Condition Monitoring of Wind Turbine Converters with Limited and Unbalanced SCADA Data

Borong Hu[1], Chunjiang Jia[1], Chi Zhang[1], Subhash Lakshminarayana[1], Biyun Chen[1], Paul Mckeever[1], Chong Ng[1], Teng Long[1], and Li Ran[1]

[1]Affiliation not available

July 02, 2024

Condition Monitoring of Wind Turbine Converters with Limited and Unbalanced SCADA Data

Borong Hu, Chunjiang Jia, Chi Zhang, Subhash Lakshminarayana, Biyun Chen, Paul McKeever, Chong Ng, Teng Long, Li Ran

Abstract- **This paper proposes a condition monitoring (CM) framework to detect the health-to-fault operational behavior of wind turbine (WT) converters based on SCADA data from early-stage operation. A novel method for characterizing operating data distribution is proposed to exploit the limited and unbalanced SCADA data. By providing weighting factors for cost functions, this method significantly improves the robustness and generalization of the CM method. The CM framework is also enabled for the full life cycle including early-stage and long-term operation, by employing an online learning algorithm. Validation with both healthy and faulty WTs demonstrates the potential to detect abnormality a few days before actual converter faults occur.**

***Index Terms* – Condition monitoring, wind turbine converters, SCADA, neural network, data-driven.**

I. Introduction

Wind energy is one of the most promising renewable solutions for achieving low carbon emissions. Wind turbine (WT) manufacturers are increasingly adopting new power semi-conductors and converter techniques to improve the power rating and efficiency. However, the difficulty and high expense of maintenance demand for developing field-deployable condition monitoring (CM) methods suitable for the full life cycle including early-stage and long-term operation.

In addition to the slow aging process driving to the wear-out phase of the bathtub curve, early faults in the “infant fatality” phase also contribute to high failure rates due to component defects, immature design, or improper assembly [1]. However, existing CM methods, based on thermal networks, electrical signals, or TSEPs (temperature-sensitive electrical parameters) [2], are primarily designed to detect long-term gradual aging processes. These methods often depend on the degradation mechanism learned from a large quantity of model-based or sample tests, requiring complex modeling processes and additional sensor installations. In addition, multi-physics domain modeling is costly and time-consuming [3], posing challenges for online applications. To address these challenges, this paper introduces a data-driven approach to detect the health-to-fault operational behavior of WT converters, utilizing limited and unbalanced SCADA (Supervisory Control and Data Acquisition) data from early-stage operations.

Data-driven solutions have demonstrated superior abilities in modeling and characterizing complex systems, owing to groundbreaking developments in AI technology. Various intelligent CM methods using SCADA data have shown promise in monitoring mechanical components [4], but their application to WT converters remains limited. SCADA data usually provides limited information in terms of fault patterns and the degradation process of converters, particularly in the early-stage operation. This limitation hinders the training of supervised learning CM methods, which require detailed information as training labels [5]. Unsupervised learning, on the other hand, possesses the potential to estimate changes in converter operation characteristics relative to a baseline without the necessity for training labels of fault conditions. This can be achieved by training deep learning neural network (DNN) architectures, including autoencoders, fully connected neural networks (FCNN), and long short-term memory (LSTM) networks, given a sufficient amount of data [6-8]. However, in WT systems, early-stage SCADA data often fails to cover the full range of operating scenarios, like varying wind speeds and environmental temperatures. In addition, the data is frequently skewed towards lower power regions due to frequent trial runs and debugging tests of WT converters, resulting in an unbalanced distribution of operating points. Data distribution could be enhanced by various approaches, such as scaling, rotating and color shifting for image processing [9], but these techniques are not directly applicable to SCADA data. Moreover, conventional cost functions such as mean absolute error (MAE) and mean square error (MSE) inherently assign equal weight to all data points and can reduce the generalization capability of neural networks due to bias from unbalanced data distributions, leading to more accurate predictions around operating points with abundant data than those with less. The CM method, however, is required to capture the health state accurately no matter what the converter operating points are.

