Received May 21, 2019, accepted July 1, 2019, date of publication July 9, 2019, date of current version July 29, 2019.
Digital Object Identifier 10.1109/ACCESS.2019.2927488

A Data-Driven Health Prognostics Approach for Steam Turbines Based on Xgboost and DTW

ZIJUN QUE AND ZHENGGUO XU, (Member, IEEE)
College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
Corresponding author: Zhengguo Xu (xzg@zju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61751307 and Grant 61473254.

ABSTRACT A steam turbine is one of the critical components in a power generation system whose failure may result in unexpected consequences, even catastrophic losses. Thus, the reliability of steam turbines needs to be guaranteed all the time, which requires that its health state can be monitored and predicted effectively. Due to various failure modes, it is difficult to build physics-of-failure models used for health prognostics for steam turbines. In this paper, a data-driven integrated framework for health prognostics for steam turbines, which is based on extreme gradient boosting (XGBoost) and dynamic time warping (DTW), is proposed. The proposed framework includes two modules: anomaly detection and remaining useful life (RUL) prediction. The anomalies refer to the overall abnormal operation of steam turbines. In the process of anomaly detection, the temporal variables which can represent the operating conditions of the considered steam turbine are selected first. Appropriately selected temporal variables can reduce the input dimension and will improve real-time performance. Then, XGBoost is used to detect anomalies based on learning historical data. In the process of RUL prediction, a similarity-based algorithm with DTW is used to gain the RUL by contrasting the measured temporal variables with those in the historical cases. The similarity-based algorithm can predict the RUL without establishing a degradation path model, which can avoid the difficulties in parameter estimation for the degradation model and model generalization. The proposed framework is validated by real case studies from an industrial steam turbine. The results show that the proposed approach can detect the anomalies successfully and predict the RUL effectively.

INDEX TERMS Dynamic time warping, extreme gradient boosting, remaining useful life prediction, steam turbine.

I. INTRODUCTION

Steam turbines are the major equipment of electricity production in thermal power plants. A steam turbine is a type of heat engine which transforms heat into mechanical energy and is used in industry to generate electricity by driving an electric generator [1]. To keep a steam turbine system operating well, periodic maintenance is usually provided. However, the maintenance cost is large, which has caused many technical interests in preventive maintenance [2]. Monitoring and health prognostics for steam turbines are effective technologies to reduce avoidable and costly turbine maintenance [3]. The health prognostics includes not only anomaly detection and diagnosis, but also remaining useful life (RUL) prediction before tripping [4]. It can provide a controller with a suitable approach to extend the turbine's availability and prevent unanticipated trip losses during operation [5].

In general, the health prognostics approaches can be classified into two categories: model-based approaches [6] and data-based approaches [7]. Model-based approaches for health prognostics are based on the identification of potential failure mechanisms and failure locations for a system [8]. However, a steam turbine is a system with highly complicated structures and diverse operating conditions. For the degradation processes under varying operating conditions, obtaining well-modeled physics-of-failure (PoF) models is by no means an easy task, and hence other health prognostics approaches are needed to be utilized [9]. In this case, data-driven based health prognostics framework, which is capable of detecting anomalies and predicting undesired failures, is needed [10].

The data-driven approaches are to make use of large numbers of historical data to enable knowledge discovery and precise decision making [11]. The data-driven approaches are widely used in medicine [12], industry [13] and other fields. The data-driven health prognostics approaches are developing rapidly. Lotfi proposed a vibration-based prognostic and

The associate editor coordinating the review of this manuscript and approving it for publication was Chong Leong Gan.

health monitoring methodology for wind turbine high-speed shaft bearing using a spectral kurtosis indices and support vector machine (SVM) [14]. Dong used Brownian motion and particle filter to estimate battery state of health and RUL [15]. Jason proposed a deep learning approach for RUL prediction of rotating components with big data [16]. Neural networks have also been widely used in health prognosis [7].

Many data-driven approaches, such as support vector machine and Bayesian inference, are used for steam turbine health prognostics. The health prognostics framework for steam turbine contains anomaly detection and RUL prediction.

Anomaly detection aims to estimate the health state and the performance of a steam turbine based on process monitoring data [17]. Aiming at fault detection and diagnosis for steam turbines, Salahshoor fused an SVM classifier with an adaptive neuro-fuzzy inference system (ANFIS) to enhance the fault detection and diagnosis performance [18]. Dhini constructed a neural network (NN) based classifier to detect and diagnose faults for steam turbines [1]. Both SVM and NN have good performance on anomaly detection. However, these algorithms are sensitive to the input data. The presence of irrelevant or redundant features might reduce the speed and accuracy of the learning algorithms [19]. To avoid this, additional algorithms, such as principal component analysis (PCA), are often required to select features. Nevertheless, additional algorithms might increase the complexity of the whole approach. Extreme gradient boosting (XGBoost) is a scalable machine learning system for tree boosting [20]. The XGBoost is an improvement on gradient boosting, where the gradient boosting algorithm has the ability to retrieve importance scores for each feature after the boosted trees are constructed [21]. The importance scores can be used to select features without additional algorithms.

RUL of a steam turbine is defined as the period from the current time to the time when the steam turbine will trip. It is important and informative on the operating performance and the planning of maintenance activities [22], [23]. Choi presented a damage growth model and a RUL prediction method for aged steam turbines by using Bayesian inference [4], but the method requires to construct a degradation model. In practice, the accuracy of RUL prediction relies heavily on the selected function of the model and the initialization of the model parameters. Furthermore, different training data may lead to different models, which will bring great uncertainties to RUL prediction. To avoid this difficulty, Khelif proposed a framework that predicted RUL directly from sensor measures [19]. In this framework, the time series was first decomposed by windows. Then, the average value and the trend coefficient of the linear regression model were extracted from each window. In the proposed framework, the windows in both training and testing sets had the same length. For RUL prediction, a similarity-based algorithm, which doesn't need any prior knowledge about the components' characteristics or the degradation models [24], is proposed. The dynamic time warping (DTW) algorithm, which can calculate the similarity between two time series with different lengths, is used to locate the period in the historical case in which the data characteristics were most similar to those in the current period. The RUL can be predicted according to the located period in the historical case.

