Temporal Attention Convolutional Neural Network for Estimation of Icing Probability on Wind Turbine Blades

Xu Cheng, *Member, IEEE*, Fan Shi, *Member, IEEE*, Meng Zhao, *Member, IEEE*, Guoyuan Li, *Senior Member, IEEE*, Houxiang Zhang, *Senior Member, IEEE*, and Shengyong Chen, *Senior Member, IEEE*

***Abstract*—Wind farms are usually located in high-latitude areas, which brings a high risk of icing. Traditional methods of anti-blade-icing are limited by extra costs and potential damages to the original mechanical structure. Model-based methods are heavily dependent on mathematical models of the blade icing, which are prone to produce erroneous estimation. As data-driven models are better able to achieve competitive performances for the blade icing estimation, this paper proposes a temporal attention-based convolutional neural network (TACNN). This novel data-driven model introduces a temporal attention module into a convolutional neural network, with the goal of determining the importance of sensors and timesteps and automatically identifying discriminative features from raw sensor data. Benchmark experiments on ten public datasets of multivariate time series classification show competitive performance against the state-of-the-art methods. Compared with ten baseline networks and three widely used attention mechanisms, the TACNN shows significant advantages applying to three real-world datasets. These datasets are logged by the supervisory control and data acquisition system and contain operational and environmental measurements such as power and temperature. The ablation study and sensitivity study demonstrate the effectiveness of the key components of the TACNN. The practicability of the TACNN is further verified through online estimation testing.**

***Index Terms*—Wind Turbine, Icing Detection, Convolutional Neural Network, Temporal Attention, Time Series Classification**

Manuscript received Month xx, 2xxx; revised Month xx, xxxx; accepted Month x, xxxx. This work was in part supported by National Natural Science Foundation of China (62020106004, 92048301, 61906133, 61703304) and in part by the Opening Foundation of Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, China (TJUT-KLICNST-K20180002). *(Corresponding author: Fan Shi, Shengyong Chen.)*

Xu Cheng, Fan Shi, Meng Zhao and Shengyong Chen are with Engineering Research Center of Learning-Based Intelligent System (Ministry of Education), the Key Laboratory of Computer Vision and System (Ministry of Education), the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, 300384, China. (e-mail: shifan@email.tjut.edu.cn. xu.cheng@ieee.org, mzhao@ieee.org, sy@ieee.org).

Guoyuan Li and Houxiang Zhang are with the Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Aalesund, 6009 Norway.(e-mail: guoyuan.li@ntnu.no. hozh@ntnu.no).

Xu Cheng and Meng Zhao are equal contribution.

I. Introduction

WIND energy, which is abundant and inexhaustible, has for decades been an important energy source [1]. In order to provide a reliable source of energy, wind turbines must be able to operate under various climate conditions. This may entail icing, especially in winter. Ice accreting on the blades of wind turbines changes their aerodynamic efficiency and torque, resulting in a reduction in power generation, as well as intensification of fatigue loads [2]. Additionally, severe icing brings potential safety hazards, affecting the economic benefits and stable operation of wind farms. Improving the detection of ice accretion on blades is of paramount importance to the proper maintenance of wind farms [3], [4].

Traditionally, there have been three main types of anti-icing/de-icing for wind turbine blades: passive, active, and hybrid. Passive methods involve the application of certain materials to prevent blade surfaces from icing. Special materials such as liquid-infused surfaces [4], [5] have been employed to prevent icing. Advantages of passive methods include reduced operating costs and the ability to keep wind turbine blades ice-free without the need for control systems [6]. Active methods depend on the external system (e.g., thermal or mechanical methods) to remove ice. They have garnered increased attention because they work in a controllable fashion. In active methods, electro-thermal and ultrasonic de-icing devices are external sensors and equipment widely used for blade icing detection [7], [8]. Hybrid methods have become increasingly popular in recent years, as they offer a combination of the advantages of both active and passive methods [2], [9]. Conventional anti-icing/de-icing methods often incur high costs and additional energy demands. In addition, these methods suffer from internal unreliability that may cause inaccurate estimations of icing conditions. To make matters worse, these traditional methods may require mechanical replacement of the wind turbine, a process that requires significant human effort and may cause damage to the original mechanical structure [10].

In order to deal with the disadvantages of traditional methods, several researchers have made extensive efforts to identify ice conditions on the basis of the operational measurements of a given wind turbine. This approach has largely involved model-based, data-driven, and hybrid methods. Model-based approaches have been proposed to establish mathematical or numerical models, with the help of human domain knowledge [11]. However, these methods depend heavily on assumptions, which can lead to the misidentification of icing conditions. They also require costly external experimental tools (e.g., wind tunnels) to establish accurate models. Conversely, data-driven approaches directly mine useful information hidden in operational measurements [12], [13]. The advantage of data-based methods is that they do not rely on prior domain knowledge and only need to use existing sensors; this saves on cost. Hybrid methods integrate the advantages of both model-driven and data-driven methods [14].

Data-driven methods, which can be further roughly divided into shallow machine learning and deep learning-based methods [15], have been used widely in fault detection in key components of wind turbines such as gearboxes [16] and main shafts [17], and corresponding conditions monitoring systems have been introduced. Shallow machine learning methods identify blade icing by extracting the representative features characterizing icing conditions and then creating classification models from those extracted features. Commonly used shallow machine learning models include logistic regression, support vector machine, artificial neural networks, and random forest [12], [13], [18]–[20]. The limitation of shallow machine learning methods is that the process for obtaining such features is usually time-consuming and can be very expensive. Deep learning-based methods attempt to model high-level representations of sensor data and identify icing conditions via a hierarchical structure [10], [16], [21], which is more competitive in terms of performance than shallow machine learning methods are.

