Renewable Energy 211 (2023) 412–419

Contents lists available at ScienceDirect

Renewable Energy

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

Wind turbine blade icing diagnosis using RFECV-TSVM pseudo-sample processing

Xinjian Bai [a], Tao Tao [b], Linyue Gao [c], Cheng Tao [a], Yongqian Liu [a,*]

[a] State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources (NCEPU), School of New Energy, North China Electric Power University, Beijing, 102206, China
[b] China Southern Power Grid Technology Co, Ltd, Guangzhou, 510080, China
[c] Department of Mechanical Engineering, University of Colorado Denver, Denver, CO, 80204, USA

ARTICLE INFO

Keywords:
Wind turbine
Icing diagnosis
Supervisory control and data acquisition
Transductive support vector machine
Small-sample data

ABSTRACT

Wind turbine blade icing seriously affects turbine power generation and fatigue life, and an accurate diagnosis of blade icing is beneficial for wind turbines to make in-time adjustments. However, the high dimensional and unbalanced original data recorded by Supervisory Control and Data Acquisition (SCADA) systems pose great challenges to the accurate diagnosis of blade icing. To effectively address the challenges of difficult feature extraction and small number of fault samples, we propose a data processing method based on pseudo-sample processing. Specifically, Recursive Feature Elimination and Cross-Validation (RFECV) is used to analyze the influence of various SCADA features on the diagnostic model and select the most compelling feature set. A Transductive Support Vector Machine (TSVM) is implemented to regenerate unlabelled samples. The labeled pseudo samples and ice data are combined to form the training set. The effectiveness of the proposed method is examined using the three most commonly used classifier algorithms, i.e., Random Forest (RF), Support Vector Machine (SVM), and XGBoost, for four utility-scale wind turbines. The results show that this method can effectively obtain the optimal selection, utilize unlabelled samples, and improve the diagnostic accuracy of the model, especially for small sample data, with an average accuracy improvement of 10.06%.

1. Introduction

With the increasing depletion risks of fossil fuels and the rapid development of renewable energy technologies, the rational usage of renewable energy plays an essential role in ensuring global energy security [1,2]. As one of the fastest-growing renewable energy sources in the world, wind power has become one of the most commercially viable forms of renewable energy [3]. However, approximately 20% of worldwide wind turbines are located in ice-prone areas with appreciable blade icing issues during winters [4]. The significant negative effects of blade ice include annual energy production loss, shortened wind turbine lifespan, and threats to personnel safety. Specifically, the annual power generation loss associated with the ice-induced stops and degradations can each up to 50% [5]. The amount of ice accreted at different blade spanwise or edgewise locations of the blades are uneven, and thus the unbalanced loads formed may accelerate the turbine fatigue damage. In addition, blade ice can be thrown out by centrifugal force and gravity from the rotating turbine blades, seriously threatening the safety of personnel [6–8].

Researchers in the wind energy community have conducted extensive studies on blade icing diagnosis, which can be divided into direct and indirect methods. The direct method is the direct monitoring of the blade surface by equipment or instruments, including but not limited to various icing sensors [9], camera imaging [10], and hyperspectral imaging [11]. The indirect method links blade characteristics and icing features, including the physics-driven approaches [12,13] and the data-driven approaches. The physics-driven approaches highly rely on small-scale measurement results and such trends may not be able to apply to utility-scale wind turbines. The data-driven strategies have fewer requirements for the experimental facilities and mainly depend on the recorded data in the utility-scale turbine Supervisory Control and Data Acquisition (SCADA) systems [14–16]. Such approaches are more suitable for large-scale wind turbine applications. However, the researchers may often encounter issues, such as high dimensionality, data imbalance, or incorrect data labels, when dealing with the SCADA data.

Consequently, feature engineering is a promising tool to improve

* Corresponding author.
E-mail address: yqliu@ncepu.edu.cn (Y. Liu).

https://doi.org/10.1016/j.renene.2023.04.107
Received 28 August 2022; Received in revised form 20 April 2023; Accepted 22 April 2023
Available online 24 April 2023

diagnostic accuracy for data-driven approaches [15]. The SCADA systems record more than 100 wind turbine parameters. Such features may have different sensitivities to the icing process, while they may have linear relationships among them. The direct use of the original features is likely to cause feature redundancy and affect the accuracy of icing diagnosis. The feature screening and reconstruction are of vital importance. The current feature engineering for blade icing can be classified into two main categories: 1) The icing mechanism: Tao et al. [16] established the icing diagnosis model based on the icing and wind energy conversion principles. Xu et al. [17] selected features from the process of blade icing to build the diagnostic model. 2) The importance of features: Xiao et al. [18] ranked the importance of features based on the Chi-square test and eliminated the last feature in turn to form a feature subset. Wen et al. [19] reduced the data dimensionality using the ReliefF algorithm, and then further reduced the redundant data using principal component analysis.

Imbalanced data is another barrier to accurate icing diagnostics. The SCADA system contains a large amount of historical data. However, in comparison to the non-icing dataset, the icing dataset is relatively small. Research on such imbalanced data can be divided into two following categories. First, from the data perspective, imbalanced data could be processed via under-sampling and over-sampling methods. The most representative one is the Synthetic Minority Over-Sampling Technique (SMOTE) algorithm proposed by Chawla et al. [20]. Han et al. [21] further improved the SMOTE algorithm by interpolating different classes of boundary data to generate new samples. Second, from the algorithmic perspective, such methods include cost-sensitive learning and integration learning. Ren et al. [22] considered the sample distribution, the convergence trend of samples, and the adaptive sample cost to balance the data samples. Liu et al. [23] proposed two integrated learning algorithms, i.e., EasyEnsemble and BalanceCascad, to handle the imbalanced data. Note that integration learning cannot be explicitly used to deal with imbalanced data, which requires combining it with SMOTE algorithm or cost-sensitive learning [24].

Although the aforementioned studies significantly advance the wind turbine blade icing diagnosis in terms of feature extraction and imbalanced data processing, the following problems still need to be solved to achieve better icing diagnostics.