This paper proposes an online deployable CM method for WT converters, tailored for the full life cycle and based on limited and unbalanced SCADA data. An unsupervised CM framework is presented to detect the health-to-fault operational behavior of WT converters, with no need to label health states. Additionally, this paper introduces a novel method for characterizing operating data distribution, which provides weighting factors for cost functions to mitigate the impact of data imbalance on model training and generalization. An online learning algorithm is implemented to continually update the framework with newly logged real-time data. The effectiveness of the method is demonstrated through early-stage fault detection, with the potential to extend its application to long-term, full life cycle CM.

II. Condition Monitoring Framework

The proposed condition monitoring framework comprises two main processes: modeling and condition monitoring, as shown in Fig. 1. The modeling process uses SCADA data from a healthy operating period to train a DNN to capture the healthy state characteristics of the WT converter. In the CM process, each newly logged real-time data is fed into the trained model whose prediction error indicates to what extent the WT converter has deviated from the healthy state, i.e., the health-to-fault process. The development of the framework contains the following steps:

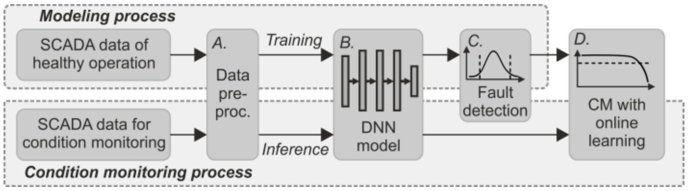


Fig. 1. Condition monitoring framework

A. Data pre-processing

The SCADA system in our case study records a wide range of physical measurements across aerodynamic, electrical, and thermal domains. This raw data, logged every 10 minutes, undergoes a pre-processing phase where channels with null ('zero' or 'NA') values are removed. The samples during WT shutdown periods are also removed to focus on operating behavior. Each channel's data is normalized using z-score normalization. Specific details about the wind farm, turbine, and power converter types are omitted for confidentiality.

B. Network model

The network aims to differentiate the faulty state from a healthy reference by analyzing variations in thermal and electrical measurements during failure development. The thermal response, influenced by electrical operating points and cooling conditions, is crucial for fault characterization. Thus, a regression neural network is proposed to model the thermal response, correlating converter temperature measurements as output channels with the other SCADA measurements as input channels. The number of SCADA measurement channels is donated as *n*.

This study selects two common network architectures for comparison analysis, as shown in Fig. 2. The first is a fully connected neural network (FCNN) consisting of three hidden layers to extract the complex thermal response. It uses ReLU as the activation function and Adam optimizer [10]. The second architecture is an LSTM type, designed to capture the aerodynamic inertia and thermal capacitance effects from time sequence features. In alignment with the system's thermal time constant, the input data is resampled by a 12-hour rolling window, i.e., 72 SCADA samples, into a 2-D sequence format. A sequence-to-one LSTM is then established to predict the converter temperature at the final timestamp within the rolling window.

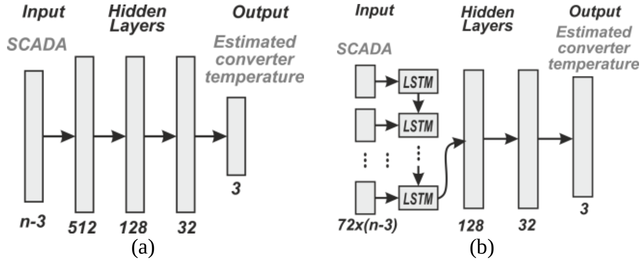


Fig. 2. Network structure of (a) FCNN and (b) LSTM

C. Fault detection criterion and indicator

When a fault occurs in the converter, the network's ability to accurately represent its thermal characteristics diminishes. Thus, the percentage accuracy, *A*, has potential signatures for fault detection and can be calculated as

$$A = \left(1 - \frac{(y-t)}{t}\right) \times 100\% \qquad (1)$$

where *y* is the network output prediction, *t* is the actual measured data from SCADA.