However, to the best of our knowledge, their use in detecting icing on wind turbine blades has not yet been extensively studied [10], [12], [21]. There are even fewer studies of deep learning-based methods for icing detection [10]. This may be because of four main challenges to apply deep learning-based methods to this type of task. The first is that wind turbines usually work in varying environmental conditions, and therefore the measured sensor data are characterized by high nonlinearity and non-stationarity quality. Thus, automatically extracting useful features from raw sensor data for icing detection is quite difficult. The second challenge is the substantial imbalance between the non-icing (i.e., normal) and icing statuses in such data. It is not easy to properly process the raw sensor data to avoid biased identification and make a deep learning model learn all the possible features of the icing status. The third challenge is how to achieve early predictions of blade icing on wind turbines. Though difficult, predicting icing as early as possible and identifying icing trends in advance would be very beneficial for engineers and maintenance personnel, sparing the additional time needed to activate anti/de-icing systems. The fourth challenge is how to identify the importance of sensors and timesteps. Detection of icing conditions can be modeled as a time series classification (TSC) task, but most researches on TSC have focused on the architectural design of deep learning models. Few studies have investigated the importance of particular sensors and timesteps. Intuition indicates that the temperature sensor would be significant in distinguishing between icing and non-icing conditions, but scant research has provided a roadmap for evaluating whether this is actually the case. Likewise, the features of some timesteps, such as those already in an icing state or those that are about to freeze, may show a more salient pattern than others, but research has not delineated such distinctions.

To address these challenges, a convolutional neural network (CNN) is adopted in the present research, due to its excellent learning capabilities. To overcome the data imbalance, a specially designed imbalanced data processing approach is utilized. A temporal attention (TA) module is integrated into a CNN to learn the importance of sensors and timesteps. A TA-based CNN (TACNN) is then designed to automatically learn and discover discriminative features from balanced data and classify icing or non-icing conditions of wind turbine blades. The proposed TACNN can learn the relationships between different timesteps, so as to predict the icing probability of wind turbine blades at an early stage.

The contributions of this research can be summarized as follows:

1) A novel deep learning network, TACNN, is presented by introducing a TA module into a conventional CNN. The TACNN ensures effective features extraction through its ability to learn the importance of sensors and timesteps and overcome the limitations of conventional CNNs that treat each sensor equally. An end-to-end framework is developed based on the proposed TACNN for icing detection on wind turbine blades. The framework successfully processes highly imbalanced sensor data and simultaneously ensures both automatic discriminative feature learning and effective icing conditions identification.

2) The performance of the proposed TACNN is evaluated according to ten benchmark datasets of multivariate TSC and real supervisory control and data acquisition (SCADA) data from three wind turbines. Compared to the state-of-art TSC methods in the ten benchmark datasets, TACNN achieves better performance. The comparisons of baseline networks and other attention modules in the SCADA data demonstrate the superiority and significance of the proposed model. The generalizability and practicability of the proposed model are further verified through online estimation testing.

The rest of this research is structured as follows. Section II reviews the related work for wind turbine blade icing detection and TSC. The proposed TACNN is presented in Section III. The performance of the proposed approach is evaluated in Section IV. Section V emphasizes the conclusions and future work.

II. Related Work

Detection of icing conditions can be modeled as a TSC task, it is, therefore, necessary to review the methods used in TSC. The primary TSC algorithms are distance-based, feature-based, or deep learning-based. Orsenigo et al. proposed a distance-based method that combines discrete SVM and warping distances [23]. Feature-based methods classify time series data based on the patterns extracted from time series. Models of feature-based approaches mainly involve a bag-of-features

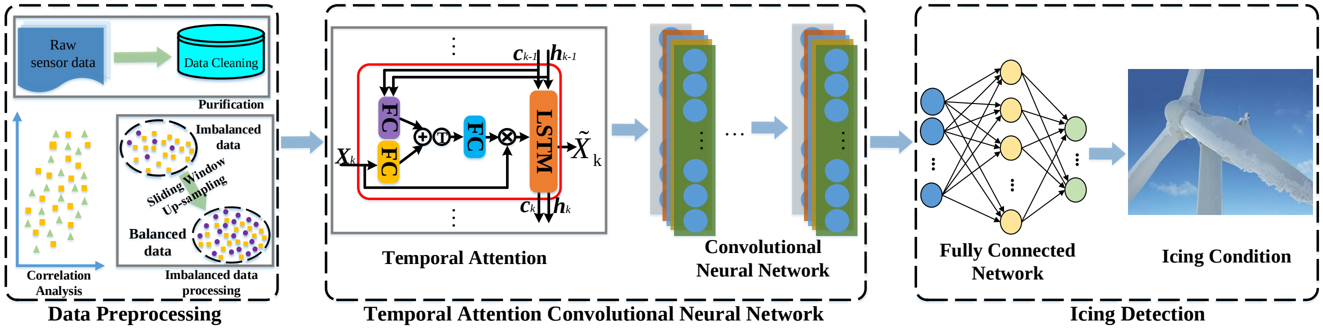


Fig. 1: Framework of wind turbine blade icing detection. The photo of frozen wind turbine blades is from [22].

framework [24], bag-of-SFA-symbols structure [25], or hidden unit logic model [26]. One limitation of feature-based methods is that extracting effective features requires substantial human effort and domain knowledge. Deep learning-based methods have been thoroughly explored in recent years, motivated by the need to overcome the shortcomings of feature-based methods. Several alternatives to deep learning models have been proposed by various researchers. For example, Wang et al. recommended several baseline models for TSC [27]. Fazle et al. proposed a parallel structure of long short-term memory (LSTM) and a fully convolutional network (FCN) (LSTM-FCN) [28]. A similar deep learning model was proposed by Cheng et al., but an additional spectral branch was added to the LSTM-FCN [29]. Moreover, the authors presented a novel model combining dense connections and a CNN to achieve state-of-the-art performance [30]. Zhang et al. proposed a prototype-based deep learning model (TapNet) for TSC. This model showed a competitive performance over others. The literature exploring these deep learning models in TSC applications has mainly focused on model structural design; the attention mechanisms used widely in computer vision and natural language processing have yet to be thoroughly investigated.

III. The Temporal Attention Convolutional Neural Network for Icing Detection

A. Structure

The proposed approach consists of three components: data preprocessing, a temporal attention convolutional neural network, and icing detection, as depicted in Fig. 1. The performance of data-driven models depends almost exclusively on access to high-quality data. To reduce uncertainty regarding data quality, researchers have typically begun with data cleaning and processing and correlation analysis. Wind turbines typically operate in conditions without ice. Thus, the problem of imbalanced data remains. In the proposed methodology, processed sensor data are fed into the TACNN model. The sensor data sent to the TACNN model are first processed by the TA module. The TA module has been designed to identify key sensors and important timesteps. The weighted sensor data are then sent to the CNN for feature extraction. Finally, the features learned by the TACNN are utilized to calculate the probability of icing in the fully connected (FC) network.

