Papadopoulos, V., Giovanis, D. G., Lagaros, N. D., & Papadrakakis, M. (2012). Accelerated subset simulation with neural networks for reliability analysis. *Computer Methods in Applied Mechanics and Engineering*, **223**, 70–80. https://doi.org/10.1016/j.cma.2012.02.013.

Papadopoulos, V., Soimiris, G., Giovanis, D. G., & Papadrakakis, M. (2018). A neural network-based surrogate model for carbon nanotubes with geometric non-linearities. *Computer Methods in Applied Mechanics and Engineering*, **328**, 411–430. https://doi.org/10.1016/j.cma.2017.09.010.

Pham, H., & Olafsson, S. (2019). Bagged ensembles with tunable parameters. *Computational Intelligence*, **35**, 184–203. https://doi.org/10.1111/coin.12198.

Ren, C., Aoues, Y., Lemosse, D., & Souza De Cursi, E. (2022). Ensemble of surrogates combining Kriging and artificial neural networks for reliability analysis with local goodness measurement. *Structural Safety*, **96**, 102186. https://doi.org/10.1016/j.strusafe.2022.102186.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **8**, e1249. https://doi.org/10.1002/widm.1249.

Schrader, K., & Schauer, M. (2021). Multi-directional design control of plastic crash components by means of domain-predictive feed-forward neural networks. *Structural and Multidisciplinary Optimization*, **64**, 4115–4128. https://doi.org/10.1007/s00158-021-03031-w.

Shahhosseini, M., Hu, G., & Pham, H. (2022). Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications*, **7**, 100251. https://doi.org/10.1016/j.mlwa.2022.100251.

Shao, M., Zhu, X., Cao, H., & Shen, H. (2014). An artificial neural network ensemble method for fault diagnosis of proton exchange membrane fuel cell system. *Energy*, **67**, 268–275. https://doi.org/10.1016/j.energy.2014.01.079.

Shaw, A. R., Smith Sawyer, H., LeBoeuf, E. J., McDonald, M. P., & Hadjerioua, B. (2017). Hydropower optimization using artificial neural network surrogate models of a high-fidelity hydrodynamics and water quality model. *Water Resources Research*, **53**, 9444–9461. https://doi.org/10.1002/2017WR021039.

Silverman, B. (2018). *Density estimation for Statistics and Data Analysis*. Routledge, New York, USA. https://doi.org/10.1201/9781315140919.

Sollich, P., & Krogh, A. (1996). Learning with ensembles: How overfitting can be useful. *Advances in Neural Information Processing Systems*. **8**, 190–196. https://papers.nips.cc/paper/1044-learning-with-ensembles-how-overfitting-can-be-useful.pdf.

Thrampoulidis, E., Mavromatidis, G., Lucchi, A., & Orehounig, K. (2021). A machine learning-based surrogate model to approximate optimal building retrofit solutions. *Applied Energy*, **281**, 116024. https://doi.org/10.1016/j.apenergy.2020.116024.

Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, **34**, 2639–2649. https://doi.org/10.1016/j.eswa.2007.05.019.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, **5**, 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.

Xia, Y., Chen, K., & Yang, Y. (2021). Multi-label classification with weighted classifier selection and stacked ensemble. *Information Sciences*, **557**, 421–442. https://doi.org/10.1016/j.ins.2020.06.017.

Yan, S., Zou, X., Ilkhani, M., & Jones, A. (2020). An efficient multiscale surrogate modelling framework for composite materials considering progressive damage based on artificial neural networks. *Composites Part B: Engineering*, **194**, 108014. https://doi.org/10.1016/j.compositesb.2020.108014.

Yılmaz, İ., Yelek, İ., Özcanan, S., Atahan, A. O., & Hiekmann, J. M. (2021). Artificial neural network metamodelling-based design optimization of a continuous motorcyclists protection barrier system. *Structural and Multidisciplinary Optimization*, **64**, 4305–4323. https://doi.org/10.1007/s00158-021-03080-1.

Yu, L., Lai, K. K., & Wang, S. (2008). Multistage RBF neural network ensemble learning for exchange rates forecasting. *Neurocomputing* **71**, 3295–3302. https://doi.org/10.1016/j.neucom.2008.04.029.

Zaier, I., Shu, C., Ouarda, T. B. M. J., Seidou, O., & Chebana, F. (2010). Estimation of ice thickness on lakes using artificial neural networks ensembles. *Journal of Hydrology*, **383**, 330–340. https://doi.org/10.1016/j.jhydrol.2010.01.006.

Zhang, G. P. (2007). A neural network ensemble method with jittered training data for time series forecasting. *Information Sciences*, **177**, 5329–5346. https://doi.org/10.1016/j.ins.2007.06.015.

Zhang, X., Xie, F., Ji, T., Zhu, Z., & Zheng, Y. (2021). Multi-fidelity deep neural network surrogate model for aerodynamic shape optimization. *Computer Methods in Applied Mechanics and Engineering*, **373**, 113485. https://doi.org/10.1016/j.cma.2020.113485.

Zhang, H., Zhou, C., Zhao, H., & Zhang, Z. (2022). An ensemble model-based method for estimating failure probability function with application in reliability-based optimization. *Applied Mathematical Modelling*, **108**, 445–468. https://doi.org/10.1016/j.apm.2022.03.040.

Zhou, Z. H., Jiang, Y., Yang, Y. B., & Chen, S. F. (2002). Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine*, **24**, 25–36. https://doi.org/10.1016/S0933-3657(01)00094-X.

Zhu, Y., Newbrook, D. W., Dai, P., de Groot, C. H. K., & Huang, R. (2022). Artificial neural network enabled accurate geometrical design and optimisation of thermoelectric generator. *Applied Energy*, **305**, 117800. https://doi.org/10.1016/j.apenergy.2021.117800.

Received: March 13, 2023. **Revised:** June 26, 2023. **Accepted:** June 26, 2023