International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00595-w

REGULAR PAPER

Storage of weights and retrieval method (SWARM) approach for neural networks hybridized with conformal prediction to construct the prediction intervals for energy system applications

Waqar Muhammad Ashraf[1] · Vivek Dua[1]

Received: 13 November 2023 / Accepted: 20 June 2024

Abstract

The prediction intervals represent the uncertainty associated with the model-predicted responses that impacts the sequential decision-making analytics. Here in this work, we present a novel model-based data-driven approach to construct the prediction intervals around the model-simulated responses using artificial neural network (ANN) model. The loss function is modified with least mean square error and standard deviation between the model-simulated and actual responses for the online-training mode of ANN model development. The parameters (weights and biases) stored during the model development are extracted and are deployed to construct the prediction intervals with 95% confidence level for the test datasets of the three energy systems-based case studies including: crease recovery angle, energy efficiency cooling & energy efficiency heating and gas turbine power plant & coal power plant which are taken from literature, benchmark datasets and industrial-scale applications, respectively. The developed ANN models present root-mean-squared error of 1.20% and 0.52% on test dataset for energy efficiency cooling and energy efficiency heating, respectively. The width of prediction intervals made by the proposed approach, called as Storage of Weights And Retrieval Method (SWARM), incorporates the information available for each test observation during the model training and the SWARM-based prediction intervals are compared to those of inductive conformal prediction (ICP) technique. It is noted that SWARM technique offers better locally adaptive prediction intervals than those of ICP, highlighting the effectiveness of the SWARM technique for the estimation of prediction intervals for the case studies. This research presents a novel data-driven approach to construct the prediction intervals using the model-based information that can be applied on different real-life applications.

Keywords Prediction interval · Uncertainty quantification · Conformal prediction · SWARM

Abbreviations

ANN	Artificial neural network
CRA	Crease recovery angle
ENC	Energy efficiency cooling
ENH	Energy efficiency heating
HLN	Weight links from hidden to output layer
ICP	Inductive conformal prediction
R^2	Coefficient of determination
RMSE	Root-mean-squared-error
TCP	Transductive conformal prediction

List of symbols

b_1	Bias introduced on hidden layer neurons
b_2	Bias introduced on output layer neuron
D	True value
E_{cal}	Non-conformity value corresponding to calibration dataset
E_{epoch}	Non-conformity value corresponding to epoch
$\mathcal{L}$	Loss function
m	Number of hidden layer neurons
N	Number of observations
$\widehat{q}_{1-\alpha}(\mathrm{E_{cal}})$	Quantile value on $1-\alpha$ from E_{cal}

✉ Waqar Muhammad Ashraf
Waqar.ashraf.21@ucl.ac.uk

✉ Vivek Dua
v.dua@ucl.ac.uk

[1] The Sargent Centre for Process Systems Engineering, Department of Chemical Engineering, University College London, Torrington Place, London WC1E 7JE, UK

Published online: 05 July 2024

$\widehat{q}_{1-\alpha}(\mathrm{E}_{\mathrm{epoch}})$	Quantile value on $1-\alpha$ from E_{epoch}
s	Number of input layer neurons
V	Velocity matrix
W_1	Weight connections from the input to hidden layer neurons
W_2	Weight connections from the hidden to output layer neuron
X	Set of input variables
Z	Model-simulated value

Greek letters

$1-\alpha$	Confidence level
η	Learning rate
β	Momentum parameter

1 Introduction

In the current era, when machine learning-based models are increasingly applied for modelling the systems with varying complexity and design space of the input variables [1, 2], it becomes equally important to estimate the uncertainty/prediction intervals associated with the model-based predictions. In most of the recent popular applications domain of machine learning like natural language processing and computer vision, the literature is heavily focused on classification based problems [3, 4] and the problems analysed by regression are less reported in these published articles. Many problems associated with the real-time operation optimization of industrial systems like oil refineries, chemical processing plants, energy systems, etc., can be effectively handled with regression-based techniques [5–11]. The policy makers and system engineers can incorporate the range of variability in the machine learning model-simulated responses, while making the effective operational strategies and informed decision-making, to enhance the operation excellence of the industrial systems [12, 13].

The techniques used to draw the prediction intervals for classification and regression-based problems differ significantly. The probability estimation as predicted by the classification methods serve as the starting point for making the prediction intervals. On the other hand, the point-predictor methods, which provide one summary statistic for the conditional distribution, are conventionally used for regression-based problems. The potential disadvantage of the point-predictors is the lack of the information about the confidence the method may express in the predictions. The prediction intervals can be constructed by modelling the conditional distribution via Bayesian method [14–17] or ensemble method [18]. There is an alternative paradigm for estimating the prediction intervals using direct interval estimation method [19] or conformal prediction intervals [20] that does not require to model the conditional distribution.

1.1 Literature review

The Bayesian method attempts to model the conditional distribution by a prior distribution, the available data and the likelihood function. The prior estimate is updated using Bayes' rule and posterior distribution is computed. Gaussian process model works efficiently with Bayesian method and may incorporate the domain knowledge into the prior distribution to characterize the distribution underlying the data generating process that is helpful to compute the prediction intervals [15, 21]. However, the Gaussian process model assumes that data distribution is Gaussian and the reliability of the prediction intervals made by the Gaussian process model can be unreliable if the assumption is not valid on the given data. Ensemble method is another popular method that trains multiple machine learning models and aggregates their predictions to estimate the mean prediction against the input conditions [22, 23]. The ensemble method can be considered as an approximation of Bayesian method where each trained machine learning model represents a sample in the parameters space. Thus, the prediction made by the ensemble method contains the notion of uncertainty that lacks the probabilistic interpretation.

Conformal prediction is another class of prediction bound construction methods and it can estimate the prediction intervals upon providing the dataset and the non-conformity measure [24–27]. The transductive conformal prediction (TCP) method is computationally intensive since the training of an underlying model in the data must be redone for every data point [28]. Furthermore, the error associated during the model training is also stacked up which leads to error propagation in the prediction intervals. To this end, inductive conformal prediction (ICP) method is introduced that decouples the training of 'conformalization' phase [29]. However, ICP exhibits less strong theoretical guarantees than the original *transductive* approach [30]. But, the computational speed is improved for making the prediction intervals by the ICP method [31]. Direct interval estimation method involves training a machine learning model on a loss function tailored to produce the prediction intervals [32]. Since this technique is specifically designed to produce predictions bounds, thus, it is anticipated to perform comparatively better than the modified point estimators. The potential disadvantage of direct interval estimation method is to incorporate a pre-defined confidence level in the loss function and producing prediction intervals on different confidence level requires to retrain the model [33].

The four commonly used prediction bound construction techniques as discussed above have their own merits as well as drawbacks depending upon their working principles and the type of the dataset. Direct interval estimation method directly targets the construction of prediction intervals and the technique can be applied for the given data-driven application. However, the loss function of model being trained can be modified and the procedure for drawing the prediction intervals for the particular confidence level can be updated in order to overcome the limitation of this technique for its widespread utilization for the real-life applications.

1.2 Contribution of this research work

In this work, we present a novel data-driven prediction intervals construction technique called Storage of Weights And Retrieval Method (SWARM) that is inspired by the direct interval estimation method and is hybridized with ICP technique to leverage the power of the two techniques for constructing the accurate prediction intervals. The proposed loss function of the ANN model in the SWARM approach consists of the least mean square error and standard deviation between the actual and model-simulated response—the new loss function is written differently as traditionally specified for the direct interval estimation method [34]. However, the loss function is minimized considering the minimization of standard deviation between the actual and model-simulated responses through the online training for the ANN model development. It allows to minimize the loss function for each observation of the output variable as opposed to the batch training mode, and the parameters (weights and biases) stored in each iteration are deployed for the construction of the prediction intervals where confidence level can be selected by the user (generally 95%), which is a different approach to compute the prediction intervals compared from the traditional direct interval estimation technique.

The SWARM is applied on three energy systems-based case studies taken from the literature (crease recovery angle for resin finishing [35]), benchmark datasets (energy efficiency cooling and energy efficiency heating of residential buildings [36]) and the industrial applications (power generation from gas turbine power plant and coal power plant). Furthermore, the width of the prediction intervals computed by the SWARM on the case studies is compared with those computed by the traditional ICP technique; please note that ICP is computationally inexpensive technique compared with TCP yet producing valid prediction intervals with reasonably high theoretical guarantees [30]. The SWARM approach utilizes the model-based information for the construction of the prediction intervals that is computationally inexpensive and eliminates the need to design additional experiments for the construction of prediction intervals. Thus, the training algorithm of the ANN model can be supplemented to estimate the prediction intervals around the model-simulated responses by the SWARM once the ANN model has been trained. The proposed SWARM technique leverages the power of ICP to compute the valid prediction intervals by its hybridization with the ANN parameters and can be applied for real-life applications involving ANN used for the modelling tasks.