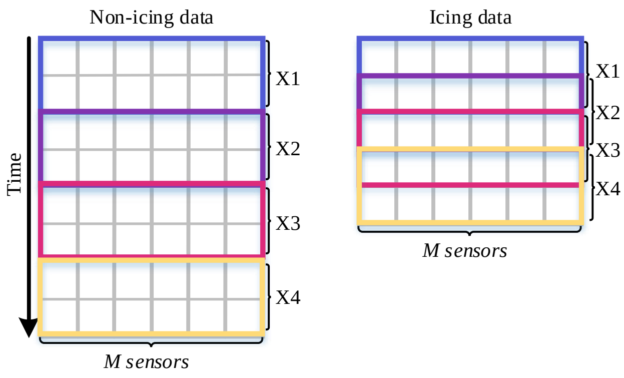


Fig. 2: An illustration example of SWU.

B. Data Preprocessing

As illustrated in Fig. 1, data preprocessing mainly involves data cleaning, correlation analysis, and imbalanced data processing. To reduce the impact of outliers and noise in the sensor data, it is necessary to clean the raw sensor data. Correlation analysis is employed to study the relationships between sensors, which is helpful in reducing the amount of redundant information. In this work, the Pearson correlation analysis is utilized. Various methods have been developed for handling class-imbalanced learning problems. A common method of rebalancing is to undersample the majority samples or oversample the minority samples. However, undersampling may eliminate potentially useful information, and improper over-sampling may lead to overfitting [28]. Another solution for rebalancing is to use a specially designed loss, such as the focal loss function [29]. However, it is necessary to optimize some hyper-parameters in order to use these methods. Even worse, in practical use, focal loss has not greatly improved model performance. To address this challenge of data imbalance, in the proposed method, a sliding window upsampling (SWU) method is employed [3]. The idea of SWU is simple but effective. Non-icing data are partitioned in a non-overlapping way with a window of predefined size, while an overlapped sliding window is utilized for the icing data. As illustrated in Fig. 2, more non-icing data points are collected than icing data points (eight points for non-icing and five points for icing). By using SWU, four samples can be obtained for non-icing data with a non-overlapping sliding window, and four samples can be obtained with an overlapping sliding window, if the window size of samples is set to two.

C. Temporal Attention Convolutional Neural Network

CNN is utilized for automatic feature learning because wind turbines usually work under varying environmental conditions. Previous works focused on improving CNN's automatic feature learning capability (i.e., structural design). However, traditional CNNs can't focus on adaptively learning more distinguishing features and ignoring those that are irrelevant. To address these challenges, the attention mechanism is usually utilized.

Attention has been recognized as an important aspect of human perception. To the best of our knowledge, squeeze-and-excitation (SE) is the first attention module to be integrated into an FCN model and applied to TSC [28]. There have been some other attention modules, such as the convolutional block attention module (CBAM) [31] and global contexts (GC) [32], which are specially designed for these applications in computer vision. A common characteristic of these attention modules is to re-weight the features extracted by a CNN to obtain their importance. In other words, these attention modules are attached after the CNN layers. Although such attention mechanisms can effectively improve the performance of a CNN, they cannot recognize the importance of each sensor and timestep in raw time series data.

To address this issue with attention modules, the present work proposes a TA module. The main difference between the proposed TA module and attention modules (e.g., SE, CBAM, GC) is that our proposed TA module is placed before the CNN layer, and thus can directly calculate the importance of each sensor and timestep.

To be able to calculate the importance of each sensor and time step, the cell state c_{t-1} and hidden state h_{t-1} of previous timesteps in LSTM are utilized [33]. As shown by the middle panel in Fig. 1, the two states (c_{t-1} and h_{t-1}) are transformed by an FC layer. A similar operation is also applied to the input. As shown in Fig. 1, the encoded input and states are then transformed by certain operations such as add, tan, and FC, before they are sent to the LSTM.

Assuming the number of sensors is N, the number of timesteps is T, and the number of hidden units in LSTM is M, the shape of the purple FC is defined as $W_p \in \mathbb{R}^{2M\times T}$, the yellow as $W_y \in \mathbb{R}^{T\times T}$, and the blue as $W_b \in \mathbb{R}^{T\times 1}$. The subscripts p, y, and b represent the colors of the three FC layers, as shown in Fig. 1. The calculation of these three FCs can be represented as follows:

$$\begin{aligned} z_p &= W_p([h_{t-1}; c_{t-1}]), \\ z_y &= W_y(X_k), \\ z_b &= W_b(tan(z_p + z_y)), \end{aligned} \tag{1}$$

where z_p, z_y, and z_b donate the output of the three FC layer. Then, the importance of each sensor can be computed as:

$$\alpha_k = softmax(z_b) = \frac{exp(z_b)}{\sum_{j=1}^{n} exp(z_b^j)} \tag{2}$$

where α_k is the weight for each timestep in X_k. Thus, the input for each timestep can be represented as:

$$\widetilde{X_k^t} = \alpha_k * X_k^t = (\alpha_k^1 X_k^1, \alpha_k^2 X_k^2, \cdots, \alpha_k^n X_k^n) \tag{3}$$

The whole time series is weighted by LSTM and fed into the CNN for feature extraction. There is one basic convolutional (CONV) layer, one batch normalization (BN), and one rectified linear unit (RELU) sequentially stacked in each CNN layer. In this work, 1D CNN is applied. The convolution operation in each CNN layer can be summarized as follows:

$$\begin{aligned} \mathbf{Y} &= CONV(\mathbf{X}, \mathbf{W}, \mathbf{b}) \\ \mathbf{Y} &= BN(\mathbf{Y}) \\ \mathbf{Y} &= ReLU(\mathbf{Y}) \end{aligned} \tag{4}$$

where $\mathbf{X} \in \mathbb{R}^{T\times M}$ is the weighted input, $\mathbf{Y}$ is the intermediate features, and $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters in the 1D CNN. The CNN block is constructed by sequentially stacking several CNN layers. The optimized layers of the CNN block are discussed in Section IV.