1) Existing feature extraction methods have difficulties in extracting the best combination of features directly from the original SCADA data. Specifically, feature extraction based on the icing mechanism often requires new features to be modeled, resulting in an unrealistic linear relationship between different features. The selection of features according to their importance may not be the best combination. Some feature extraction methods are highly model-dependent. Besides, the complexity of the turbine SCADA data may also hinders the feature extraction process.
2) Existing imbalanced data processing methods do not fully consider the useful information in the SCADA data in the pre-icing and post-icing periods. From the data perspective, simple undersampling tends to lose important information for most samples, while the oversampling that generates samples outside the original samples tends to cause overfitting issues. From the algorithmic perspective, cost-sensitive learning may encounter difficulties in accuracy and total cost in practice, while integrated learning needs to be used in conjunction with other techniques.

To address the above issues, we propose a data pre-processing approach based on Recursive Feature Elimination with Cross-Validation (RFECV) and Transductive Support Vector Machine (TSVM). First, the RFECV method is applied to the high-dimensional SCADA data to extract the best combination of features based on feature importance ranking (i.e., via the number of reliable features provided by cross-validation). Second, the TSVM is adapted to label the pseudo-samples to effectively complement to the minority samples. The remaining sections are organized as follows. Section 2 explains the specific process of feature extraction using RFECV and tagging of pseudo-samples using TSVM. Section 3 presents the test model, the evaluation criteria and the complete flowchart of the proposed method applied to wind turbine icing diagnostics. Section 4 shows the icing diagnostic performance of the proposed method using the SCADA data from four wind turbines, followed by the discussion on dealing with the small samples. Section 5 concludes the major findings of the study.

2. Feature extraction and pseudo-sample processing

2.1. Optimal feature subset extraction based on SVM-RFECV

Considering the dimensionality of features and the correlation between features, raw SCADA data cannot be directly used to train the classification models. Deep mining or/and a suitable combination of features might be able to be used to determine optimal features for subsequent classification and improve model generalization capability. In this study, we use SVM-RFECV to determine an optimal feature subset from the original ones, which can avoid additional constructed features and improve the model performance simultaneously. SVM-RFECV has two components, i.e., SVM-RFE and CV, as shown in Fig. 1. SVM-RFE uses the SVM as the base model to rank each feature and omit the lowest scoring features [25,26]. SVM-RFE first constructs a linear SVM model based on the input training data set with the initial feature set. The discrimination function for SVM is given by Eq. (1). Then, the feature set is sorted in descending order according to their weight, and the feature with the lowest weight value is removed. The weight vector is computed by Eq. (2). In the next iteration, a new SVM is built based on the training data set of the remaining features, and the process is repeated until all features are removed. Finally, the features are ranked according to the order of omitted features. But this does not determine the number of features in the final feature set. Therefore, the CV algorithm is introduced into the RFE algorithm, and a 5-fold crossover operation is used to determine the optimal number of features.

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

where x is an input sample, b is a bias, ω^T represents the weight vector.

$$\omega = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{2}$$

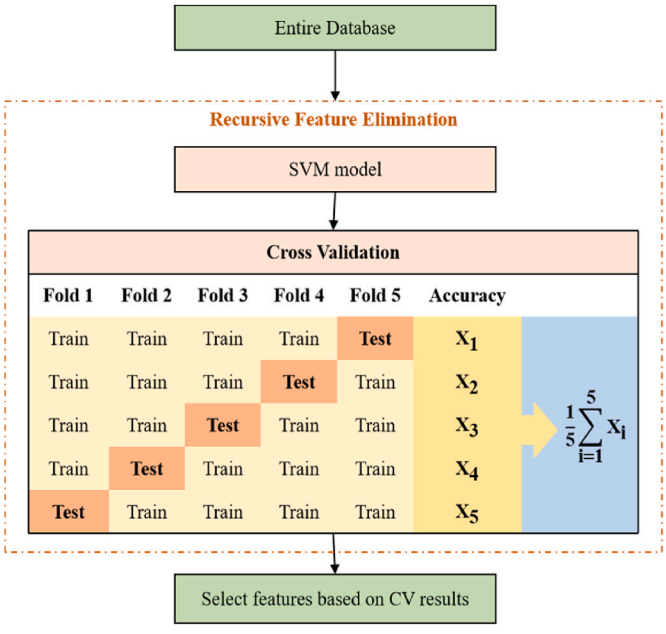


Fig. 1. Feature extraction using SVM-RFECV.

where α_i is the Lagrange multipliers.

2.2. Pseudo-sample processing based on TSVM

Wind turbine blade icing is a complex process including pre-icing, icing, and post-icing periods. The previous studies usually omit the influence of the pre-icing and post-icing periods by not using such data in the model training process. Or some studies directly label the data in such periods as normal (i.e., no-icing)_data in the model training, which could affect the training accuracy. In this paper, the data in the pre-icing and post-icing periods are referred to as pseudo-samples and treated as unlabelled data. The icing data are referred to as absolute icing data, and the normal data are referred to as absolute non-icing data. These pseudo-samples are reclassified using TSVM to utilize all sample data.

The TSVM, similar to the standard SVM, is a learning method for binary classification problems. The major difference is how they handle the unlabelled data. The SVM attempts to find the maximum interval division plane without considering unlabelled samples. In contrast, the TSVM, with unlabelled samples, attempts to find the hyperplane that separates the two classes of labeled samples and passes through the low-density region of the data [27], as shown in Fig. 2.

TSVM assigns various possible markers to the pseudo-samples, tries to use each pseudo-sample as a positive or a negative example, and then determine a hyperplane that maximizes the interval over all samples among all results, with the optimization objective and constraints as described in Eq. (3).

$$\begin{aligned} &\min_{\omega,b,\hat{y},\xi} \frac{1}{2}\|\omega\|_2^2 + C_l\sum_{i=1}^{l}\xi_i + C_u^+\sum_{i=l+1}^{k}\xi_i + C_u^-\sum_{i=k}^{m}\xi_i \\ &s.t.\ y_i\left(\omega^T x_i + b\right) \geq 1-\xi_i, i = 1,2,...,l \\ &\hat{y}_i\left(\omega^T x_i + b\right) \geq 1-\xi_i, i = l+1, l+2,...,m \\ &\xi_i \geq 0, i = 1,2,...,m \end{aligned} \tag{3}$$

where (ω, b) defines a hyperplane. $\xi_i (i = 1,2,...,l)$ corresponds to marked samples and $\xi_i (i = l+1, l+2,..., m)$ refers to unmarked samples. C_l and C_u (including C_u^+ and C_u^-) describe the importance of marked and unmarked samples. C_u^+ and C_u^- correspond to the importance of unmarked samples based on pseudo markers used as positive and negative examples, respectively.