As the model training removes the dependency on the operating point by using the customized cost function in Section III, the distribution of *A* from a healthy period is statistically close to a normal distribution, which will be presented in Section IV. If the converter remains healthy, the distribution of *A* in both modeling and CM processes should be identical. This can be validated by two-sample *t*-test [11]. A rejection from the test indicates that the converter has deviated from its healthy state and is subjected to degradation. Given that the degradation increases monotonically, considering fault detection sensitivity, the ‘2-σ’ boundary of the *A*’s distribution is designated as the fault alarm threshold.

Meanwhile, squared error accuracy, *SA*, is used to derive the fault detection indicator, as it provides exclusively positive values, enabling unidirectional judgment:

$$SA = \left(1 - \frac{(y-t)^2}{t^2}\right) \times 100\% \qquad (2)$$

Furthermore, to detect the monotonical trend with a better signal-to-noise ratio, a 30-day moving average is applied to *SA*. The resulting smoothed square accuracy, *SSA*, is used as the indicator of the converters’ health. During the CM process, the *SSA* should be close to 100 for healthy WT converters. In the event of a fault, the *SSA* will drop below the threshold defined by the ‘2-σ’ boundary of *A*’s distribution.

D. CM with online learning

An online learning method is put forward to achieve real-time condition monitoring during long-term operations. Initially, the model undergoes training with a substantial amount of SCADA data, preparing it to forecast the converter's performance for the ensuing 30-day period, referred to as the CM process. With each new 10-minute interval, as fresh SCADA data is logged, the oldest data point in the CM rolling window is transferred to a designated data pool reserved for online learning. This pool, combined with the previous training dataset, is utilized to re-train the model after each online learning period, e.g., a 30-day cycle set to the same length as the CM window. As a result, the model undergoes periodic updates to perform CM on the upcoming 30-day period. Notably, both the frequency of the online re-training and the duration of the CM rolling window are adjustable individually, allowing customization to meet specific system requirements.

E. Data sufficiency analysis

It is also essential to evaluate whether the training data is sufficient for the network to accurately depict a healthy WT converter and to ensure robust generalization ability. This involves ensuring consistent prediction performance across both training and testing datasets. The consistency of this performance can be quantified using the consistency indicator *C*, defined as:

$$C = \frac{R_{tst}^2}{R_{trn}^2} \times \left(1 - \frac{|\sigma_{err,tst} - \sigma_{err,trn}|}{\sigma_{err,trn}}\right) \qquad (3)$$

where $\sigma_{err,tst}$ and $\sigma_{err,trn}$ represent the standard deviation of the prediction error on the testing and training datasets, respectively. The R^2_{tst} and R^2_{trn} donate the prediction performance on testing and training data, respectively, defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{size}(y_i - t_i)^2}{size \times Var(t)} \qquad (4)$$

where $Var(t)$ is the variation of actual measurement and *size* is the number of data points in the testing or training dataset. Therefore, the closer the value of *C* approaches 1, the stronger the network's generalization ability is inferred. A data sufficient analysis will be conducted across various DNNs, to assess their generalization capabilities and data volume requirements. The analysis involves incrementally increasing the training data by 30-day accumulation intervals and calculating the value of *C* at each step to determine how the amount of training data influences the model's performance. The results will be discussed in Section IV.

III. Characterization of Operating Data Distribution

In order to mitigate the data imbalance on regression model training, this paper proposes a novel method to characterize the operating data distribution and use the characterization method to generate probability density weights (PDW) to optimize two cost functions MSE and MAE. The characterization and PDW calculation process is shown in Fig. 3.