D. Icing Detection of Wind Turbine Blades

In this work, the icing detection of wind turbine blades is considered to be a binary classification problem. The feature representations obtained by the TACNN are fed into a global average pooling (GAP) layer and an FC network. The softmax function is utilized in the output layer to output a probabilities for icing and normal (non-icing) statuses.

Assuming the features obtained by the TACNN are $\mathbf{X} \in \mathbb{R}^{K\times T}$, where K is the number of filters in the CNN layer and T is the length of the time series, the output probability $P_k \in [0, 1]$ of the corresponding icing condition for icing and normal statuses can be computed as follows:

$$P_k = softmax(\Phi(GAP(s))) = \frac{exp(\Psi^k \Phi(GAP(s)))}{\sum_{k=1}^{n} exp(\Psi^k \Phi(GAP(s)))} \tag{5}$$

where Φ is the parameter of the FC network, Ψ denotes the parameter of the output layer, and $\sum_{k=1}^{n} P_k = 1, k \in [0, 1]$. It is worth noting that these parameters, Φ and Ψ, are automatically updated and optimized during training on the basis of training samples.

In this paper, the back-propagation algorithm is utilized to train the model in gradient descent and the cross entropy is chosen as the loss function. The Adam algorithm is used to optimize the loss function to achieve efficient calculation and minimize memory usage.

IV. Experiments and Discussion

All experiments are implemented on a server equipped with Intel processors (64GB) and TITAN V (12GB). Pytorch is used for the implementation of the models. Throughout the training process, the learning rate is set to 1e-4.

A. Benchmark Comparison

The proposed TACNN can be considered as a general solution for TSC. Thus, the TACNN is firstly evaluated in ten public benchmark datasets [34]. The ten datasets consist

TABLE I: Accuracy Comparison in UEA Multivariate Time Series Dataset.

Dataset	TACNN	TapNet	MLSTM -FCN	WEASEL +MUSE	ED -1NN	DTW- 1NN-1	DTW- 1NN-D	ED-1NN (norm)	DTW–1NN -I(norm)	DTW–1NN -D(norm)
ArticularyWordRecognition	0.983	0.987	0.973	**0.99**	0.97	0.98	0.987	0.97	0.98	0.987
AtrialFibrillation	**0.467**	0.333	0.267	0.333	0.267	0.267	0.2	0.267	0.267	0.22
BasicMotions	0.975	**1**	0.95	**1**	0.675	**1**	0.975	0.676	**1**	0.975
FaceDetection	**0.629**	0.556	0.545	0.545	0.519	0.513	0.529	0.519	0.5	0.529
HandMovementDirection	**0.446**	0.378	0.365	0.365	0.279	0.306	0.231	0.278	0.306	0.231
Heartbeat	**0.756**	0.751	0.663	0.727	0.62	0.659	0.717	0.619	0.658	0.717
NATOPS	**0.961**	0.939	0.889	0.87	0.86	0.85	0.883	0.85	0.85	0.883
PenDigits	**0.988**	0.98	0.978	0.948	0.973	0.939	0.977	0.973	0.939	0.977
SelfRegulationSCP2	**0.572**	0.55	0.472	0.46	0.483	0.533	0.539	0.483	0.533	0.539
StandWalkJump	**0.533**	0.4	0.067	0.333	0.2	0.333	0.2	0.2	0.333	0.2
Average Value	**0.731**	0.687	0.617	0.657	0.585	0.638	0.624	0.584	0.637	0.626
Wins&Ties	**8**	1	0	1	0	1	0	0	1	0

of various applications, such as FaceDetection, Heartbeat, and PenDigits. The number of output classes also varies. For example, FaceDetection has only two classes, while ArticularyWordRecognition has 25. The size of the ten datasets also varies from hundreds of kilobytes to several megabytes. Classification accuracy is employed as the evaluation metric. The average accuracy and the number of Wins/Ties are calculated for the comparison of different approaches.

The state-of-the-art approaches used for evaluation are as follows: 1) **TapNet** [35]: TapNet is a recently proposed model based on prototypes for TSC. The TapNet can extract features automatically by utilizing its parallel structure of LSTM and CNN, and then identify the class according to the distance from the extracted features to the prototype. 2)**MLSTM-FCN** [28]: This model consists of two parallel structures, LSTM and FCN, and the FCN is equipped with an SE module. 3) **WEASEL-MUSE** [36]: WEASEL-MUSE is a feature-based model for TSC. 4) **ED-1NN**, **DTW-1NN-I**, **DTW-1NN-D**, **ED-1NN(norm)**, **DTW-1NN-I (norm)**, and **DTW-1NN-D (norm)**: In these methods, Euclidean distance (ED) and dynamic time warping (DTW) represent two distance measurement methods. I means that the DTW treats each dimension individually and D denotes that the data normalization is applied. The hyper-parameters of the TACNN are as follows: the number of hidden units in LSTM $M = 64$, and three layers of CNN are utilized based on the number of filters $\{128, 256, 128\}$. Only one FC layer is used in the icing detection network in this experiment. The best accuracy for each dataset is denoted with boldface.

In terms of average accuracy, the proposed TACNN outperforms all state-of-the-art methods. The TACNN gets the best average accuracy of 0.731, a significant improvement over the existing state-of-the-art approach, TapNet, with an average accuracy of 0.687. In terms of the number of wins/ties, our model achieves eight, which is the best among the nine methods, while TapNet and WEASEL+MUSE achieve only one win or tie each. These results suggest that our model can achieve better performance in most datasets, especially in those datasets with small amounts of data such as Heartbeat and HandMovementDirection, which only contain hundreds of training samples.

TABLE II: SCADA Data Specification

No.	Parameter	No.	Parameter
1	Wind speed	14	Temperature of pitch motor 1
2	Generator speed	15	Temperature of pitch motor 2
3	Active power	16	Temperature of pitch motor 3
4	Wind direction	17	Horizontal acceleration
5	Average wind direction angle within 25s	18	Vertical acceleration
6	Yaw position	19	Environment temperature
7	Yaw speed	20	Internal temperature of nacelle
8	Angle of pitch 1	21	Switching temperature of pitch 1
9	Angle of pitch 2	22	Switching temperature of pitch 2
10	Angle of pitch 3	23	Switching temperature of pitch 3
11	Speed of pitch 1	24	DC power of pitch 1 switch charger
12	Speed of pitch 2	25	DC power of pitch 2 switch charger
13	Speed of pitch 3	26	DC power of pitch 3 switch charger

B. Dataset and Evaluation Metrics

In this work, icing data for the wind turbine blades are obtained from Goldwind Inc., one of the largest manufacturers of wind turbines in the world. We have access to the operational data of three wind turbines logged by the SCADA system in Inner Mongolia, China. The recorded running times for the three turbines are 305.77, 695.59, and 329.28 hours, respectively.