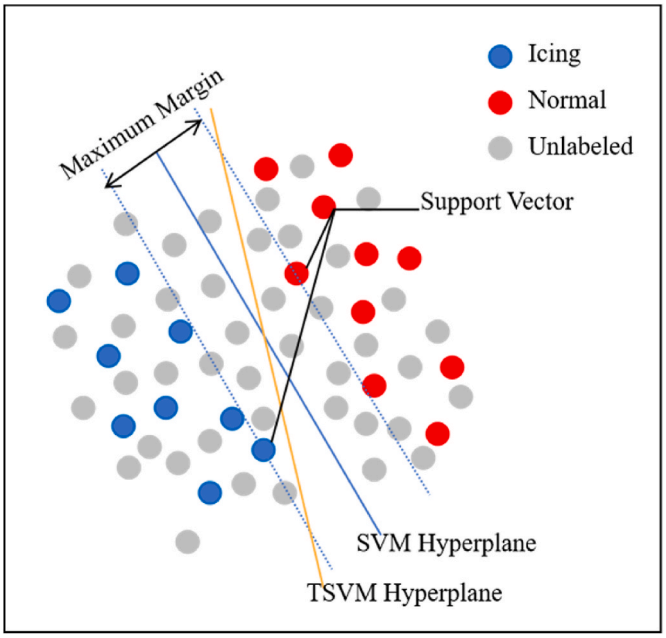


Fig. 2. SVM and TSVM hyperplane segmentation.

Initialize C_u^+ and C_u^- according to Eq. (4).

$$C_u^+ = \frac{u_-}{u_+} C_u^- \tag{4}$$

where u_+ and u_- denote the numbers of unlabelled samples used as positive and negative examples based on pseudo-labelling, respectively.

3. Model evaluation criteria

3.1. Data normalization

To eliminate the adverse effects of different feature sizes on the training model and speed up the training process, the original data is normalized according to Eq. (5) to reduce the range of values to [0, 1].

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{5}$$

where x presents the data under a particular feature, and $max(x)$ and $min(x)$ refer to the maximum and minimum values under such a feature, respectively.

3.2. Classifiers

To verify the performance improvement effect after feature extraction and labelling of pseudo-samples using TSVM, we select three most commonly used classifiers, including Random Forest (RF), Support Vector Machine (SVM) and XGBoost.

RF is an extended variant of Bagging [28]. RF further introduces random attribute selection in the training process of decision trees and uses decision tree as the base learner to build Bagging integration. Assuming that each sample data has N features, for each node of the base decision tree, a subset of n features is randomly selected from all the features of that node ($n \leq N$). Then an optimal attribute is selected from this subset for division so that RF can handle high-dimensional feature samples well.

SVMs have shown excellent performance in classification tasks over time. SVMs are widely used in diagnosis-oriented classification tasks [29,30]. As shown in Fig. 2, the binary classification's blue and red circles represent icing and normal data. Assuming that many hyperplanes separate the two classes of samples, the samples located on the margins are called support vectors. The SVM seeks to find the maximum sum of the distances from the two different support vectors to the hyperplanes. Based on the location of the test sample, it is possible to discern which class it belongs to.

XGBoost is a new algorithm based on Gradient Boosted Decision Trees (GBDT) proposed by Chen et al. [31]. Unlike GBDT, the objective function of XGBoost consists of two parts: the loss function and the regularization term. The complexity of the model is controlled by introducing the regularization term to prevent overfitting. The constant term is removed by second-order Taylor expansion to optimize the loss function term and the regularization term.

3.3. Evaluation

According to the diagnosis results and the actual situation of wind turbine blade icing, the diagnostic result can be characterized into the

Table 1
Confusion matrix of classification results.

Predicted status Actual status	−1	1
−1	True Positive (TP)	False Negative (FN)
1	False Positive (FP)	True Negative (TN)

following four categories, as shown in Table 1. "-1" and "1" represent icing status and normal status, respectively. TP refers to icing data being correctly diagnosed as icing status. FP refers to normal data being incorrectly diagnosed as icing status. FN refers to icing data being incorrectly diagnosed as normal data. TN refers to normal data being correctly diagnosed as normal data.

Based on the confusion matrix, four evaluation indicators can be obtained, as shown in Eqs. (6)–(9).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$Accuracy = \frac{\text{TP} + \text{TN}}{TP + TN + FP + FN} \tag{9}$$

where *Precision* indicates the proportion of correct icing diagnostic data out of all data diagnosed as icing. *Recall* indicates the correct icing diagnostic data proportion out of all true icing data. F1 score indicates the weighted average of *Precision* and *Recall*. *Accuracy* indicates the proportion of correct diagnostic data out of all data.

3.4. Technical frame

Fig. 3 shows the complete flowchart of the proposed method applied to wind turbine icing diagnostics. Firstly, the original data is normalized to eliminate the influence of data dimension. Secondly, the SVM-RFECV mentioned in the second section is used for feature extraction. The feature set is verified by the training set and the test set. Thirdly, TSVM is used to process pseudo-samples to reduce the impact of data imbalance, and the method is verified. Finally, the performance improvement of the proposed method on small sample data is further verified.

4. Results and discussion

4.1. Dataset

The SCADA data of four doubly-fed asynchronous and actively pitch-controlled wind turbines, denoted as A1, A2, A3 and A4, from a mountainous wind farm in Yunnan Province, China, are used in this paper. The average annual wind speed at the four turbines numbered A1, A2, A3, and A4 are respectively 6.39 m/s, 6.95 m/s, 6.47 m/s and 7.66 m/s. The SCADA data contains 18 commonly used variables, as described in Table 2.