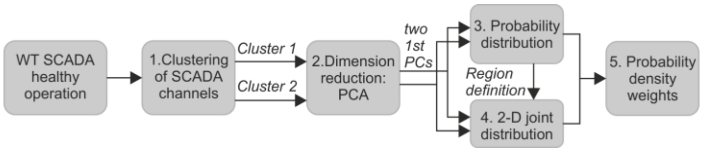

WT SCADA healthy operation
1.Clustering of SCADA channels
Cluster 1
Cluster 2
2.Dimension reduction: PCA
two 1st PCs
3. Probability distribution
Region definition
4. 2-D joint distribution
5. Probability density weights

Fig. 3. The flowchart of the characterization process

Step 1: It is first necessary to identify SCADA measurement channels that significantly influence the converter temperature channels. This is achieved by conducting hierarchical clustering analysis on all SCADA channels based on their pairwise distance:

$$dist(\boldsymbol{x}_p, \boldsymbol{x}_q) = \|\boldsymbol{x}_p - \boldsymbol{x}_q\|, p, q \in (1, \dots, n) \qquad (5)$$

where *dist* is the Euclidean distance between the p^{th} channel and q^{th} channel of the SCADA measurements. $\mathbf{x}_p$ and $\mathbf{x}_q$ are the normalized data in these two channels. Pairs of channels that are in the closest proximity will be linked as a cluster. The distance $d(r,s)$ between two clusters *r* and *s* is further calculated by the nearest neighbor method as

$$d(r, s) = min(\, dist(\boldsymbol{x}_{si}, \boldsymbol{x}_{tj})), i \in (1, \dots, n_s), j \in (1, \dots, n_t) \qquad (6)$$

where n_s (or n_t) is the number of channels in cluster *s* (or cluster *t*), and $\mathbf{x}_{si}$ (or $\mathbf{x}_{tj}$) is the i^{th} (or j^{th}) channel in cluster *s* (or cluster *t*). The closest clusters can be grouped further into larger clusters at the next level until all clusters of channels are linked together as a hierarchical tree by the agglomerative process.

Fig. 4 presents the dendrogram resulting from this analysis for a healthy WT's SCADA channels. The channels closer to the temperature channels have more influence, and based on their types of measurements, they are manually summarized into two clusters: 1st is cooling conditions and 2nd is electrical conditions.

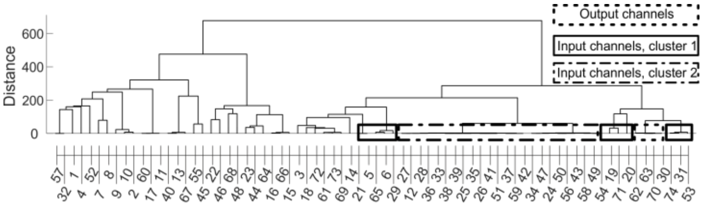


Fig. 4. The hierarchical clustering dendrogram of SCADA channels

Step 2: This step involves reducing the dimension (the number of channels) of these two identified clusters to a single dimension using principal component analysis (PCA) [12], respectively. The aim is to summarize each cluster's measurement information by a one-dimensional representation. PCA uses orthogonal linear transmission to find a coordinate wherein the scalar projection of time-sequence SCADA data, denoted as $\boldsymbol{X}$, exhibits the maximum variance compared to its projections on all other coordinates. To find the first principal components (1st PC), the weight matrix $\boldsymbol{W}$ has to satisfy the equation

$$\boldsymbol{W} = \underset{\|\boldsymbol{W}\|=1}{\operatorname{argmax}}\{\boldsymbol{W^T X^T X W}\} \qquad (7)$$

Then the 1st PC $\boldsymbol{T}$ can be calculated as

$$\boldsymbol{T} = \boldsymbol{X} \cdot \boldsymbol{W} \qquad (8)$$

The ratio of the variance in 1st PC to the total variance in the cluster's SCADA data quantifies the extent to which the 1st PC can represent the measurement information in each cluster. Remarkably, the explanation percentage exceeds 99% according to the results, indicating the operating conditions are effectively characterized by these two 1st PCs from the perspectives of cooling and electrical conditions, respectively.

Step 3: The cumulative density functions (CDF) of these two 1st PCs are calculated from their statistical distribution histograms, respectively. The PC's value where its CDF reaches 0.9 is selected as the threshold of the boundary to divide the uneven data distribution into two parts, dense and sparse.