2 Methodology

In this paper, we have proposed the SWARM that hybridizes the neural network with the inductive conformal prediction technique to construct the prediction intervals around the model-simulated responses. We have also compared the width of the prediction intervals and the coverage ratio for the SWARM and inductive conormal prediction (ICP) techniques for the considered case studies. More details about the procedure of prediction interval construction by the SWARM and ICP techniques are described in the following section.

2.1 ANN model training and computing the prediction intervals by the SWARM approach

Let X be the matrix of input variables, say $X = [X_1, X_2, \ldots, X_s]$ having the dimension of s by N where s is the number of inputs, N is the total number of observations in X and is deployed to construct the functional map by ANN with an output variable D that has the dimension of 1 by N. The input layer of ANN has neurons equal to the elements in X. W_1 is the weight matrix containing the weight links from the input to hidden layer of neural network having the size 'm', thus, W_1 is a matrix of m by s, while W_2 is the matrix containing the weight links from the hidden layer to output layer neuron having the dimension of 1 by m. Both X and D have the same number of observations associated with them. The observations of the input X are fed to the input layer of ANN as shown on Fig. 1 and the following computations take place:

$$p_1 = \sum W_1 \odot X^T + b_1 \tag{1}$$

$$y_1 = f_1(p_1) \tag{2}$$

The bias assigned at the hidden layer neurons is enclosed in b_1 matrix with the dimension of m by 1; $W_1 \odot X$ is the elemental wise multiplication of X with the relevant weight links and $\sum W_1 \odot X + b_1$ is calculated at the hidden layer neurons in the hidden layer; f_1 is the activation function that transforms $\sum W_1 \odot X + b_1$ calculated at the hidden layer neurons (p_1) in to the scale that depends upon the type of activation function. In this work, we have implemented the activation function of tangent hyperbolic on the hidden layer

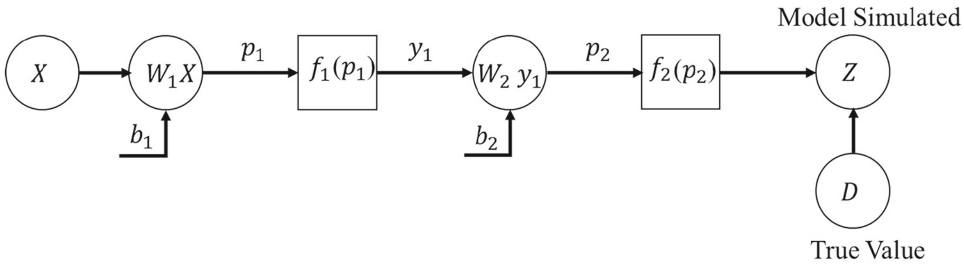

Fig. 1 The schematic depicting the flow of information and processing at different computational nodes embedded in the architecture of artificial neural network

that scales p_1 onto -1 to 1. The scaled observations at the hidden layer neurons are stored in y_1 which are fed to the neuron in the output layer and the model-simulated responses are given as:

$$p_2 = \sum W_2 \odot y_1 + b_2 \tag{3}$$

$$Z = f_2(p_2) \tag{4}$$

Here, b_2 is the bias on the output layer's neuron and it has the dimension of 1 by 1. $W_2 \odot y_1$ performs the element-wise multiplication between W_2 and y_1. The summation ($p_2 = \sum W_2 \odot y_1 + b_2$) computed on the neuron of the output layer undergoes the scale-transformation by f_2 that produces the model-simulated response (Z) calculated at the output layer of ANN. In this work, f_2 is the linear activation function that is applied on the output layer. The online-training method is used for the training of ANN model where an input vector containing one observation of each input variable is fed, the whole training dataset is utilized via sequential approach in one epoch and the parameters (weights and biases) are updated under the specified epochs. Furthermore, the online-training method is suitable for computing the prediction intervals around the model-simulated responses so that information associated with each observation in an epoch during the model training can be utilized to construct the prediction intervals.

The new loss function customized in this work is the least mean square of error and the standard deviation between the model-simulated and actual responses that is to be minimized in each training epoch. The constructed loss function is different from the standard ANN model-based loss function that generally consists of a single performance metric depending upon the type of parameter optimization algorithm and nature of the problem. The standard deviation term measures the spread of the observations around the mean value and can be minimized to achieve good simulation performance of the model. Thus, standard deviation term is made the ingredient of the new loss function which is to be minimized under the online-training mode of the ANN model development. In the online-training mode, the training data is passed on as one input vector for the parameters update and the whole training dataset is fed for the ANN model development in the sequential approach [37]. Thus, the standard deviation term is reduced to $\frac{|\mathrm{D}-Z|}{\sqrt{2}}$ (minimization of the residual between D and Z) which is written in the new loss function as follows:

$$\mathcal{L} = \frac{(\mathrm{D} - Z)^2}{2} + \frac{|\mathrm{D} - Z|}{\sqrt{2}} \tag{5}$$

The parameters, i.e. weight and bias values (W_1, W_2, b_1 and b_2) of the ANN, are optimized by gradient descent with momentum algorithm since it has stable and fast convergence compared to gradient descent algorithm [38]. Furthermore, the algorithm requires less computational efforts for the efficient parametric optimization to achieve the good simulation performance of the trained ANN model. The parameters are updated in the iterative training as governed by the specified loss function that directs to minimize the least mean square of error and the standard deviation between the model-simulated and actual responses. Thus, the tailored loss function including the standard deviation term allows to store the parameters information corresponding to each data observation in the online mode of model training that is relevant to construct the prediction intervals. The analytical expressions for the parameters update for ANN algorithm are derived considering the new loss function tailored for the construction of the prediction interval which are different than those of conventional feedforward ANN algorithm. The partial derivative of the tailored loss function is taken with respect to the parameter and the computed error signal is transmitted backward to produce an update in the parametric values during the training of the model such that the loss function is minimized.

Let us consider the update for weight links (W_1) from input to hidden layer. The new value of W_1 updated by gradient descent with momentum algorithm is given as:

$$W_1^{\mathrm{new}} = W_1 - \eta V_{\mathrm{W}_1} \tag{6}$$

where η is the learning parameter and V_{W_1} is the velocity matrix that is defined as [38]:

$$V_{\mathrm{W}_1} = \beta V_{\mathrm{W}_1} + (1 - \beta)\frac{\partial L}{\partial W_1} \tag{7}$$

The momentum parameter is denoted by β and V_{W_1} is zero matrix with the dimension as of W_1. $\frac{\partial L}{\partial W_1}$ is the partial derivative of $\mathcal{L}$ with respect W_1 that is derived by chain rule which is given as:

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial Z}\frac{\partial Z}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial p_1}\frac{\partial p_1}{\partial W_1} \tag{8}$$

$$\frac{\partial L}{\partial Z} = \frac{\partial}{\partial Z}\left(\frac{(\mathrm{D} - \mathrm{Z})^2}{2} + \frac{|\mathrm{D} - \mathrm{Z}|}{\sqrt{2}}\right) = -(D - Z) - \frac{(D - Z)}{|D - Z|\sqrt{2}} \tag{9}$$

$$\frac{\partial Z}{\partial p_2} = \frac{\partial p_2}{\partial p_2} = 1 \tag{10}$$

$$\frac{\partial p_2}{\partial y_1} = \frac{\partial}{\partial y_1}(W_2 \odot y_1 + b_2) = W_2 \tag{11}$$

$$\frac{\partial y_1}{\partial p_1} = \frac{\partial}{\partial p_1}\left(\frac{e^{p_1} + e^{-p_1}}{e^{p_1} - e^{-p_1}}\right) = 1 - y_1^2 \tag{12}$$

$$\frac{\partial p_1}{\partial W_1} = \frac{\partial}{\partial W_1}(W_1 \odot X + b_1) = X \tag{13}$$

Plugging Eq. (9–13) in Eq. 8:

$$\frac{\partial L}{\partial W_1} = -\left((D - Z) + \frac{(D - Z)}{|D - Z|\sqrt{2}}\right)W_2^T(1 - y_1^2)X^T \tag{14}$$

Equation 6 can be written with reference to Eqs. 7 and 14 as:

$$W_1^{new} = W_1 + \eta(\beta V_{\mathrm{W}_1} + (1 - \beta)\left(\left((D - Z) + \frac{(D - Z)}{|D - Z|\sqrt{2}}\right) W_2^T(1 - y_1^2)X^T)\right) \tag{15}$$