The dataset is labled with "1: absolute normal", "-1: absolute icing", and "0: ten days before and after icing (all data in between is taken if less than ten days)". The distribution of absolute normal and icing status data in the original SCADA data is imbalanced. Detailed information of the data from the four wind turbines are listed in Table 3. Wind turbines A1 and A3 have quite small datasets labeled "-1: abolute icing", which take approximately of 2.5% of the entire dataset. Even severe data imbalance is observed for wind turbines A2 and A4, i.e., 0.3%.

Table 2
Wind turbine SCADA variables used in this study.

Feature name	Feature description	Feature name	Feature description
WIND_SPEED	Wind speed	TURINTTMP	Nacelle temperature
REAL_POWER	Grid-side active power	GENAPHSA	Phase A current
CONVERTER_MOTOR_SPEED	Generator speed	GENAPHSB	B-phase current
ROTOR_SPEED	Wind turbine speed	GENAPHSC	C-phase current
WIND_DIRECTION	Wind direction	GENVPHSA	A-phase voltage
TURYAWDIR	Yaw angle	GENVPHSB	B-phase voltage
GBXOILTMP	Gear oil temperature	GENVPHSC	Phase C voltage
GBXSHFTMP	Gearbox bearing temperature	GENHZ	Motor frequency
EXLTMP	Ambient temperature	TURPWRREACT	Reactive power

4.2. Case study

This section focuses on discuss the influence of SVM-RFECV feature extraction and pseudo-sample processing on the icing diagnostic process using three different models. The Python language was used to process pseudo-samples and build the three classifiers described in Section 3.2 on computers with an Intel Core i5-8400U processor.

4.2.1. Icing diagnostics based on original data

The data imbalance is severe for all of the four wind turbines. For wind turbine A1, 1278 × 2 data samples (half of the absolute normal

Table 3
Technical parameters of the wind turbine SCADA datasets used in this study.

Wind turbine #	Total number of samples	Absolute normal data "1"	Absolute icing data "-1"	Before and after icing data "0"
A1	50,596	44,888 (88.72%)	1278 (2.53%)	4430 (8.76%)
A2	50,278	48,682 (96.83%)	167 (0.33%)	1429 (2.84%)
A3	50,500	44,824 (88.76%)	1240 (2.46%)	4436 (8.78%)
A4	50,253	45,875 (91.29%)	190 (0.38%)	4188 (8.33%)

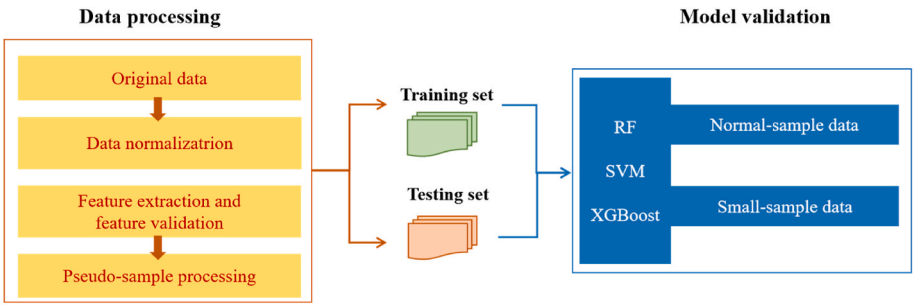


Fig. 3. The flowchart of the proposed method applied to wind turbine icing diagnostics.

data and half of the absolute icing data) are selected using a down-sampling method to fully use all the absolute icing data. It's important to note that the 1278 absolute normal data samples are selected from 44,888 data at equal interval in a time series. The data samples are further grouped into A1_train (2045 data samples for training with 50% absolute icing and 50% absolution normal) and A1_test (511 data samples for testing). The three wind turbines A2, A3 and A4 extract all the absolute icing data as the test samples, named A2_test, A3_test and A4_test. For all three typical classification algorithms, RF, SVM and XGBoost, A1_train is used as the training set, while A1_test, A2_test, A3_test and A4_test are used for model testing.

Table 4 summarizes the test results for four wind turbines using RF, SVM, and XGBoost based on the original data. The models trained by A1_train have highest accuracy in diagnosing the A1_test. The accuracy of all three models decreases when dealing with other turbine datasets, i.e., A2_test, A3_test, A4_test. The XGBoost shows the worst performance by decreasing 12.21% in average accuracy. This phenomenon might be associated with the difference turbine conditions even if they are in the same type. For example, the location of a wind turbine in a wind farm has a great effect on blade icing. On the one hand, under the influence of upstream wind turbines, the turbulence intensity of downstream wind turbines is relatively large, and the difference of turbulence intensity affects the distribution of icing data. In addition, wind speed also has an important effect on blade icing. On the other hand, due to the influence of variable pitch control, the change of wind turbine attack angle also affect the distribution of blade icing data.

4.2.2. Data pre-processing based on SVM-RFECV and TSVM

To verify the performance improvement of SVM-RFECV feature extraction and TSVM pseudo-sample processing, we use the methods proposed in Sections 2.1 and 2.2. A1_train is input as the initial set of features into the SVM-RFECV algorithm. Fig. 4 shows that the accuracy varies with the number of features during SVM-RFECV feature extraction, where the blue shaded part represents the standard deviation. As can be seen in Fig. 3, the SVM model in the SVM-RFECV algorithm achieve the highest accuracy as the number of features equals 8.

The above finding is only derived for icing diagnosis of one wind turbine using a particular model. A subset of features is still needed that can be generalized across different turbines and different models. Thus, a feature subset with the number of features from 3 to 18 is obtained according to the SVM-RFECV algorithm. Moreover, 16 feature subsets are calculated with A2_test, A3_test, A4_test as test sets for accuracy under three typical classifiers. Comparing Figs. 4 and 5, the best generalization performance of the feature subset over different classifier models and wind turbines is observed with 8 features, i.e., wind speed, generator speed, wind direction, gear oil temperature, gearbox bearing temperature, ambient temperature, nacelle temperature and reactive power. This feature subset has some commonalities with the features obtained from the icing mechanism, such as wind speed, ambient temperature, and nacelle temperature. In addition, the subset of features obtained based on the SVM-RFECV algorithm selection does not include

Table 4
Results of icing diagnosis based on original features.