There are some interruptions in the first two machines due to the stop, and only the last one machine has continuous logging. The raw sensor data collected by the SCADA system include hundreds of dimensions. Only 26 parameters related to icing blades are left; the remaining parameters are removed by the engineer based on the domain knowledge. Furthermore, the engineers also help us label the range of icing occurring. The sensor data are highly imbalanced according to the labeled range, and thus the imbalanced data processing described in Section III-B is applied. The data from the two machines whose sensor data have interruptions are mixed up for offline training and testing. The ratios for training and testing are approximately 80% and 20%, respectively. The data from the machine with continuous logging are utilized in the online estimation of icing probability.

The following metrics are used to evaluate the models: including **Precision**, **Recall**, **F1**, and **Matthews correlation coefficient (MCC)**. The definitions of these metrics are presented as follows:

$$\mathit{Precision} = \frac{TP}{TP+FP} \quad \mathit{Recall} = \frac{TP}{TP+FN} \tag{6}$$

$$\mathit{F1} = \frac{2 \times \mathit{Precision} \times \mathit{Recall}}{\mathit{Precision} + \mathit{Recall}} \tag{7}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{8}$$

where TP, FP, FN, and TN represent true positive, false positive, false negative and true negative, respectively.

C. Baseline Comparison

To verify the performance of our proposed approach to icing detection on wind turbine blades, we compare ten baseline networks. The details regarding the ten baseline networks are as follows:

- **MultiLayer Perceptron (MLP)**: There are three FC layers with 500 hidden nodes in each layer. The dropout layer is employed between FC layers.
- **Long Short-Term Memory (LSTM)**: LSTM is a powerful tool for the modeling of time series data. The LSTM used for comparison in this work is only one layer. The number of hidden units is selected from $\{8, 16, 32, 64\}$, and the best performance model is chosen for comparison.
- **Gated Recurrent Unit (GRU)**: GRU is a light-weight variant of LSTM, but its performance is not inferior to LSTM. The number of hidden units for GRU is also selected from $\{8, 16, 32, 64\}$, and the best performance model is chosen for comparison.
- **CNN**: A CNN layer is utilized with the parameters and structure described in Section III-C.
- **FCN**: FCN shows competitive performance in TSC [27]. There are three convolutional layers with filter size of $\{$128, 256, 128$\}$ in this experiment.
- **ResNet**: ResNet is a variant of CNN which is widely used in TSC. The ResNet used for comparison is adopted from [27].
- **WaveletFCNN**: WaveletFCNN is a deep neural network especially designed for blade icing detection in wind turbines. WaveletFCNN integrates the wavelet transformation and fully convolutional neural network (FCNN), which automatically learns useful features from raw sensor data. We use the same settings as described in [10].
- **MSCNN**: is a novel multi-scale deep CNN for fault diagnosis in wind turbine gearboxes. The model is re-implemented according to the settings in this research. In the present work, three scales are considered in the MSCNN [16].

The **MLSTM-FCN** and **TapNet** are described in Section IV-A. Our proposed model is an end-to-end deep learning model. Thus, only the two deep learning models (i.e., **MLSTM-FCN** and **TapNet**) described in Section IV-A are used for comparison in this section. The hyper-parameters of the TACNN are also the same as in Section IV-A, except that there are three FC layers in the icing detection network in this experiment. The window size (T) for the sensor data is 128 (almost 15 mins).

TABLE III: Performance of baseline comparison on SCADA data

Models	Recall	Precision	F1	MCC
MLP	0.607	0.718	0.714	0.420
LSTM	0.659	0.863	0.799	0.601
GRU	0.594	0.851	0.770	0.544
CNN	0.556	0.851	0.756	0.520
FCN	0.730	0.879	0.833	0.665
ResNet	0.885	0.778	0.834	0.674
TapNet	0.710	0.776	0.777	0.547
MLSTM-FCN	0.657	0.833	0.786	0.570
WaveletFCNN	0.445	**0.920**	0.732	0.498
MSCNN	0.687	0.837	0.798	0.594
TACNN	**0.922**	0.850	**0.891**	**0.784**

As illustrated in TABLE III, our proposed model achieves better results than did the other baseline methods. In terms of MCC and F1, our model has almost 16.3% and 6.83%, respectively, improvements over the best result achieved by ResNet. Compared with the worst-performing model, there is an improvement of 24.8% and 86.7% regarding to F1 and MCC, respectively. LSTM and GRU are slightly better than MLP due to their ability to learn periodic features. As for Recall, our proposed model is the only one whose accuracy is greater than 90%. However, WaveletFCNN achieves better accuracy in Precision than our proposed TACNN. ResNet and FCN obtain almost the same levels of accuracy as the F1 and MCC metrics. Surprisingly, TapNet and MLSTM-FCN do not achieve the desired accuracy. The reason for this may be that there are numerous hyper-parameters that would have needed to be optimized to obtain higher accuracy. In addition, these two methods are designed for multivariate TSC problems rather than the field of wind turbines. WaveletFCNN and MSCNN are specially designed multi-scale deep neural networks intended for use in fault diagnosis in wind turbines. Compared with WaveletFCNN and MSCNN, the performance of the TACNN has improved by 57.4% and 32.0% for MCC, respectively. For F1, there was an improvement of 21.7% and 11.7% over these two methods (WaveletFCNN and MSCNN, respectively). Importantly, WaveletFCNN and MSCNN belong to the concept of multi-scale networks. WaveletFCNN uses wavelets to generate multi-scale features, while multi-scale entropy is utilized for multi-scale feature generation by MSCNN. The performances of WaveletFCNN and MSCNN are not as expected. One possible reason is that we only select a specific variant for comparison (the scale is set to three in both methods). Furthermore, MSCNN is designed for health monitoring of wind turbine gearboxes and may not be able to fully mine the discriminative features from the sensor data for blade icing.