Similarly, W_2 is updated [38] as follows:

$$W_2^{\mathrm{new}} = W_2 - \eta V_{\mathrm{W}_2} \tag{16}$$

$$V_{\mathrm{W}_2} = \beta V_{\mathrm{W}_2} + (1 - \beta)\frac{\partial L}{\partial W_2} \tag{17}$$

Here, V_{W_2} is defined on the dimensions of W_2 as a zero matrix. The expression for $\frac{\partial L}{\partial W_2}$ is derived by chain rule as:

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial Z}\frac{\partial Z}{\partial p_2}\frac{\partial p_2}{\partial W_2} = -\left((D - Z) + \frac{(D - Z)}{|D - Z|\sqrt{2}}\right)y_1 \tag{18}$$

From Eqs. 17 and 18, Eq. 16 can be written as:

$$W_2^{\mathrm{new}} = W_2 + \eta(\beta V_{\mathrm{W}_2} + (1 - \beta))\left((D - Z) + \frac{(D - Z)}{|D - Z|\sqrt{2}})y_1^T\right) \tag{19}$$

Similarly, the update in b_1 and b_2 is given as:

$$b_1^{\mathrm{new}} = b_1 + \eta(\beta V_{\mathrm{b}_1} + (1 - \beta))\left(\left((D - Z) + \frac{(D - Z)}{|D - Z|\sqrt{2}}\right) W_2^T(1 - y_1^2)\right) \tag{20}$$

$$b_2^{\mathrm{new}} = b_2 + \eta(\beta V_{\mathrm{b}_2} + (1 - \beta))\left((D - Z) + \frac{(D - Z)}{|D - Z|\sqrt{2}}\right) \tag{21}$$

here, V_{b_1} and V_{b_2} are the zero matrices having the same dimensions as that of b_1 and b_2, respectively. Thus, the model-simulated response (Z) corresponding to the input vector can have positive or negative deviation from the true value (D) for the specified epochs. However, the difference between Z and D continues to decrease as the training of the ANN model progresses and the ANN model develops good predictive performance with the update in weights and biases introduced in the architecture of the ANN model. Moreover, gradient descent with momentum algorithm drives the smooth update in the parameters to achieve their optimal values. Generally, a few hundreds or thousands of epochs are executed for the model development depending upon the nonlinear characteristics of the output variable, the updated weights and bias values are stored corresponding to each epoch and are deployed for the construction of prediction intervals around each observation of the output variable. The values of Z with respect to input vector and the parameters (weights and biases) are simulated and the procedure is repeated for all the input vectors. This procedure allows to utilize the stored parameters to simulate observations of the output variable corresponding to the given input vector and the simulated observations of output variable with respect to the one input vector incorporate the unique information of the data distribution when the model is being trained. Thus, here we hybridize the conformal prediction technique with the online-training mode of neural network to construct the prediction bound with the locally adaptive information available for the output variable. In this work, the model stops the training upon meeting the either condition of the stopping criteria, i.e. loss function value on testing dataset is equal to zero, the gradient is less than 0.000000001 or maximum number of epochs are executed.

2.2 Prediction intervals estimation by the inductive conformal prediction technique

Let us consider that the dataset is sorted on the input vector and model-simulated responses as: (X_1, Z_1), (X_2, Z_2), …, (X_N, Z_N). We consider the example of one data observation (X_1, Z_1) to demonstrate the construction of the prediction intervals around Z_1 by SWARM technique, and then, the procedure is extended on the remaining data points of Z.

Since the online mode of neural network training is implemented, the simulated observations of Z_1 corresponding to each epoch are available which is denoted as $(Z_1)_{\text{epoch}}$. The non-conformity measure is taken as absolute difference between D_1 and $(Z_1)_{\text{epoch}}$, commonly used for regression problems [30], and is written as:

$$\mathrm{E}_{1\text{epoch}} = \left| D_1 - (Z_1)_{\text{epoch}} \right| \tag{22}$$

The quantile value on $1 - \alpha$ confidence level is computed from $(\mathrm{E}_1)_{\text{epoch}}$ as:

$$\widehat{q}_{1-\alpha}\left(\mathrm{E}_{1\text{epoch}}\right) = (1-\alpha)\frac{\text{epoch} + 1}{\text{epoch}} \tag{23}$$

here, $\widehat{q}_{1-\alpha}$ depicts the quantile value of $(\mathrm{E}_1)_{\text{epoch}}$ on $1 - \alpha$ confidence level which is used to compute the prediction interval (PI) around Z_1 as:

$$PI(Z_1)_{\text{SWARM}} = [Z_1 - \widehat{q}_{1-\alpha}\left(\mathrm{E}_{1\text{epoch}}\right),\ Z_1 + \widehat{q}_{1-\alpha}\left(\mathrm{E}_{1\text{epoch}}\right)] \tag{24}$$

Using the same procedure, the prediction interval around the remaining observations of Z can be computed.

The key difference on the computation of the prediction interval by SWARM technique and the traditional ICP is the locally adaptive prediction interval construction around each observation of Z made by SWARM while the width of the prediction interval remains fixed for ICP for the dataset [39]. In ICP technique, the dataset is split into training, testing and calibration dataset. The model is trained on training and testing dataset while calibration dataset is used for the computing of $(\mathrm{E}_1)_{\text{cal}}$ and $\widehat{q}_{1-\alpha}$ which are given as:

$$\mathrm{E}_{\text{cal}} = |D_{\text{cal}} - Z_{\text{cal}}| \tag{25}$$

$$\widehat{q}_{1-\alpha}(\mathrm{E}_{\text{cal}}) = (1-\alpha)\frac{n_{\text{cal}} + 1}{n_{\text{cal}}} \tag{26}$$

here, n_{cal} is the size of the calibration dataset. The prediction interval on Z for test dataset (Z_{test}) made by ICP technique is given as:

$$PI(Z_{\text{test}})_{\text{ICP}} = [Z_{\text{test}} - \widehat{q}_{1-\alpha}(\mathrm{E}_{\text{cal}}),\ Z_{\text{test}} + \widehat{q}_{1-\alpha}(\mathrm{E}_{\text{cal}})] \tag{27}$$

The prediction interval computed by SWARM and ICP techniques is compared for the considered case studies. The details can be found in the following sections.

2.3 Evaluation criteria

The predictive efficiency of the ANN model for training and testing dataset is computed by two rigorous statistical terms, i.e. coefficient of determination (R^2) and root-mean-squared error (RMSE). Mathematically, R^2 and RMSE are represented as follows:

$$R^2 = 1 - \frac{\sum_i^N (Z_i - \mathrm{D}_i)^2}{\sum_i^N (\mathrm{D}_i - \overline{D}_i)^2} \tag{28}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z_i - \mathrm{D}_i)^2} \tag{29}$$

here, D_i and Z_i are the actual and model-driven simulated responses for $i = 1, 2, 3, \ldots \mathrm{N}$. R^2 quantifies the predictive efficiency of the model and it varies from zero to one. $R^2 = 0$ signifies the poor predictive performance while $R^2 = 1$ indicates that model-simulated and true observations are close to each other. However, RMSE measures the error associated in the model-simulated responses and is to be minimized to achieve the good predictive performance of the trained ANN model.

3 Results

The SWARM approach proposed in this paper constructs the prediction intervals around the ANN model-based simulated responses with 95% confidence level and the methodology is implemented on the datasets taken from the literature, benchmark datasets and the industrial systems. The dataset for crease recovery angle is taken from the literature [35], whereas energy efficiency cooling (ENC) & energy efficiency heating (ENH) dataset for the buildings is taken from UC Irvine ML database repository [40]. Furthermore, dataset for power production from a 395 MW capacity gas turbine and a 660 MW capacity supercritical coal power plant is also taken to draw the prediction intervals around the model-simulated responses for the industrial-scale applications. The prediction intervals for the considered case studies are also constructed by ICP to compare the width of the prediction intervals computed by the SWARM with those of the ICP technique. This provides the comparative analysis on the estimation of the prediction intervals around the model-simulated responses by the two techniques as well as comparing the SWARM-based results with the existing benchmark technique in the literature.