Step 4: The distributions of these two 1st PCs can then be used to build a 2-D joint distribution. This aids in classifying data points into two distinct operating regions: region 1 with dense data or region 2 with sparse data. A SCADA data point will be allocated into region 1 if its 1st PCs of both cooling and electrical conditions are all lower than the predefined thresholds; otherwise, it will be placed into region 2.

Step 5: The probability density weight (PDW) for the SCADA data points falling into region *r* (1 or 2) is then derived as w_r,

$$w_r = \frac{\sum_{r=1}^{2} m_r}{m_r} \tag{9}$$

where m_r is the number of training data belonging to the region r. The SCADA dataset can be weighted accordingly since the basic MSE and MAE cost functions are then optimally weighted by PDW as

$$WMSE = \frac{1}{uv}\sum_{i=1}^{u}\sum_{j=1}^{v} w_{r,ij} \times (y_{ij} - t_{ij})^2,$$
$$WMAE = \frac{1}{uv}\sum_{i=1}^{u}\sum_{j=1}^{v} w_{r,ij} \times |y_{ij} - t_{ij}| \tag{10}$$

where u is the batch size, and v the number of output channels. Hence, each data point is assigned a weighting factor as well as newly logged data in the CM window for online learning. This approach allows the training process to prioritize less populated data points, enhancing the model's performance.

IV. Condition Monitoring Results and Analysis

The SCADA data of five WTs in the same wind farm are collected for the study. Four of those turbines, Ref A to Ref E, have been operating normally since commission, while WT Ref E reported a converter fault, which is treated as the detection target in this section.

For instance, from Ref A WT, the raw SCADA data of the first 150 days has 22000 sampling points, but only 16855 are valid and retained in the dataset after pre-processing.

The two 1st PCs obtained from PCA can explain 99.84% and 99.99% of the measurement information in clusters of cooling and electrical conditions, respectively. Their histograms, CDFs, and thresholds of 0.9 are shown in Fig. 5 (a) and (b), respectively. In the joint distribution shown in Fig. 5(c), most operating points fall into region 1, while the defined threshold can clearly differentiate the uneven distribution. Based on (9) and (10), the weights for region 1 and 2 are 1.13 and 8.61, respectively.

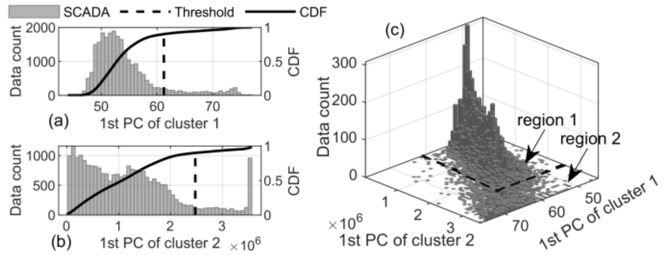


Fig. 5. The histogram and CDF of the 1st PC for (a) cluster 1 and (b) cluster 2, and (c) the joint probability distribution

The adequacy of the training data is evaluated from initial training data of 60-days. The consistency C of FCNN and LSTM with four cost functions and the different amounts of training data are calculated by (3) and (4) and plotted in Fig. 6. Without PDW, the FCNN with MAE and MSE fail to generalize on new operating points despite being trained with 150-days of training data. After accumulating over 150-days of data, the FCNNs with PDW have better performance with the C above 0.9, while the WMSE has stronger constraints on minority data and is easier to solve compared with WMAE. However, the LSTM cannot achieve such consistency when new operating points appear in a later period because the LSTM has difficulty in extracting patterns from the discontinuous sequence caused by the data removal process. Going by these observations, this study selects the FCNN with WMSE and starts the CM process 150 days after the initial commissioning.