D. Comparison with Other Attention Mechanisms

To further illustrate the performance of the proposed approach, three widely used attention mechanisms are used for comparison with the icing datasets, presented in TABLE IV.

TABLE IV: Comparison with other attention modules

Models	Recall	Precision	F1	MCC
SE	0.727	0.848	0.818	0.633
CBAM	0.863	0.780	0.829	0.660
GC	**0.996**	0.746	0.845	0.729
TACNN	0.922	**0.850**	**0.891**	**0.784**

For a fair comparison, these attention modules are attached after CNN layer. The structure and parameters for the CNN are the same as those used in TACNN. It is worth noting that there is one attention module attached to every CNN layer. The details of the attention modules used are as follows. **SE** [37]: is a famous attention module proposed for CNN. **CBAM** [31]: is sequentially comprised of channel and spatial attention module. **GC** [32]: is a lightweight attention module that can model the global context.

TABLE IV clearly indicates that the proposed TACNN outperforms the other attention modules in terms of MCC, F1, and Precision. GC achieved the second-best performance among the three attention modules because it is equipped with better feature learning capabilities. Specifically, the proposed TACNN has 23.9%, 18.8%, and 7.54% better performances than the SE, CBAM, and GC for MCC, respectively. With respect to F1, SE and CBAM obtain almost the same classification accuracy, and there is approximately a 5.44% improvement between the TACNN and GC. Interestingly, for Recall, the GC achieves almost a 100% accuracy, which is better than the proposed TACNN. We believe that the reason for this is that the GC has the ability to explore the relationships among the features extracted in a global context way.

The main difference between our proposed TACNN and these widely used attention modules is that our proposed attention mechanism is faced with the raw sensor data, which is applied before the CNN, while these widely used attention mechanisms reweight the features extracted by the CNN. The experimental results indicate that the proposed TACNN can effectively improve performance. The reason might be that our proposed model can directly evaluate the importance of sensors and time steps, while these other attention mechanisms used for TSC can not. In short, these attention mechanisms need to rely on the features extracted by the CNN, while our method first processes the sensor data and then is fed into the CNN. In summary, the superiority and significance of the proposed TA module relate to 1) reinforcing the icing status learning mechanism and 2) exploiting the discriminant feature learning mechanism.

E. Ablation Study and Sensitivity Analysis

An ablation study is conducted to illustrate the importance of the proposed TA module. To perform the ablation study, the TA module is removed, and thus it is called **TACNN_TA**. The performance of TACNN and TACNN_TA is compared in four different datasets when the window size is set to {32, 64, 128, 256}. As shown in Fig. 3, the F1 and MCC are presented for TACNN and TACNN_TA within this different window size. From Fig. 3, we can observe that 1) TA do improve the performance of traditional CNN, 2) the biggest accuracy drop happens when the window size is 128, and 3) with the change in window size, the performance of TACNN increases and then decreases. TACNN achieves the best performance when the window size is 128. However, the performance of the CNN gradually declines. In sum, the proposed TACNN is determined to be sensitive to the length of the time series.

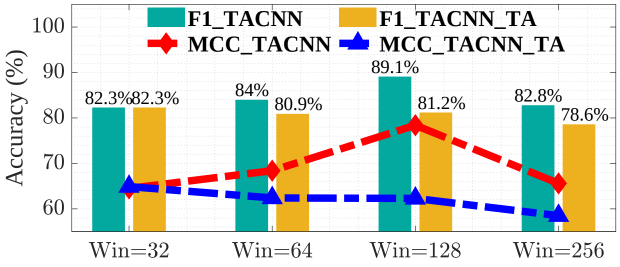


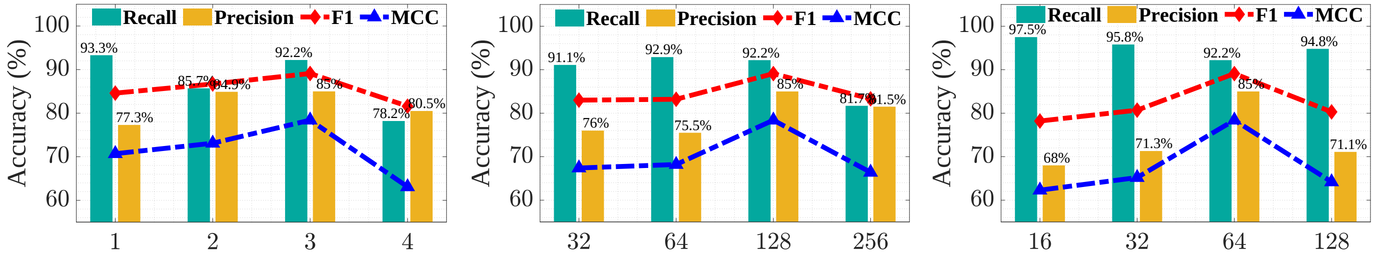
Fig. 3: Ablation analysis.

To study the influence of important factors on the TACNN and optimize its structure, a sensitivity analysis is performed. The results are presented in Fig. 4. To explore the impact of the number of CNN layers in TACNN, four different networks are created. The number of filters for the four networks are $\{128\}$, $\{128, 256\}$, $\{128, 256, 128\}$, and $\{128, 256, 128, 128\}$, respectively. To investigate the influence of filter size in TACNN, we set the number of layers to three, and vary the filter size as such: {32, 64, 128, 256}. To understand the importance of hidden units in TA module, we set the layers of CNN to three with a filter size of $\{128, 256, 128\}$ and vary the number of hidden units from 16 to 128.

As illustrated in Fig. 4a, the highest Precision, F1, and MCC values occur when there are three CNN layers. It also can be seen from Fig. 4a that the worst performance happens when the number of CNN layers is four and not one. The explanation for this result is the impact of the number of filters. If we optimize the number of filters, the results might change. As in Fig. 4b, the best Precision, F1, and MCC can also be found when the filter size of the CNN is 128. There is a trend in the accuracy increasing for both F1 and MCC when the filter size of the CNN increases from 32 to 128. The performance decreases when the filter size is 256. As depicted in Fig. 4c, the best performance occurs when the number of hidden units of the TA module is 64. There is also a trend in which the accuracy increases when the number of hidden units in TA module increases from 16 to 64. Then, the performance significantly drops when the number of hidden units is 128 and grow even worse than when the number of hidden units is 32.