The dataset for the considered case studies is split into training, testing and calibration dataset on the split ratio of 0.7, 0.15 and 0.15, respectively. While training and testing dataset is primarily used for the ANN model development, the calibration dataset serves to compute the prediction intervals for the test dataset by the ICP technique [41]. Since SWARM is a locally adaptive technique for the prediction

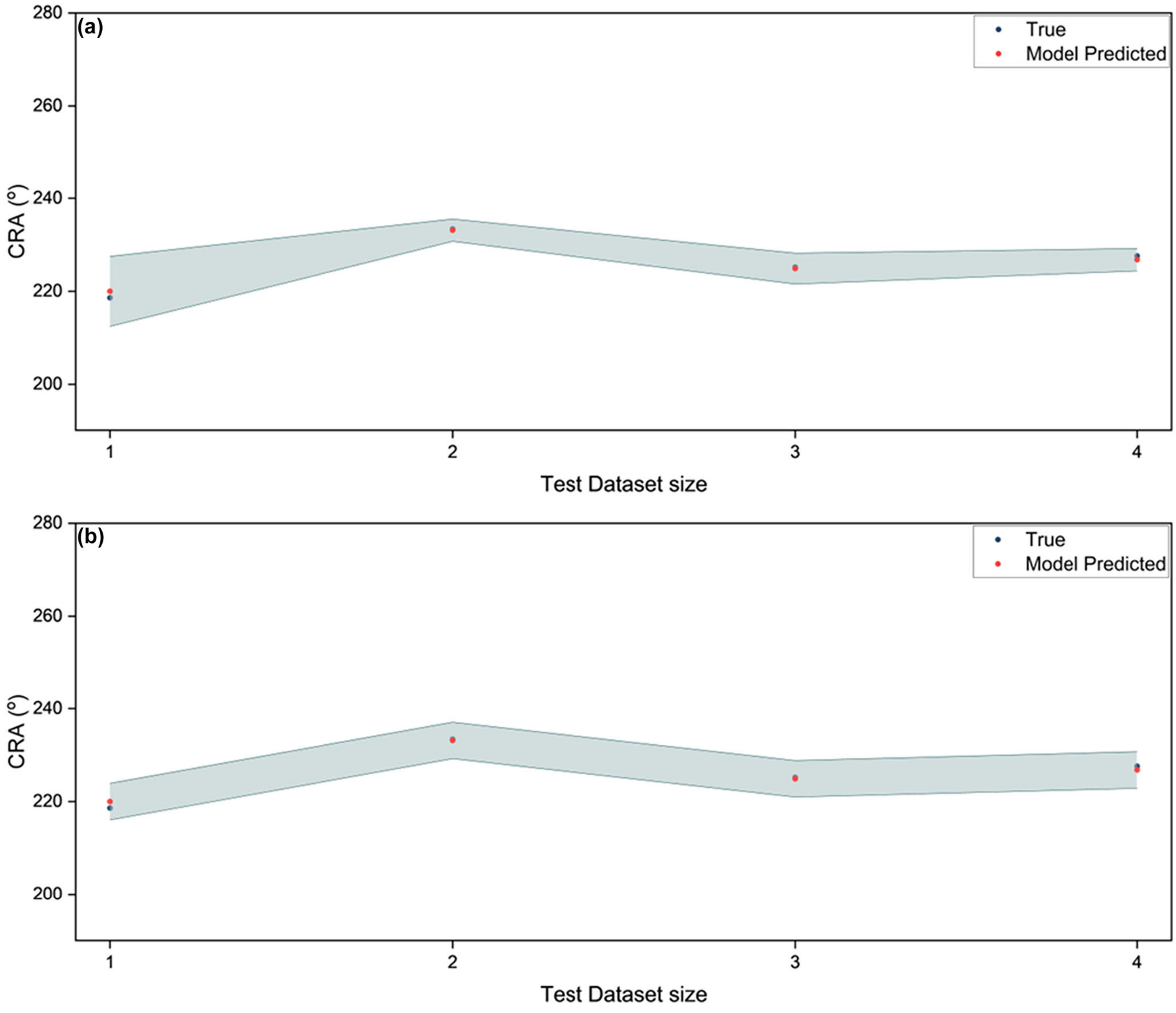

Fig. 2 Computing the prediction intervals for CRA by (**a**) SWARM and (**b**) ICP techniques. The prediction intervals are computed on the test dataset with 95% confidence level

interval estimation, we have computed the prediction intervals for the test dataset by the SWARM technique and have compared the results with those of ICP technique for the same test dataset for case studies. The analysis presented in the paper is conducted in MATLAB 2019 b software installed on a system having a processor of 11th Generation Intel Core i7 1185G7 @ 3.0 GHz 1.8 GHz and 32 GB installed RAM.

3.1 Case study-1: crease recovery angle

The crease recovery angle (CRA) is modelled by five input variables, namely resin, polyethylene softener, catalyst, curing temperature and curing time. The dataset taken from the literature [35] consists of 27 experiments designed on different levels of input variables.

During the training of ANN model, learning rate and momentum coefficient are taken as 0.01 and 0.9, respectively. Tangent hyperbolic and linear activation functions are applied on the hidden and output layer of ANN, respectively. The hidden layer neurons are taken as 10 which are reasonably large number to approximate the function space. The predictive performance of the model on the training and testing dataset is computed which is as follows: R^2_train = 0.98, RMSE_train = 1.1 (°), R^2_test = 0.99, RMSE_test = 0.83 (°).

Having trained the ANN model for CRA, the prediction intervals are computed by SWARM and ICP techniques and are presented on Fig. 2. Comparing the prediction intervals computed by the two techniques, it is evident that SWARM-based prediction intervals for test dataset observation number

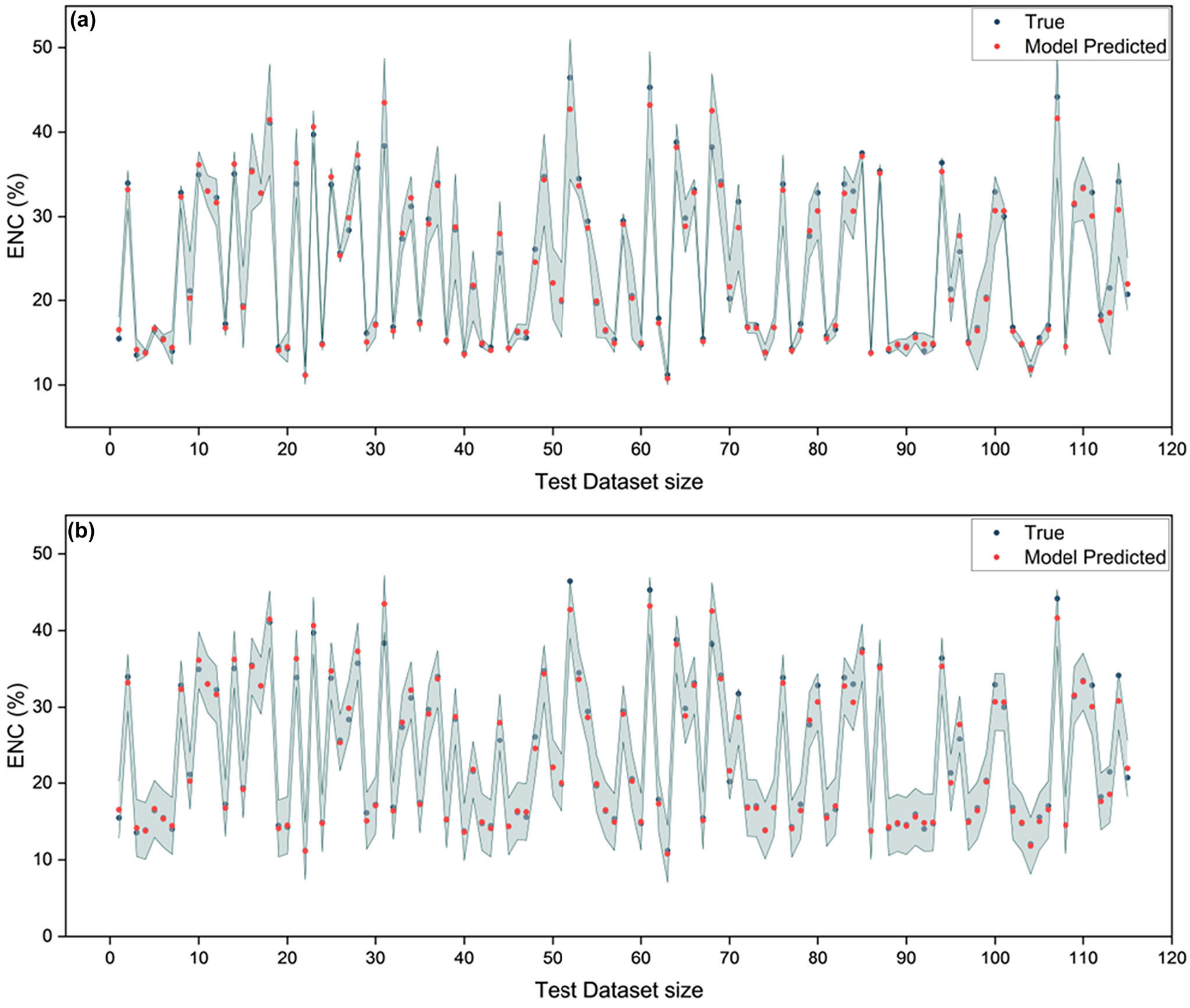

Fig. 3 The prediction intervals estimation for ENC dataset by (**a**) SWARM and (**b**) ICP techniques. The ICP technique-based prediction intervals seem to be more spread out than those of SWARM, yet providing full data coverage on the test dataset

2, 3 and 4 are comparatively tighter yet the true observations lying within the prediction intervals of ICP, thereby validating the efficacy of the estimated prediction intervals made by the SWARM. However, the prediction interval for observation 1 as made by SWARM is relatively spread out which can be attributed to the modelling inaccuracy of model in the localized region. Moreover, full data coverage is observed for SWARM and ICP, indicating the good performance of the techniques to construct the prediction intervals for the CRA application.