Test set	Models	Precision	Recall rate	F1 score	Accuracy
A1_test	RF	99.57%	100.00%	99.80%	99.77%
	SVM	100.00%	99.59%	99.79%	99.80%
	XGBoost	99.63%	100.00%	99.79%	99.84%
A2_test	RF	95.71%	93.41%	94.55%	94.61%
	SVM	90.21%	99.40%	94.58%	96.21%
	XGBoost	74.55%	100.00%	85.42%	88.62%
A3_test	RF	93.17%	96.85%	94.98%	94.87%
	SVM	85.14%	96.53%	90.69%	89.82%
	XGBoost	86.36%	95.48%	90.69%	90.20%
A4_test	RF	82.46%	98.95%	89.95%	88.95%
	SVM	84.79%	96.84%	90.42%	89.74%
	XGBoost	79.81%	89.47%	84.37%	83.87%

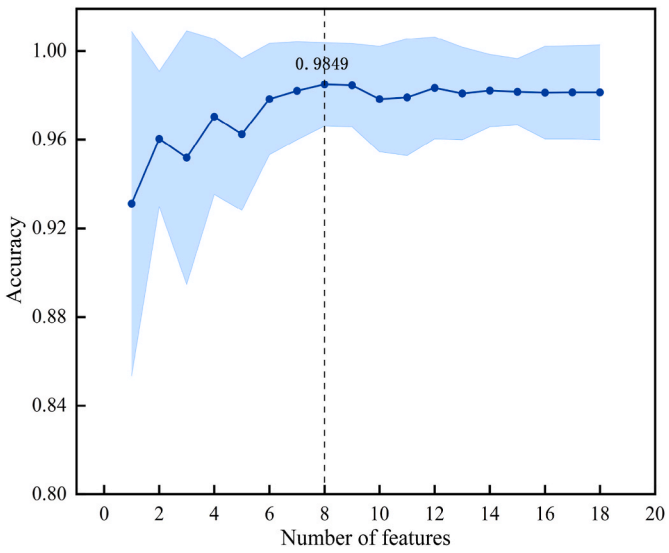


Fig. 4. Accuracy versus the number of features.

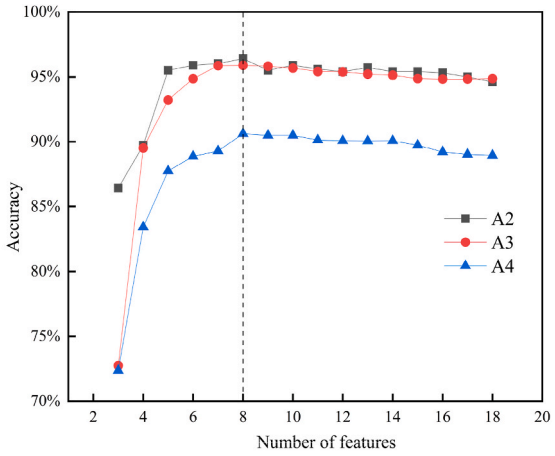
actual active power but rather the generator speed and reactive power. The combination of the latter reflects not only the actual active power to a certain extent but also the state and performance of wind turbines.

With the selected features, pseudo sample processing is performed on the original SCADA data. Wind turbine blade icing is a time-dependent process from ice accretion to ice natural melting, lasting from a few hours, a few days or even several months [32]. Considering the influence of environmental factors on the accretion rate of icing in different regions, we use ten days of data before and after absolute icing as a pseudo-sample to include as much icing information as possible. In the initialization stage, the pseudo-labels of unmarked samples are likely to be inaccurate, so the value of C_l should be much larger than C_u so that the labeled samples can play a greater role. As the iterative process proceeds, the value of C_u should increase and gradually approach C_l so that marked samples and pseudo samples play the same role in model training. The above process is implemented by $C_u = \min\{2C_u, C_l\}$, and the iterative process can be accelerated by using the 2-fold relation. In the process of label assignment and adjustment for unmarked samples, the problem of class imbalance may occur. Therefore, on the basis of $C_u = C_u^+ + C_u^-$, C_u^+ and C_u^- is initialized according to Eq. (4). The recalibrated pseudo-sample not only solves the problem of insufficient original absolute icing data but also positively impacts the icing diagnosis rate. Table 5 shows the training time of the model after feature selection and pseudo-sample processing. This is respectively compared to the time taken to train the model directly with the original data and the time taken to train the model with only the feature-selected data. The comparison results show that the feature-selected and pseudo-sample processed data takes less time to train the model. This effectively demonstrates that the selected features and pseudo-sample processed data are model friendly and reduce the computational complexity.

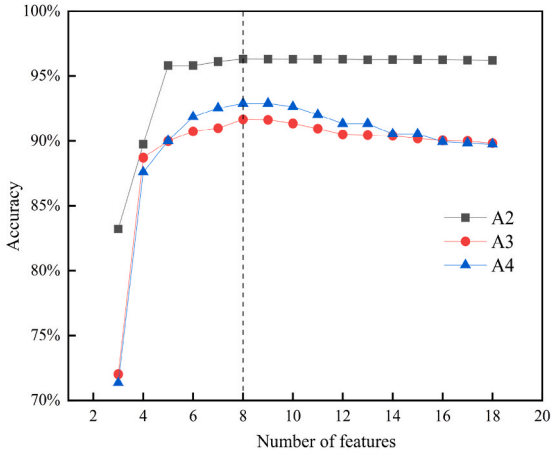
Fig. 6 shows the icing diagnosis accuracy with the proposed feature extraction and pseudo-sample processing for the three models. With SVM-RFECV feature extraction, RF, SVM, and XGBoost have better performance with increments in average accuracy of 1.49%, 1.70% and 2.99%. With pseudo-sample processing on top of feature extraction, the average accuracy of the three classifiers are further improved by 1.56% (RF), 3.84% (SVM) and 6.24% (XGBoost), with the most significant improvement for XGBoost.