F. Online Estimation

An online estimation scheme is proposed to provide real-time identification of the icing conditions of wind turbine blades on wind farms. The model is trained in an offline fashion as presented in the previous section, beginning with segmenting the historical sensor data into a fixed window size to train the icing detector. For the online identification scheme, a sliding window with the same length as the segments



(a) (b) (c)

Fig. 4: SA of (a) CNN layers, (b) filters size in CNN, and (c) hidden units in TA.

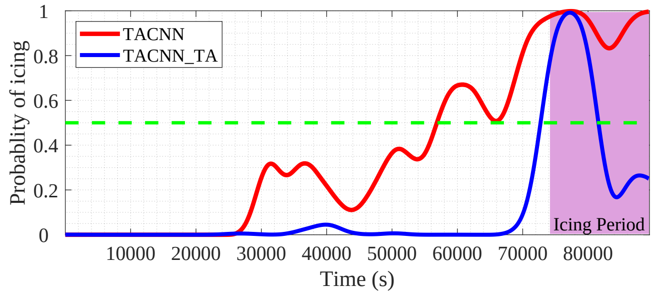


Fig. 5: Results of online estimation of TACNN and TACNN_TA.

moves as new sensor data become available. Then the trained model provides a predicted probability of icing and normal conditions. The majority vote algorithm is utilized for the real-time estimation of icing condition. It is verified that the majority vote algorithm can ensure robustness and eliminate accidental errors [10]. As mentioned in Section III-B, this machine, along with continuous logging sensor data , can be utilized for online icing detection. In this experiment, the hyper-parameters are the same as with previous experiments. The model is able to predict output approximately every 4 minutes and the entire testing time is around 23 hours.

To further illustrate the performance of the proposed TA module, the online estimation results of the TACNN and TACNN_TA are presented in Fig. 5. The magenta background indicates the icing period (i.e., when icing has occurred), a condition that is labeled by the domain engineer. The green dash line presents the classical threshold (0.5) for icing detection. During the icing period, it is easy to ensure that the proposed TACNN achieves 100% accuracy, while the TACNN_TA has only approximately 50% accuracy. During normal condition, the output probability of the TACNN increases gradually, while the TACNN_TA changes suddenly. This suggests that these results can be attributed to the proposed TA module, which employs the relationships between time steps. Ice accumulation, of course, is generally a slow process. Compared with the CNN, our proposed model allows engineers more time to prevent icing, facilitating mitigation in advance of negative outcomes.

There are only two states (i.e., normal/icing) in recorded SCADA data; the ice accretion on blades is a gradual process with various intensities, which makes it difficult for deep learning models to detect the severity of icing on wind turbine blades. In icing detection, it is of practical meaning to identify icing of wind turbine blades at an early stage. The simplest way is to interpret the output probability, which is defined in Eq. 5. In other words, we can set different thresholds to adjust the sensitivity of the proposed model for early icing identification. For example, if the threshold is set to 0.3, compared with the TACNN_TA, the proposed TACNN can identify icing at a very early stage, though there are some misidentifications during the normal period. If the threshold is 0.75, the proposed TACNN also achieves competitive performance both in the normal and icing periods. In practice, the threshold can be determined by observing the relationship between the threshold and amount of icing.

V. Conclusion

This research presents a novel deep learning network for identifying icing conditions for wind turbine blades. The proposed model integrates the traditional CNN with a TA module, with the goal of learning the importance of sensors and timesteps and automatically learning and discovering discriminative features from raw temporal sensor data. The effectiveness of the proposed TACNN in dealing with the four challenges mentioned in Section I has been demonstrated. The proposed TACNN can also be considered as a general solution for TSC, which is verified by ten public benchmark datasets. The TACNN has been applied to the icing datasets of three wind turbines obtained from one of the largest wind power companies in the world. Compared with ten state-of-the-art baseline networks, the TACNN shows significant advantages in terms of accuracy. Compared with three widely used attention mechanisms, the proposed model achieves competitive results. The generalization and practicability of the proposed model are further verified by online estimation testing.

It is worth noting that there are two assumptions for the proposed model: 1) Feature space and label space of training and testing samples should be the same, that is, the window size and sampling frequency of training and testing samples should be the same. 2) The distribution of training and testing samples also should be the same. There are some limitations on the proposed TACNN. The first is that it does not indicate the severity of the icing on wind turbine blades. One potential solution may be structural modification of the proposed model to ensure robust feature learning of icing conditions. The second limitation is that the proposed TACNN relies on individual training for every individual wind turbine, limiting generalization of the TACNN. As transfer learning can be used to extrapolate to other sizes of wind turbines, the application and improvement of the proposed model for domain adaptability of different types of wind turbines should also be further explored. The third limitation is that it is dangerous to blindly apply the TACNN to vibrating systems. There is no guarantee that the underlying model necessarily aligns with the dynamics, rendering extrapolation difficult. A possible solution is to combine the mathematical modeling methods or Gaussian processes [38] to capture the dynamics in an interretable or easy to supervise fashion. Future work should also study the hyper-parameters optimization to obtain the best network structure.

References

[1] Z. Ren, A. S. Verma, Y. Li, J. J. Teuwen, and Z. Jiang, "Offshore wind turbine operations and maintenance: A state-of-the-art review," *Renewable and Sustainable Energy Reviews*, vol. 144, p. 110886, 2021.

[2] Y. Wang, Y. Xu, and Y. Lei, "An effect assessment and prediction method of ultrasonic de-icing for composite wind turbine blades," *Renewable Energy*, vol. 118, pp. 1015–1023, 2018.

[3] K. Wei, Y. Yang, H. Zuo, and D. Zhong, "A review on ice detection technology and ice elimination technology for wind turbine," *Wind Energy*, vol. 23, no. 3, pp. 433–457, 2020.

[4] E. Madi, K. Pope, W. Huang, and T. Iqbal, "A review of integrating ice detection and mitigation for wind turbine blades," *Renewable and Sustainable Energy Reviews*, vol. 103, pp. 269–281, 2019.