3.2 Case study-2: energy efficiency cooling & energy efficiency heating

Energy efficiency cooling (ENC) and energy efficiency heating (ENH) are the two performance indicators of the residential buildings and are modelled on orientation, surface area, relative compactness, roof area, overall height, wall area, glazing area and glazing area distribution [36]. The compiled dataset for the input and output variables has 768 observations that are deployed to train the ANN-based models for ENC and ENH separately.

In the literature [42], a feedforward neural network model is constructed for ENC on ten number of hidden layer neurons; tangent hyperbolic and linear activation function are applied on hidden and output layer of the model, respectively. Thus, we also apply the same settings to initialize the ANN architecture incorporating the loss function as proposed in this study in Eq. 5 and have trained the ANN model. The predictive performance of the ANN model trained in this work is as follows: R^2_train = 0.99, RMSE_train = 1.1%, R^2_test = 0.98, RMSE_test = 1.2%.

The prediction intervals on the test dataset of ENC with 95% confidence level are computed by the SWARM and ICP techniques and are presented on Fig. 3. The SWARM technique-based prediction intervals around some observations, as shown on Fig. 3a, are quite tight that represent the reduced modelling inaccuracy of the trained ANN in the localized regions. However, the ICP technique-based prediction intervals seem to be comparatively more spread out, as shown on Fig. 3b, than those of SWARM. However, ICP-based computed prediction intervals offer full data coverage, demonstrating the accuracy of the technique to predict the test observation with good estimate of prediction intervals.

Similarly, the ANN model for ENH is trained using the same set of initializations as established for ENC-based ANN model. The predictive performance of the ANN model for the training and testing dataset is computed as: R^2_train = 0.99, RMSE_train = 0.44%, R^2_test = 0.99, RMSE_test = 0.52%. The computed prediction intervals on 95% confidence level by the SWARM and ICP techniques for ENH are depicted on Fig. 4. Nearly the same width of the prediction intervals is observable for ENH as computed by the two techniques for most of the test observations. We also observe the full data coverage on ICP technique-based prediction intervals, thereby validating their accuracy. A few test observations have comparatively large width of the prediction intervals as computed by the SWARM with those of the ICP technique. This is attributed to the deviation of the true test values with those of the model-simulated responses, resulting in the computation of the prediction intervals which are adaptive to locally available information on the model-simulated responses.

3.3 Case study-3: power generation from gas turbine & coal power plant

The power generation from a gas turbine power plant is modelled on flow rate of fuel gas, air temperature at the outlet of compressor, air pressure at the outlet of compressor, fuel gas temperature at the outlet of performance heater, ambient pressure, ambient temperature and ambient humidity. A total of 578 observations compiling the dataset for input–output variables is deployed for the development of ANN model. The initialization settings for ANN model training are the same as discussed in case study 2 except the hidden layer neurons are kept at 16. The modelling performance of ANN for the training and testing dataset is as follows: R^2_train = 0.99, RMSE_train = 1.42 MW, R^2_test = 0.99, RMSE_test = 1.94 MW.

The prediction intervals are calculated for the gas turbine power on 95% confidence interval with the SWARM and ICP techniques and are shown on Fig. 5. Overall, the two techniques present the comparable width of the prediction intervals for the test dataset of the gas turbine power. We observe nearly full data coverage by the ICP technique with only one observation lying outside the estimated prediction intervals. However, overall, the estimated prediction intervals by the ICP technique are quite tight with good data coverage. The comparable width of the prediction intervals, as visualized on Fig. 5a, b for SWARM and ICP techniques, respectively, demonstrates the accuracy of the SWARM technique to estimate the prediction intervals with good accuracy for the gas turbine power as they have been compared with the benchmarked ICP technique.

Similarly, power generation from a coal power plant is modelled by input variables, namely coal flow rate, total air flow rate, main steam pressure, main steam temperature, main steam flowrate, feed water temperature, reheat steam temperature and condenser vacuum. A total of 639 observations are taken for the input and output variables and same initialization settings as those of gas power plant-based ANN model are applied for the training of ANN model for power generation from coal power plant except hidden layer neurons are taken as 10. The performance metrics for the trained ANN are as follows: R^2_train = 0.99, RMSE_train = 2.64 MW, R^2_test = 0.99, RMSE_test = 2.20 MW.

The computed prediction intervals by the SWARM and ICP techniques for coal power on test dataset at 95% confidence interval are shown on Fig. 6a, b, respectively. Nearly, comparable width of the prediction intervals is observed for the coal power by the SWARM and ICP techniques. Furthermore, full data coverage is achieved for the test dataset by ICP technique. The comparable width of the prediction intervals demonstrates the usefulness of exploiting the parameters (weight and biases) information for drawing the prediction intervals by the SWARM approach.

4 Discussion

4.1 Visualization of parameters and model-simulated responses distribution

In the previous section, we have presented the results on drawing the prediction intervals by using the information stored for the parameters (weights and biases) of ANN model and have compared the width of prediction intervals for testing datasets with those of the ICP method. Here, we provide the data-distribution profiles of some of the parameters taken from the trained ANN model for gas turbine power as shown on Fig. 7. The weight links from hidden to output layer (HLN) are updated during the iterative training of the model. The weight connections for HLN are visualized since linear activation function is applied on the output layer, and thus, weight connections of HLN have direct impact on the computation of the model-simulated responses. Thus, the impact of the update in the HLN-weight connections

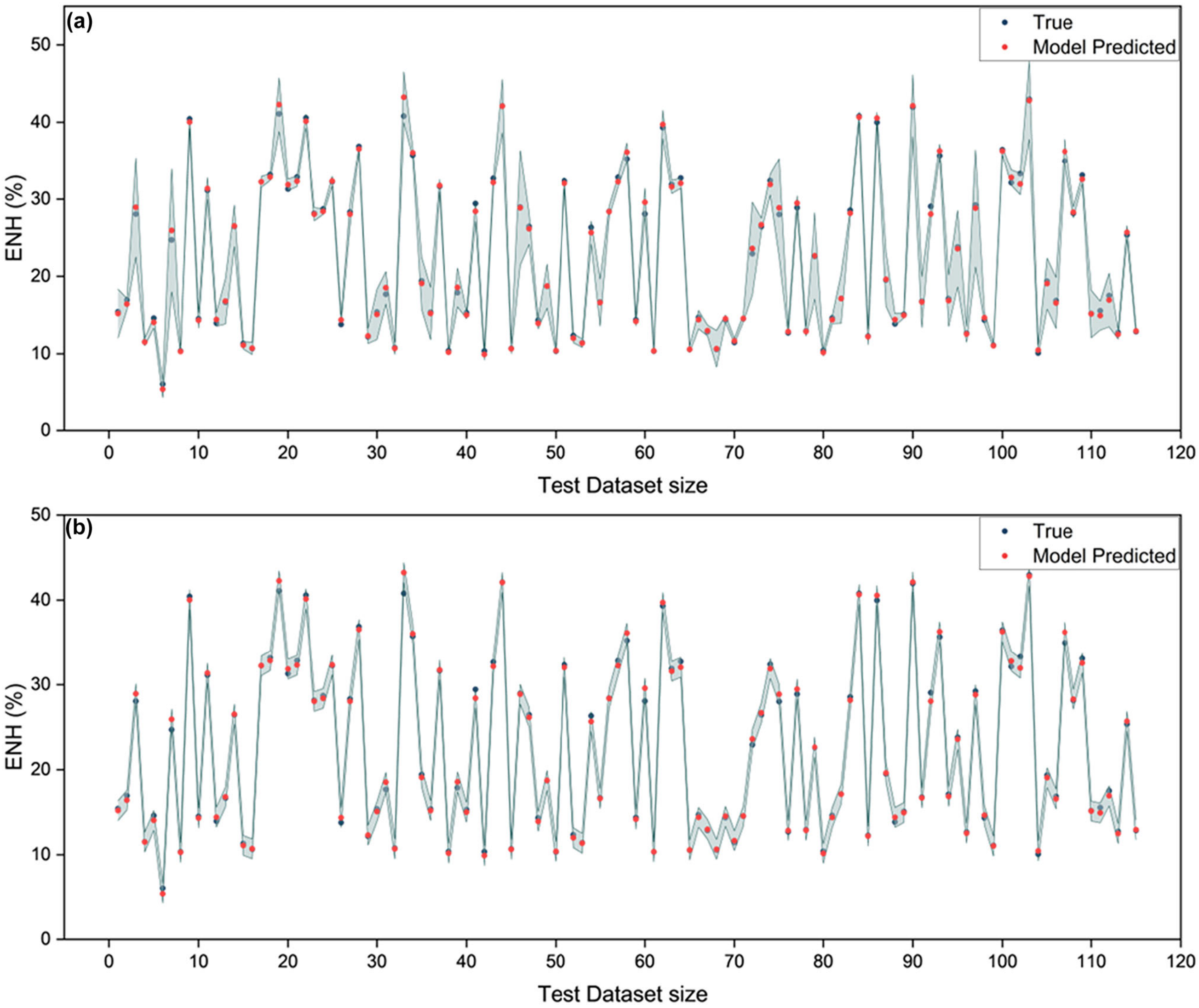

Fig. 4 Estimation of prediction intervals with 95% confidence level on test dataset for ENH by (**a**) SWARM and (**b**) ICP techniques. The ICP technique-driven prediction intervals are quite tight, yet having full data coverage for the test observations