4.2.3. Small-sample data

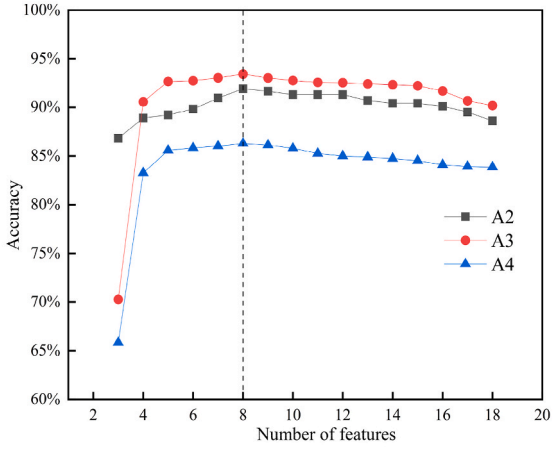
To further demonstrate the effectiveness of the proposed method for small-sample data, all absolute icing data (167 samples) are extracted from the A2 dataset as training samples, and the training set A2_train is



a) Accuracy of different feature subsets using RF classification algorithm.



b) Accuracy of different feature subsets using the SVM classification algorithm.

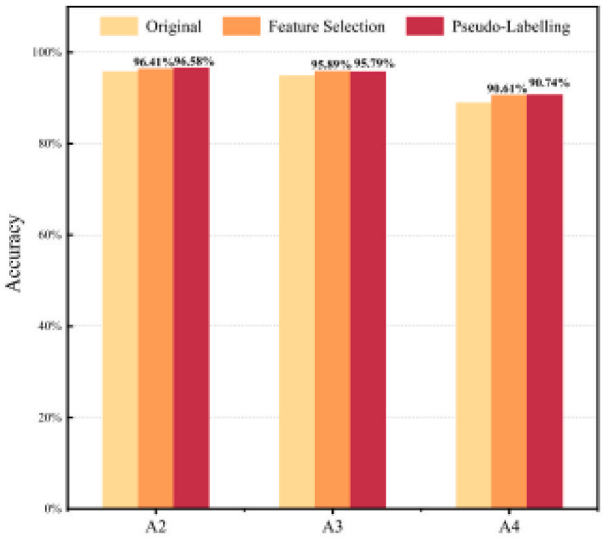


c) Accuracy of different feature subsets using XGBoost classification algorithm.

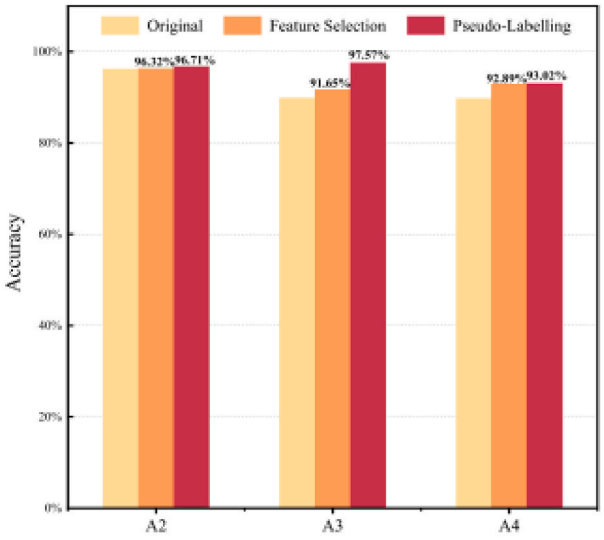
Fig. 5. Variation of different feature subsets under different classifiers.

Table 5
Comparison of the time taken to train the model.

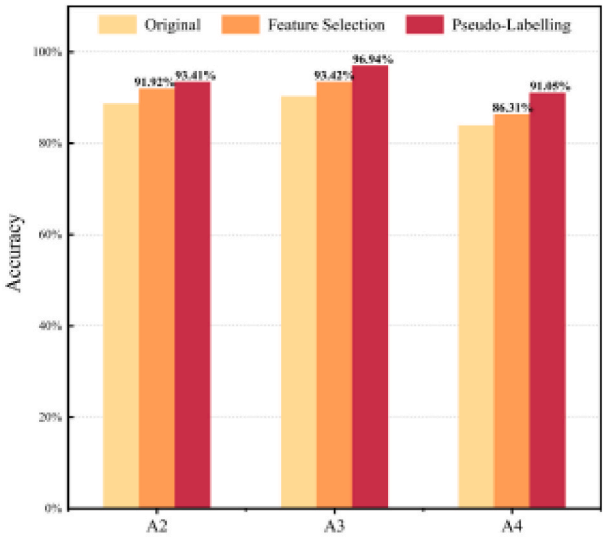
	Train time (s)		
	RF	SVM	XGBoost
Original data	0.36253	0.19948	0.30471
Feature-selected data	0.36045	0.18550	0.24963
Pseudo-sample processed data	0.35087	0.17854	0.21674



a) Icing diagnostic accuracy using RF classification algorithm



b) Icing diagnostic accuracy using SVM classification algorithm.



c) Icing diagnostic accuracy using XGBoost classification algorithm.

Fig. 6. Performance of the three classification algorithms.

constructed according to the principle of a 50/50 split between normal data and icing data. The same method is used to extract data from the rest three wind turbines, A1, A3 and A4, to form the new test sets A1_test, A3_test, A4_test. The feature subsets of the training and testing sets are the same as those extracted in Section 4.2.2. Fig. 7 shows the icing diagnosis accuracy of the three typical models after feature extraction and pseudo-sample processing. The results show that the average accuracy of the three typical classification models, RF, SVM and XGBoost, increase by 8.25%, 7.86% and 14.07%, respectively. Such trend indicates the proposed method is beneficially in dealing with small-sample data.

5. Conclusions

This paper presents an effective data pre-processing method for wind turbine blade icing diagnostics. Such a method can address the existing difficulties in dealing with the high-dimensional, complex wind turbine SCADA data, such as feature selection, data imbalance and the improper data labelling. Specifically, we use RFECV to dissect the influence degree of various SCADA features on the diagnostic model, select the most compelling features, and calibrate the pseudo-samples. TSVM is then implemented to regenerate the pseudo-samples to fully consider the pre-icing and post-icing periods. The effectiveness of the proposed method is systematically examined using the three most commonly used classifiers algorithms, i.e., Random Forest (RF), Support Vector Machine (SVM), and XGBoost, for four utility-scale wind turbines and further examined for the applications in small-sample data.