It may be concerned that the operating condition would be influenced not only by the cooling and electrical conditions but also other unseen facts that may cause false positive fault detection. Hence, from the data-driven perspective, this study uses DBSCAN [13] to cluster every logged data set into different classes representing different operating states. During 300-days of operating, in total 21 classes are identified, while the -1 indicates that these data points are not able to cluster into any others, as shown in Fig. 7. The consistent prediction accuracy A on each class can be examined by the bar figure using dot and bar to donate the mean value and standard deviation, respectively. It can be observed that starting from the 150th day, new classes appear, i.e. 7th to 21st, but the model can still provide consistent accuracy distribution compared to previous classes. This demonstrates the robustness and generalization capabilities of the method.

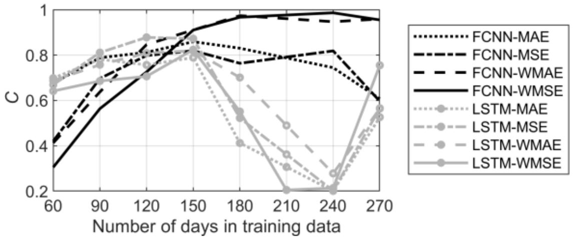


Fig. 6. The consistency C of FCNN and LSTM with four cost functions on the different amounts of training data

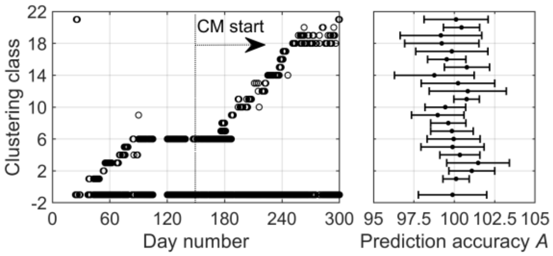
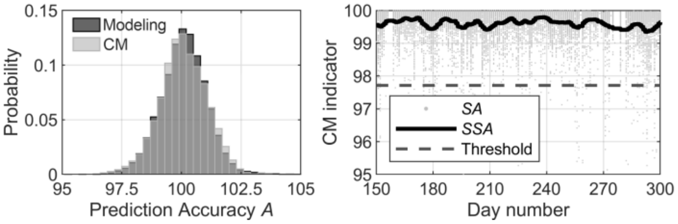


Fig. 7. The clustering results and corresponding error bar of each class

With the CM starting from the 150th day, the prediction accuracy A of the healthy WT Ref A on the training and testing dataset has the same normal distribution, which is accepted by t-test at a 5% significance level (p=0.4490), as shown in Fig. 8(a). The 30-day smoothed squared accuracy SSA keeps above the predefined fault detection threshold during the entire CM process, as shown in Fig. 8(b). The other three healthy WTs also have the same results. It indicates that the proposed method would barely result in a false positive.

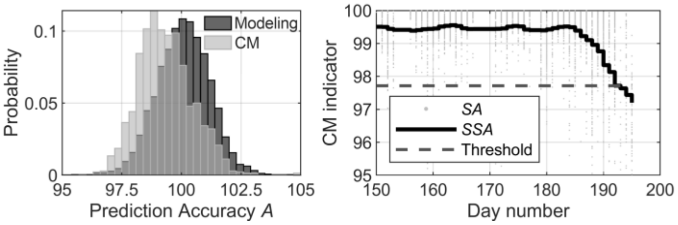
The accuracy A of Ref E WT on training and testing data cannot be considered as the same distribution by t-test, with p≅0 at 5% significance level, which indicates that the converter is no longer within the original healthy state, as shown in Fig. 9(a). On day 195, the WT is reported a converter fault and shut down, while the indicator SSA dropped below the threshold a few days before, as shown in Fig. 9(b). The proposed CM method could produce a warning message several days ahead of the converter fault, which could help coordinate predictive maintenance.



(a) (b)

Fig.8. (a) The prediction accuracy distribution and (b) condition monitoring results of healthy WT Ref A

The online learning method requires model training to be completed within 10 min. The CM algorithm of FCNN-WMSE, including the online learning process, can be completed by a PC with i7-CPU/GTX1080-GPU within one minute. Furthermore, cloud and distributed computing can be employed for large scale wind farms and effectively boost the computing for long-term CM.