on the computation of model-simulated responses can be visualized which is presented on Fig. 7a. Initially, the slope of weight-epoch graph is high until the epochs are around 200 for gas turbine power, and then, it starts stabilizing for most of the weight connections. To further confirm the distribution profiles of HLN-weight connections of the trained ANN model for gas turbine power, HLN-1, 6, 10, 14 are selected, and their weight-connection distributions are presented on Fig. 7b. We observe that weights distribution profile is asymmetric yet following almost log-normal distribution profile for the weight connections. The similar observation is presented in literature that connection weights in the neural network follow log-normal distribution which can help design the schemes for the efficient and cost-effective training of neural networks [43].

We have taken two observations from the test dataset of gas turbine power and the iterative variation in the model-predicted responses is depicted on Fig. 7c, d for true value of 364 and 290 MW, respectively. Referring to Fig. 7c, the model-predicted responses start from 423 MW and are improved during the iterative training of the model after the parametric update. The distribution of model-predicted responses during the model training is depicted where the model estimated the value of 373 MW corresponding to the true value of 364 MW upon stopping of its training. It depicts that model training was localized in a region where model-predicted responses are far away from the true value to be simulated, and thus, a significant deviation between the true and model-simulated responses exists. The significant deviation between the true and model-predicted responses directs to incorporate the locally available information to construct

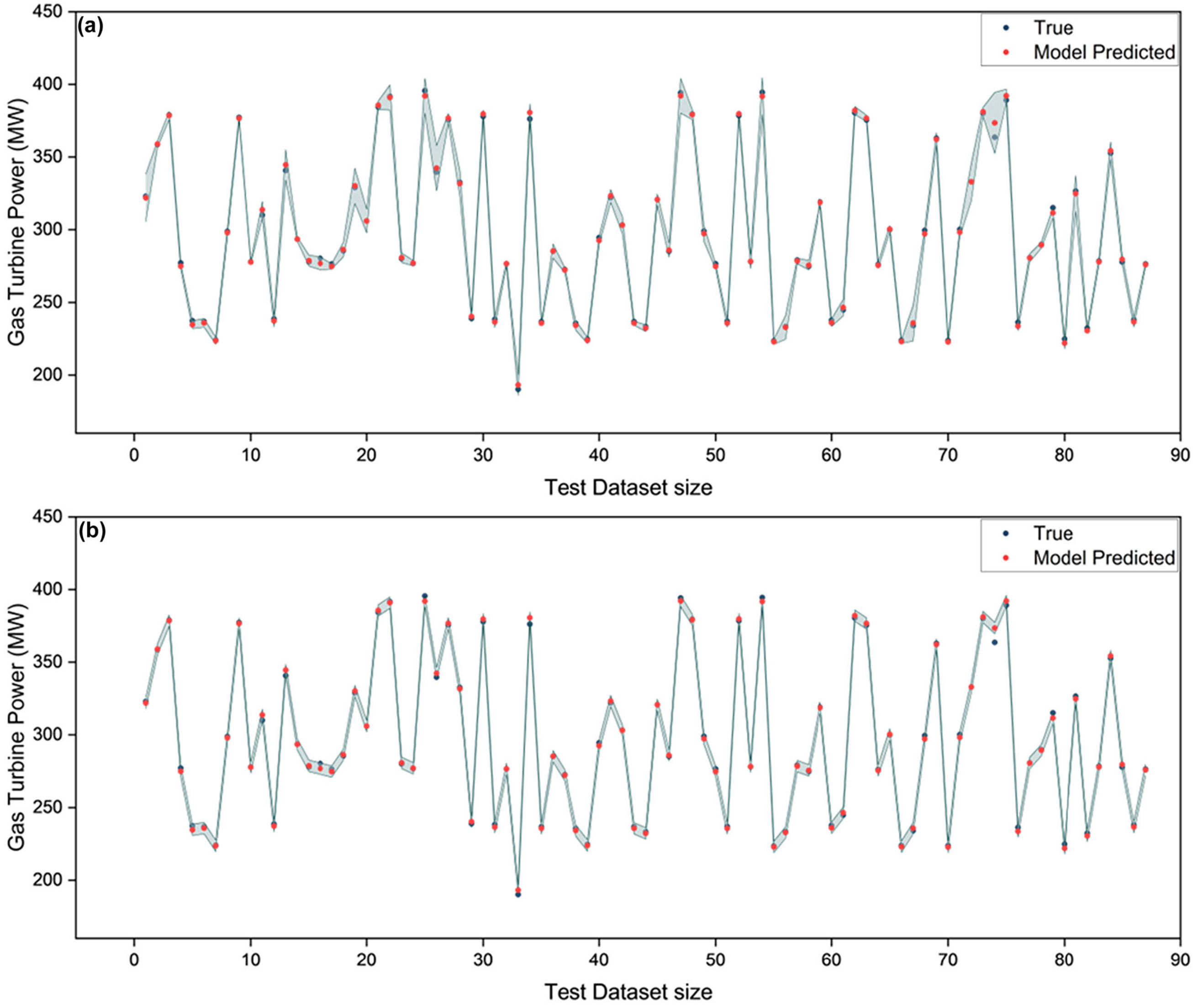

Fig. 5 Comparison of the prediction intervals for gas turbine power computed on the test dataset by (**a**) SWARM and (**b**) ICP techniques on 95% confidence interval

the prediction interval around the model-simulated response. Thus, the SWARM technique calculates $\widehat{q}_{1-\alpha}\left(\mathrm{E}_{\mathrm{epoch}}\right) = 21$ on 95% confidence interval from the computed non-conformity score corresponding to the true value of 364 MW. However, $\widehat{q}_{1-\alpha}(\mathrm{E}_{\mathrm{cal}}) = 3.94$ is calculated on 95% confidence interval by ICP technique. From Fig. 5, the true value is taken corresponding to test observation number of 74 and it is evident that the SWARM-based prediction interval provides data coverage for the true value, whereas ICP technique-based estimated prediction interval has missed the data coverage for the considered test observation. The missed data coverage can be attributed to fixed width of the prediction interval that can be conservative enough to capture the true value.

Similarly, referring to Fig. 7d, the model-predicted responses start from 249 MW and are iteratively updated. The model achieves the predicted value of 289.4 upon the termination of model training corresponding the true value of 290 MW. The true and model-predicted responses are comparably close to each other, meaning that SWARM-based quantile value is reasonably tight, i.e. $\widehat{q}_{1-\alpha}\left(\mathrm{E}_{\mathrm{epoch}}\right) = 2.1$ to capture the true test value. Similarly, the ICP technique-based $\widehat{q}_{1-\alpha}\left(\mathrm{E}_{\mathrm{epoch}}\right) = 3.94$ is also wide enough to capture the true value. The SWARM-based estimation of prediction intervals incorporates the iterative values of the model-predicted responses to estimate the prediction intervals for each observation that offers unique and locally compliant prediction intervals to construct the prediction intervals. The ICP technique-driven fixed-width prediction intervals can sometimes be wide enough to satisfy the guarantees of the data coverage, but they result in significantly large width of the prediction intervals depicting the high uncertainty in the model-predicted responses [39]. Therefore, adapting the