(1) The best combination of features obtained based on the SVM-RFECV algorithm effectively improves the icing diagnosis accuracy by avoiding the construction of complex features and simplifying feature extraction. For the three test wind turbines, the icing diagnosis accuracy of the three algorithms is improved by 2.06% on average.
(2) The pseudo-sample processing method based on the TSVM algorithm uses the SCADA data before and after absolute icing and effectively extracts the icing information from this data. Compared to the original data, the icing diagnosis accuracy of the three algorithms is improved by an average of 3.88%.
(3) The proposed methods based on the original feature combination and TSVM pseudo-sample data pre-processing are further validated on a small sample dataset. The results showed that the proposed methods have significant advantages, with an average improvement of 10.06% in the accuracy of the three algorithms.

The proposed model is suitable for utility-scale inland wind turbines. It should be cautioned to apply the proposed model to the offshore wind turbine icing diagnostics, and the small-scale wind turbines.

Data availability

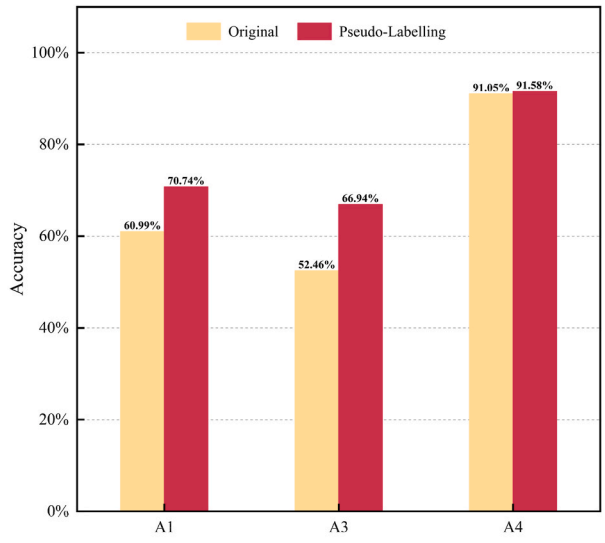
The data used in this study are confidential at the request of the wind farm operators.

CRediT authorship contribution statement

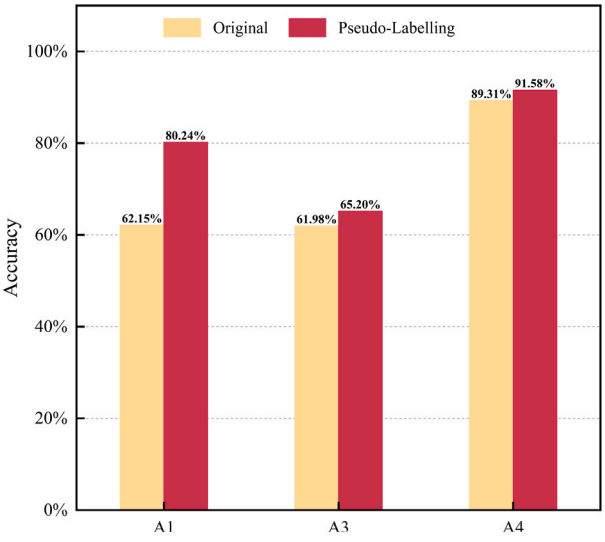
Xinjian Bai: Conceptualization, Methodology, Software, Writing – original draft. **Tao Tao:** Conceptualization, Writing – review & editing. **Linyue Gao:** Conceptualization, Writing – review & editing. **Cheng Tao:** Conceptualization, Formal analysis, Data curation. **Yongqian Liu:** Conceptualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

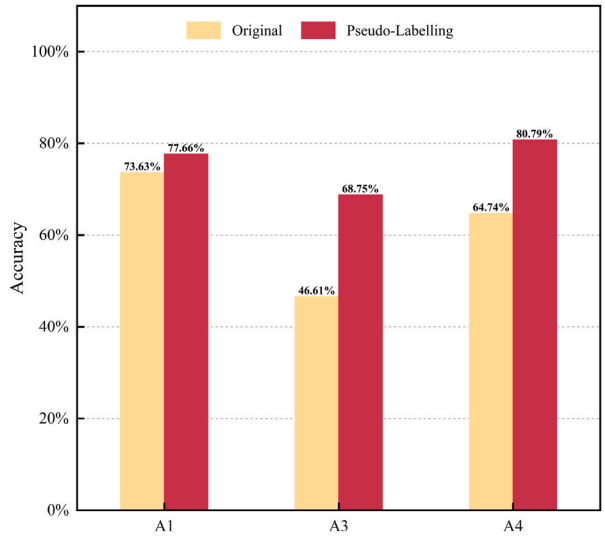
The authors declare that they have no known competing financial



a) Icing diagnostic accuracy using RF classification algorithm.



b) Icing diagnostic accuracy using SVM classification algorithm.



c) Icing diagnostic accuracy using XGBoost classification algorithm.

Fig. 7. Performance of the three classification algorithms for small-sample datasets.

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work presented in this paper is part of the project “Research on smart operation control technologies for offshore wind farms” supported by the National Key Research and Development Program of China (No.2019YFE0104800).