[5] M. Zhang, J. Yu, R. Chen, Q. Liu, J. Liu, D. Song, P. Liu, L. Gao, and J. Wang, "Highly transparent and robust slippery lubricant-infused porous surfaces with anti-icing and anti-fouling performances," *Journal of Alloys and Compounds*, vol. 803, pp. 51–60, 2019.

[6] O. Fakorede, Z. Feger, H. Ibrahim, A. Ilinca, J. Perron, and C. Masson, "Ice protection systems for wind turbines in cold climate: characteristics, comparisons and analysis," *Renewable and Sustainable Energy Reviews*, vol. 65, pp. 662–675, 2016.

[7] C. Q. G. Muñoz, F. P. G. Márquez, and J. M. S. Tomás, "Ice detection using thermal infrared radiometry on wind turbine blades," *Measurement*, vol. 93, pp. 157–163, 2016.

[8] J. Zeng and B. Song, "Research on experiment and numerical simulation of ultrasonic de-icing for wind turbine blades," *Renewable Energy*, vol. 113, pp. 706–712, 2017.

[9] L. Gao, Y. Liu, L. Ma, and H. Hu, "A hybrid strategy combining minimized leading-edge electric-heating and superhydro-/ice-phobic surface coating for wind turbine icing mitigation," *Renewable energy*, vol. 140, pp. 943–956, 2019.

[10] B. Yuan, C. Wang, F. Jiang, M. Long, P. S. Yu, and Y. Liu, "Waveletfcnn: A deep time series classification model for wind turbine blade icing detection," *arXiv preprint arXiv:1902.05625*, 2019.

[11] L. Shu, G. Qiu, Q. Hu, X. Jiang, G. McClure, and H. Yang, "Numerical and field experimental investigation of wind turbine dynamic de-icing process," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 175, pp. 90–99, 2018.

[12] A. A. Jiménez, F. P. G. Márquez, V. B. Moraleda, and C. Q. G. Muñoz, "Linear and nonlinear features and machine learning for wind turbine blade ice detection and diagnosis," *Renewable Energy*, vol. 132, pp. 1034–1048, 2019.

[13] L. Zhang, K. Liu, Y. Wang, and Z. B. Omariba, "Ice detection model of wind turbine blades based on random forest classifier," *Energies*, vol. 11, no. 10, p. 2548, 2018.

[14] H. Badihi, Y. Zhang, and H. Hong, "Fault-tolerant cooperative control in an offshore wind farm using model-free and model-based fault detection and diagnosis approaches," *Applied Energy*, vol. 201, pp. 284–307, 2017.

[15] F. Cheng, J. Wang, L. Qu, and W. Qiao, "Rotor-current-based fault diagnosis for dfig wind turbine drivetrain gearboxes using frequency analysis and a deep classifier," *IEEE transactions on industry applications*, vol. 54, no. 2, pp. 1062–1071, 2017.

[16] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3196–3207, 2018.

[17] L. Saidi, J. B. Ali, E. Bechhoefer, and M. Benbouzid, "Wind turbine high-speed shaft bearings health prognosis through a spectral kurtosis-derived indices and svr," *Applied Acoustics*, vol. 120, pp. 1–8, 2017.

[18] S. Chang, M. Leng, H. Wu, and J. Thompson, "Aircraft ice accretion prediction using neural network and wavelet packet transform," *Aircraft Engineering and Aerospace Technology: An International Journal*, 2016.

[19] E. Ogretim, W. Huebsch, and A. Shinn, "Aircraft ice accretion prediction based on neural networks," *Journal of Aircraft*, vol. 43, no. 1, pp. 233–240, 2006.

[20] H. H. Yildirim and M. Yavuz, "Evaluation of wind energy investment with artificial neural networks," *An International Journal of Optimization and Control: Theories & Applications (IJOCTA)*, vol. 9, no. 2, pp. 142–147, 2019.

[21] Y. Liu, H. Cheng, X. Kong, Q. Wang, and H. Cui, "Intelligent wind turbine blade icing detection using supervisory control and data acquisition data and ensemble deep learning," *Energy Science & Engineering*, vol. 7, no. 6, pp. 2633–2645, 2019.

[22] S. SHOJA, "Guided wave propagation in composite structures," Master's thesis, Chalmers University of Technology, 2016.

[23] C. Orsenigo and C. Vercellis, "Combining discrete svm and fixed cardinality warping distances for multivariate time series classification," *Pattern Recognition*, vol. 43, no. 11, pp. 3787–3794, 2010.

[24] M. G. Baydogan, G. Runger, and E. Tuv, "A bag-of-features framework to classify time series," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2796–2802, 2013.

[25] P. Schäfer, "Scalable time series classification," *Data Mining and Knowledge Discovery*, vol. 30, no. 5, pp. 1273–1298, 2016.

[26] W. Pei, H. Dibeklioğlu, D. M. Tax, and L. van der Maaten, "Multivariate time-series classification using the hidden-unit logistic model," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 920–931, 2018.

[27] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1578–1585. IEEE, 2017.

[28] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate lstm-fcns for time series classification," *Neural Networks*, vol. 116, pp. 237–245, 2019.

[29] X. Cheng, G. Li, R. Skulstad, S. Chen, H. P. Hildre, and H. Zhang, "Modeling and analysis of motion data from dynamically positioned vessels for sea state estimation," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6644–6650. IEEE, 2019.

[30] X. Cheng, G. Li, A. L. Ellefsen, S. Chen, H. P. Hildre, and H. Zhang, "A novel densely connected convolutional neural network for sea state estimation using ship motion data," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 5984–5993, 2020.

[31] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[32] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[33] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in *IJCAI*, 2017.

[34] A. Bagnall, H. A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, and E. Keogh, "The uea multivariate time series classification archive, 2018," *arXiv preprint arXiv:1811.00075*, 2018.

[35] X. Zhang, Y. Gao, J. Lin, and C.-T. Lu, "Tapnet: Multivariate time series classification with attentional prototypical network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[36] P. Schäfer and U. Leser, "Multivariate time series classification with weasel+ muse," in *Proceedings of ACM Conference*, pp. 0–0, 2017.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[38] W. Weijtjens, L. D. Avendaño-Valencia, C. Devriendt, and E. Chatzi, "Cost-effective vibration based detection of wind turbine blade icing from sensors mounted on the tower," in *9th European Workshop on Structural Health Monitoring, EWSHM 2018*, 2018.