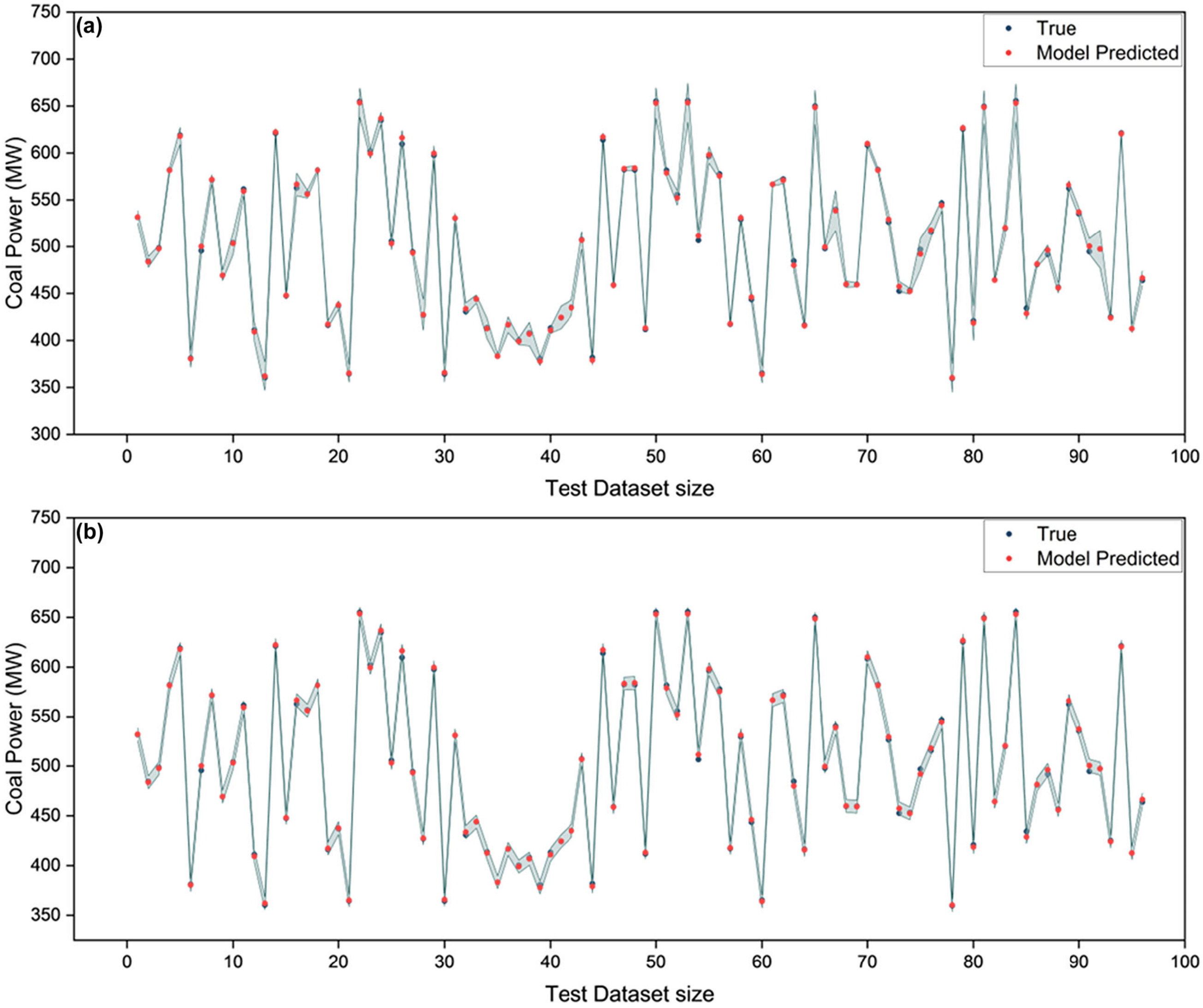

Fig. 6 Prediction intervals comparison for coal power on test dataset at 95% confidence interval by (**a**) SWARM and (**b**) ICP techniques. Comparable width of the prediction intervals is observable as computed by the two techniques

prediction intervals with the locally available information retrieved during the ANN model training can provide the adjustable width of the prediction intervals for the given confidence level for each observation of the test dataset.

5 Conclusions

Constructing the prediction intervals around the model-simulated responses presents the range of variability in the predictions. In this work, we present a novel data-driven approach to construct the prediction intervals around the model-simulated responses using the artificial neural network—a commonly used algorithm of machine learning for function approximation applications. The loss function comprises of the least mean square of error and the standard deviation between the model-simulated and actual responses embedded with the online-training method for the development of the ANN model. The online mode of ANN model development allows to store the parameters information in each epoch during the model development. The SWARM approach is built on hybridizing the ANN parameters with conformal prediction technique to construct the prediction intervals. Three case studies, namely CRA, ENC & ENH for buildings, and gas turbine power & coal power are considered in this work for the construction of prediction intervals by the SWARM approach. Furthermore, the SWARM-based prediction intervals are compared with those of traditional ICP technique.

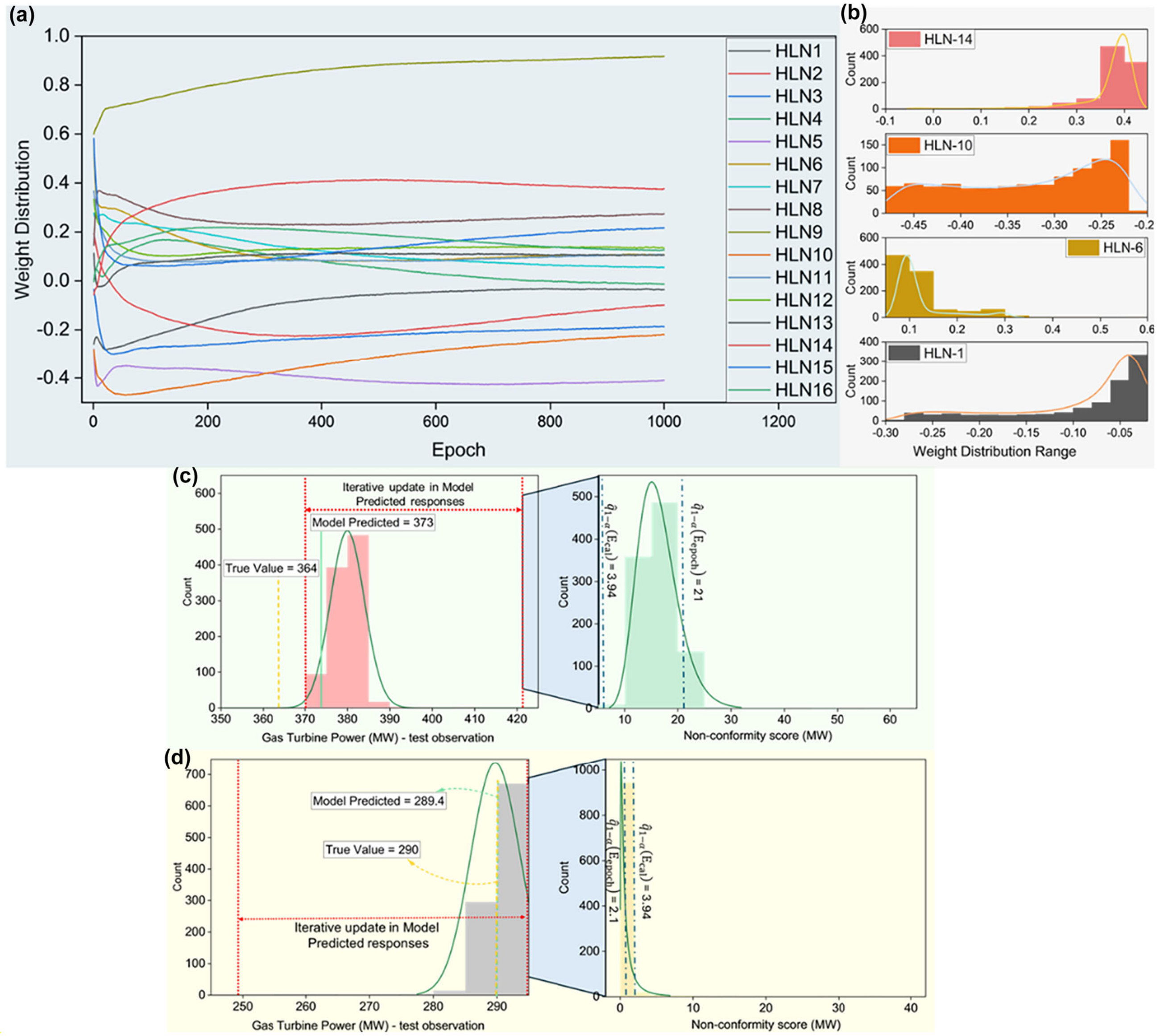

Fig. 7 Visualization of parameters distribution during the model training for gas turbine power. (**a**) The iterative update in the weight connections from the hidden to output layer of ANN shows the smooth improvement in the weight connections. (**b**) The weight-connection distribution profiles for the selected hidden layer neurons show the asymmetric distribution. The distribution of the model-predicted responses during the iterative training along with the true value is depicted on (**c**) and (**d**). Furthermore, the quantile value on 95% confidence level computed by SWARM and ICP techniques is also mentioned. It is important to note that SWARM-based quantile values are locally adaptive, thus subject to variation as opposed to the ICP-based values which remain fixed for the test observations

- The width of the prediction intervals computed by SWARM and ICP techniques is compared for the considered case studies. A comparable width of prediction intervals is noted for CRA, ENH, gas turbine power and coal power, indicating the validity of the SWARM-based prediction intervals as compared with the benchmark ICP technique.
- Slightly large spread out of the prediction intervals is noted for ICP-based prediction intervals for ENC in comparison with those of the SWARM technique. However, ICP constructs fixed-width prediction intervals that are influenced by the predictive accuracy of the model to simulate the calibration dataset.
- The SWARM technique produces adaptive prediction intervals, which are compliant with the locally available information of the observations, enclose the true values and quantify the valid uncertainty in the model-predicted responses.