References

[1] X. Lei, M. Alharthi, I. Ahmad, B. Aziz, Z. ul Abdin, Importance of International Relations for the Promotion of Renewable Energy, Preservation of Natural Resources and Environment: Empirics from SEA Nations, Renewable Energy, 2022, https://doi.org/10.1016/j.renene.2022.07.083.
[2] Z. Fareed, U.K. Pata, Renewable, non-renewable energy consumption and income in top ten renewable energy-consuming countries: advanced Fourier based panel data approaches, Renew. Energy 194 (2022) 805–821, https://doi.org/10.1016/j.renene.2022.05.156.
[3] D. Alemzero, T. Acheampong, S. Huaping, Prospects of wind energy deployment in Africa: Technical and economic analysis, Renew. Energy 179 (2021) 652–666, https://doi.org/10.1016/j.renene.2021.07.021.
[4] E. Madi, K. Pope, W. Huang, T. Iqbal, A review of integrating ice detection and mitigation for wind turbine blades, Renew. Sustain. Energy Rev. 103 (2019) 269–281, https://doi.org/10.1016/j.rser.2018.12.019.
[5] O. Parent, A. Ilinca, Anti-icing and de-icing techniques for wind turbines: critical review, Cold Reg. Sci. Technol. 65 (2011) 88–96, https://doi.org/10.1016/j.coldregions.2010.01.005.
[6] L. Gao, T. Tao, Y. Liu, H. Hu, A field study of ice accretion and its effects on the power production of utility-scale wind turbines, Renew. Energy 167 (2021) 917–928, https://doi.org/10.1016/j.renene.2020.12.014.
[7] V. Daniliuk, Y. Xu, R. Liu, T. He, X. Wang, Ultrasonic de-icing of wind turbine blades: performance comparison of perspective transducers, Renew. Energy 145 (2020) 2005–2018, https://doi.org/10.1016/j.renene.2019.07.102.
[8] M.C. Homola, P.J. Nicklasson, P.A. Sundsbø, Ice sensors for wind turbines, Cold Reg. Sci. Technol. 46 (2006) 125–131, https://doi.org/10.1016/j.coldregions.2006.06.005.
[9] S. Gantasala, J.-C. Luneno, J.-O. Aidanpaa, Detection of Ice Mass Based on the Natural Frequencies of Wind Turbine Blade, Wind Energy Science Discussions, 2016, pp. 1–17, https://doi.org/10.5194/wes-2016-30.
[10] B. Guan, Z. Su, Q. Yu, Z. Li, W. Feng, D. Yang, D. Zhang, Monitoring the blades of a wind turbine by using videogrammetry, Opt Laser. Eng. 152 (2022), 106901, https://doi.org/10.1016/j.optlaseng.2021.106901.
[11] P. Rizk, N. Al Saleh, R. Younes, A. Ilinca, J. Khoder, Hyperspectral imaging applied for the detection of wind turbine blade damage and icing, Remote Sens. Appl.: Society and Environment 18 (2020), 100291, https://doi.org/10.1016/j.rsase.2020.100291.
[12] F. Li, H. Cui, H. Su, Iderchuluun, Z. Ma, Y. Zhu, Y. Zhang, Icing condition prediction of wind turbine blade by using artificial neural network based on modal frequency, Cold Reg. Sci. Technol. 194 (2022), 103467, https://doi.org/10.1016/j.coldregions.2021.103467.
[13] A.A. Jiménez, F.P. García Márquez, V.B. Moraleda, C.Q. Gómez Muñoz, Linear and nonlinear features and machine learning for wind turbine blade ice detection and diagnosis, Renew. Energy 132 (2019) 1034–1048, https://doi.org/10.1016/j.renene.2018.08.050.
[14] X. Dong, D. Gao, J. Li, Z. Jincao, K. Zheng, Blades icing identification model of wind turbines based on SCADA data, Renew. Energy 162 (2020) 575–586, https://doi.org/10.1016/j.renene.2020.07.049.
[15] W. Chen, Y. Qiu, Y. Feng, Y. Li, A. Kusiak, Diagnosis of wind turbine faults with transfer learning algorithms, Renew. Energy 163 (2021) 2053–2067, https://doi.org/10.1016/j.renene.2020.10.121.
[16] T. Tao, Y. Liu, Y. Qiao, L. Gao, J. Lu, C. Zhang, Y. Wang, Wind turbine blade icing diagnosis using hybrid features and Stacked-XGBoost algorithm, Renew. Energy 180 (2021) 1004–1013, https://doi.org/10.1016/j.renene.2021.09.008.
[17] J. Xu, W. Tan, T. Li, Predicting fan blade icing by using particle swarm optimization and support vector machine algorithm, Comput. Electr. Eng. 87 (2020), 106751, https://doi.org/10.1016/j.compeleceng.2020.106751.
[18] J. Xiao, C. Li, B. Liu, J. Huang, L. Xie, Prediction of wind turbine blade icing fault based on selective deep ensemble model, Knowl. Base Syst. 242 (2022), 108290, https://doi.org/10.1016/j.knosys.2022.108290.
[19] X. Wen, Z. Xu, Wind turbine fault diagnosis based on ReliefF-PCA and DNN, Expert Syst. Appl. 178 (2021), 115016, https://doi.org/10.1016/j.eswa.2021.115016.
[20] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Jair 16 (2002) 321–357, https://doi.org/10.1613/jair.953.
[21] H. Han, W.Y. Wang, B.H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, in: D.S. Huang, X.P. Zhang, G.B. Huang (Eds.), Advances in Intelligent Computing, Pt 1, Proceedings, Springer-Verlag Berlin, Berlin, 2005, pp. 878–887, https://doi.org/10.1007/11538059_91.
[22] Z. Ren, Y. Zhu, W. Kang, H. Fu, Q. Niu, D. Gao, K. Yan, J. Hong, Adaptive cost-sensitive learning: improving the convergence of intelligent diagnosis models under imbalanced data, Knowl. Base Syst. 241 (2022), 108296, https://doi.org/10.1016/j.knosys.2022.108296.
[23] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 (2009) 539–550, https://doi.org/10.1109/TSMCB.2008.2007853.
[24] W. Lu, Z. Li, J. Chu, Adaptive Ensemble Undersampling-Boost: a novel learning framework for imbalanced data, J. Syst. Software 132 (2017) 272–282, https://doi.org/10.1016/j.jss.2017.07.006.
[25] D. Zhou, X. Zhuang, H. Zuo, H. Wang, H. Yan, Deep learning-based approach for civil aircraft hazard identification and prediction, IEEE Access 8 (2020) 103665–103683, https://doi.org/10.1109/ACCESS.2020.2997371.
[26] C. Wang, Z. Xiao, J. Wu, Functional connectivity-based classification of autism and control using SVM-RFECV on rs-fMRI data, Phys. Med. 65 (2019) 99–105, https://doi.org/10.1016/j.ejmp.2019.08.010.
[27] T. Joachims, Transductive inference for text classification using support vector machines, in: Proceedings of the Sixteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 200–209.
[28] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[29] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297, https://doi.org/10.1007/BF00994018.
[30] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 121–167, https://doi.org/10.1023/A:1009715923555.
[31] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[32] L. Hu, X. Zhu, J. Chen, X. Shen, Z. Du, Numerical simulation of rime ice on NREL Phase VI blade, J. Wind Eng. Ind. Aerod. 178 (2018) 57–68, https://doi.org/10.1016/j.jweia.2018.05.007.