(a) (b)

Fig. 9 (a) The prediction accuracy distribution and (b) condition monitoring results of Ref E WT with converter fault reported on day 195

V. Conclusion

This study proposes a data-driven condition monitoring method for WT converters using limited and unbalanced SCADA data. A DNN model is established to represent the characteristics of the healthy converter. The change in the model's prediction accuracy can be correlated with the abnormality during the converter operation, indicating a potentially active fault. A unique approach is proposed to characterize the distribution of operating data, addressing the challenges posed by limited and unbalanced SCADA data. The cost function is, therefore, modified by the probability density weight of the data to improve the model generalization. Moreover, the online learning process ensures the proposed CM method can be deployable in real-time and long-term operation. Demonstrated on the historical SCADA data from wind turbines, this CM method shows robust diagnosis results and predicts the converter fault a few days ahead of actual failure.

Reference

[1] A. Hanif, Y. Yu, D. DeVoto, and F. Khan, "A Comprehensive Review Toward the State-of-the-Art in Failure and Lifetime Predictions of Power Electronic Devices," *IEEE Transactions on Power Electronics,* vol. 34, no. 5, pp. 4729-4746, 2019.

[2] S. Yang, D. Xiang, A. Bryant, P. Mawby, L. Ran, and P. Tavner, "Condition Monitoring for Device Reliability in Power Electronic Converters: A Review," *IEEE Transactions on Power Electronics,* vol. 25, no. 11, pp. 2734-2752, 2010.

[3] Y. Zhang, Z. Wang, H. Wang, and F. Blaabjerg, "Artificial Intelligence-Aided Thermal Model Considering Cross-Coupling Effects," *IEEE Transactions on Power Electronics,* pp. 1-1, 2020.

[4] R. Liu, G. Meng, B. Yang, C. Sun, and X. Chen, "Dislocated Time Series Convolutional Neural Architecture: An Intelligent Fault Diagnosis Approach for Electric Machine," *IEEE Transactions on Industrial Informatics,* vol. 13, no. 3, pp. 1310-1320, 2017.

[5] B. Hu, Z. Hu, L. Ran, C. Ng, C. Jia, P. Mckeever, P. Tavner, C. Zhang, H. Jiang, and P. Mawby, "Heat-Flux Based Condition Monitoring of Multi-chip Power Modules Using a Two-Stage Neural Network," *IEEE Transactions on Power Electronics,* pp. 1-1, 2020.

[6] X. Rui, and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks,* vol. 16, no. 3, pp. 645-678, 2005.

[7] L. Wang, Z. Zhang, J. Xu, and R. Liu, "Wind Turbine Blade Breakage Monitoring With Deep Autoencoders," *IEEE Transactions on Smart Grid,* vol. 9, no. 4, pp. 2824-2833, 2018.

[8] F. Li, Q. Li, J. Zhang, J. Kou, J. Ye, W. Song, and A. H. Mantooth, "Detection and Diagnosis of Data Integrity Attacks in Solar Farms Based on Multi-layer Long Short-Term Memory Network," *IEEE Transactions on Power Electronics,* pp. 1-1, 2020.

[9] Y. Wan, and D. Shi, "Joint Exact Histogram Specification and Image Enhancement Through the Wavelet Transform," *IEEE Transactions on Image Processing,* vol. 16, no. 9, pp. 2245-2250, 2007.

[10] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning* (no. 2). MIT press Cambridge, 2016.

[11] J. K. Kruschke, "Bayesian estimation supersedes the t test," *Journal of Experimental Psychology: General,* vol. 142, no. 2, p. 573, 2013.

[12] H. Abdi, and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics,* vol. 2, no. 4, pp. 433-459, 2010.

[13] J. Shen, X. Hao, Z. Liang, Y. Liu, W. Wang, and L. Shao, "Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm," *IEEE Transactions on Image Processing,* vol. 25, no. 12, pp. 5933-5942, 2016.