Author's contribution All the authors have contributed equally to the research paper.

Funding Waqar Muhamad Ashraf got funding from The Punjab Educational Endowment Fund, funding reference number CMMS-PHD-2021-006 to obtain doctoral degree from University College London.

Data availability The code written for the research tasks of this paper can be provided on request.

Declarations

Conflict of interest No conflict of interests is observed and reported for this research.

References

1. Ashraf, W.M., Dua, V.: Artificial intelligence driven smart operation of large industrial complexes supporting the net-zero goal: coal power plants. Digit. Chem. Eng. **8**, 100119 (2023)
2. Ashraf, W.M., Dua, V.: Machine learning based modelling and optimization of post-combustion carbon capture process using MEA supporting carbon neutrality. Digit. Chem. Eng. **8**, 100115 (2023)
3. Teye, M., Azizpour, H. and Smith, K.: Bayesian uncertainty estimation for batch normalized deep networks. In: International Conference on Machine Learning. PMLR (2018)
4. Liu, J. et al.: Accurate uncertainty estimation and decomposition in ensemble learning. Adv. Neural Inform. Process. Syst. **32** (2019)
5. Krzywanski, J., et al.: Modelling of SO2 and NOx emissions from coal and biomass combustion in air-firing, oxyfuel, iG-CLC, and CLOU conditions by fuzzy logic approach. Energies **15**(21), 8095 (2022)
6. Gueddar, T., Dua, V.: Novel model reduction techniques for refinery-wide energy optimisation. Appl. Energy **89**(1), 117–126 (2012)
7. Ashraf, W.M. et al.: Artificial intelligence modeling-based optimization of an industrial-scale steam turbine for moving toward net-zero in the energy sector. ACS Omega (2023)
8. Malakouti, S.M., et al.: Predicting wind power generation using machine learning and CNN-LSTM approaches. Wind Eng. **46**(6), 1853–1869 (2022)
9. Malakouti, S.M., Menhaj, M.B., Suratgar, A.A.: The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction. Clean. Eng. Technol. **15**, 100664 (2023)
10. Malakouti, S.M.: Babysitting hyperparameter optimization and 10-fold-cross-validation to enhance the performance of ML methods in Predicting Wind Speed and Energy Generation. Intell. Syst. Appl. **19**, 200248 (2023)
11. Malakouti, S.M.: Prediction of wind speed and power with LightGBM and grid search: case study based on Scada system in Turkey. Int. J. Energy Prod. **8**(1), 35–40 (2023)
12. Liu, X., et al.: Energy management strategy based on deep reinforcement learning and speed prediction for power-split hybrid electric vehicle with multidimensional continuous control. Energ. Technol. **11**(8), 2300231 (2023)
13. Fang, Z., et al.: Temperature-field sparse-reconstruction of lithium-ion battery pack based on artificial neural network and virtual thermal sensor technology. Energ. Technol. **9**(10), 2100258 (2021)
14. Goan, E., Fookes, C.: Bayesian neural networks: an introduction and survey. Case Stud. Appl. Bayesian Data Sci.: CIRM Jean-Morlet Chair, Fall **2020**, 45–87 (2018)
15. Williams, C. and Rasmussen, C.: Gaussian processes for regression. Advances in neural information processing systems. **8** (1995)
16. Dubey, M., Palakkadavath, R., Srijith, P.: Bayesian neural Hawkes process for event uncertainty prediction. Int. J. Data Sci. Anal (2023). https://doi.org/10.1007/s41060-023-00443-3
17. Wang, X., Kadıoğlu, S.: Modeling uncertainty to improve personalized recommendations via Bayesian deep learning. Int. J. Data Sci. Anal. (2021). https://doi.org/10.1007/s41060-020-00241-1
18. Kendall, A. and Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inform. Process. Syst. **30** (2017)
19. Koenker, R., Hallock, K.F.: Quantile regression. J. Econ. Perspect. **15**(4), 143–156 (2001)
20. Romano, Y., Patterson, E. and Candes, E.: Conformalized quantile regression. Adv. Neural Inform. Process. Syst. **32** (2019)
21. Acharki, N., Bertoncello, A., Garnier, J.: Robust prediction interval estimation for Gaussian processes by cross-validation method. Comput. Stat. Data Anal. **178**, 107597 (2023)
22. Sollich, P. and Krogh, A.: Learning with ensembles: How overfitting can be useful. Adv. Neural Inform. Process. Syst. **8**(1995)
23. Lu, J., et al.: Ensemble stochastic configuration networks for estimating prediction intervals: a simultaneous robust training algorithm and its application. IEEE Trans. Neural Netw. Learn. Syst. **31**(12), 5426–5440 (2020)
24. Gammerman, A., Vovk, V. and Vapnik, V. Learning by transduction. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc.: Madison, Wisconsin. Pp. 148–155 (1998)
25. Saunders, C., Gammerman, A., and Vovk, V.: Transduction with Confidence and Credibility. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc, pp. 722–726 (1999)
26. Vovk, V., Gammerman, A. and Saunders, C.: Machine-learning applications of algorithmic randomness (1999)
27. Angelopoulos, A.N. and Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)
28. Saunders, C., Gammerman, A. and Vovk, V.: Transduction with confidence and credibility (1999)
29. Papadopoulos, H. et al.: Inductive confidence machines for regression. In Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, Aug 19–23, 2002 proceedings 13. Springer (2002)
30. Kato, Y., Tax, D.M. and Loog, M.: A review of nonconformity measures for conformal prediction in regression. Conformal and Probabilistic Prediction with Applications, p. 369–383 (2023)
31. Lei, H., Bellotti, A.: Reliable prediction intervals with directly optimized inductive conformal regression for deep learning. Neural Netw. **168**, 194–205 (2023)
32. Kuleshov, V., Fenner, N. and Ermon, S.: Accurate uncertainties for deep learning using calibrated regression. In: International conference on machine learning. PMLR (2018)

33. Dewolf, N., Baets, B.D., Waegeman, W.: Valid prediction intervals for regression problems. Artif. Intell. Rev. **56**(1), 577–613 (2023)
34. Alcántara, A., Galván, I.M., Aler, R.: Direct estimation of prediction intervals for solar and wind regional energy forecasting with deep neural networks. Eng. Appl. Artif. Intell. **114**, 105128 (2022)
35. Pervez, M.N., et al.: Sustainable fashion: design of the experiment assisted machine learning for the environmental-friendly resin finishing of cotton fabric. Heliyon (2023). https://doi.org/10.1016/j.heliyon.2023.e12883
36. Tsanas, A., Xifara, A.: Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy Build. **49**, 560–567 (2012)
37. Haykin, S.: Neural networks and learning machines, 3/E. Pearson Education India (2009)
38. Ng, A.: Improving deep neural networks: hyperparameter tuning, regularization and optimization. Deeplearning. ai on Coursera (2017)
39. Fontana, M., Zeni, G., Vantini, S.: Conformal prediction: a unified review of theory and new challenges. Bernoulli **29**(1), 1–23 (2023)
40. Tsanas, A.A.X.: Angeliki, Energy efficiency. UCI Machine Learning Repository (2012). https://doi.org/10.24432/C51307
41. Kagita, V.R., et al.: Inductive conformal recommender system. Knowl.-Based Syst. **250**, 109108 (2022)
42. Arnaldo, I., O'Reilly, U.-M. and Veeramachaneni, K.: Building predictive models via feature synthesis. In: Proceedings of the 2015 annual conference on genetic and evolutionary computation (2015)
43. Venkatasubramanian, V., Sanjeevrajan, N. and Khandekar, M.: Jaynes Machine: The universal microstructure of deep neural networks. arXiv preprint arXiv:2310.06960 (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.