In this paper, a data-driven integrated framework for health prognostics of steam turbines based on XGBoost and DTW is proposed. The advantages of the proposed framework are as follows. First, it can select temporal variables. Other algorithms, such as deep learning, often require some specific algorithms, such as autoencoder [25], to improve the effect of feature extraction. The proposed framework can select temporal variables without requiring additional algorithms. Second, the RULs of steam turbines can be predicted through the similarities between the data characteristics in the historical cases and those in the current time period. Different from [19], the constructed windows in both the training and testing sets can be with different lengths. Therefore, the time series can be decomposed flexibly.

The rest of this paper is organized as follows. Section II gives a concise description about steam turbines. Section III introduces the proposed health prognostics framework. The used XGBoost and DTW algorithms are also introduced in this section. Section IV reports the case study results from the proposed framework on an industrial steam turbine. Finally, conclusions are presented in Section V.

II. OVERVIEW OF STEAM TURBINE

A. STEAM TURBINE

Steam turbines are widely used in power generating machines because the water is easily available, the boiling point of water is moderate, and the operating cost is acceptable. About 90% of electricity generators in the US use the steam turbines to provide driving power. A turbine is a device that converts the fluid energy into the mechanical energy. A steam turbine is a kind of system that is used to generate mechanical power from the steam pressure energy [26].

The steam turbine is a complicated system that includes high-pressure, intermediate-pressure and low-pressure sections. From the function point of view, the steam turbine system contains steam extractors, feed-water heaters and the related actuators. In addition, some special types of steam turbines may have additional structures. For example, the condensing steam turbines have condensers and circulating pumps.

B. STEAM TURBINE PROCESS MONITORING

With the continuous operation of a steam turbine, some faults will inevitably occur to its internal components. Due to the complexity of the steam turbine system, the types and the mechanisms of the faults are diverse [1].

Process monitoring data can be used to assess the operating condition and detect whether there is a fault or not. For example, when there is a fault, such as misalignment or cracking in the rotor, the vibrational amplitude will increase [27].

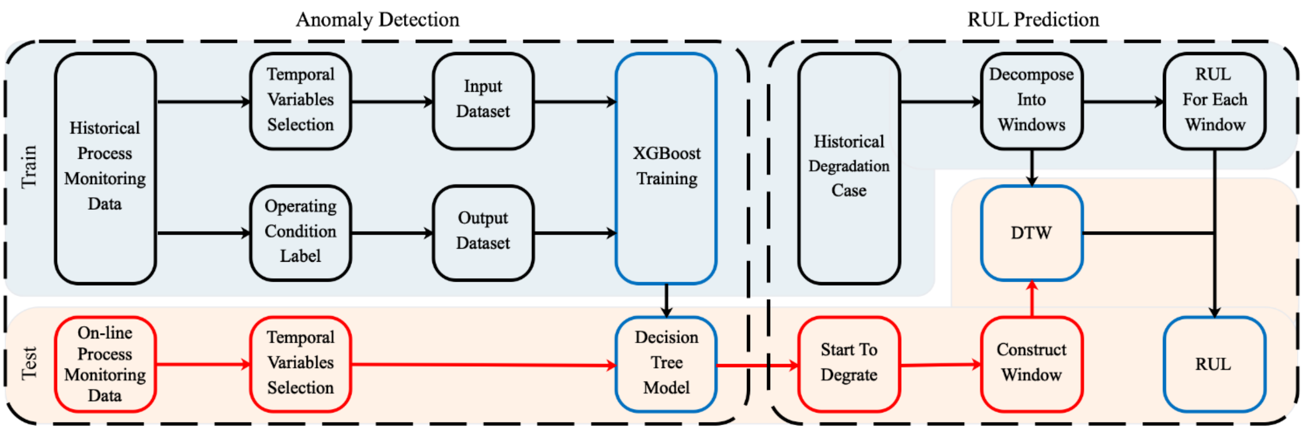


FIGURE 1. Schematic of the proposed data-driven prognostics approach.

As a result, the rotor's condition could be evaluated by monitoring the vibrational signals. In addition to the vibrational signals, the changes in thermodynamic parameters, such as temperature and pressure, directly response to the faults of the steam turbine [28]. In many cases, the fault in one component may cause changes in multiple condition parameters at the same time. An example is that the bearing misalignment may increase both vibrational levels and temperatures. Therefore, to detect the steam turbine's operating anomalies in a more comprehensive sense, it is necessary to obtain multiple process measures.

The use of process monitoring data for health prognostics of steam turbines has received extensive attentions. Tsoutsanis used several process parameters, such as temperature, to detect deviations and predict the engine behavior [29]. Rodihuez estimated the useful life by probabilistic methods based on the variables like damping and natural frequencies [30]. Choi employed the rebound hardness data to predict the RUL for aged steam turbines by using Bayesian inference [4].

The steam turbine may trip when it degrades to a certain extent. Tripping is a protection mechanism for steam turbines, and it can also be considered as the end of a useful life. When the steam turbine is detected to be in an abnormal operating condition, effective maintenance can be arranged by precise RUL prediction before the tripping of the steam turbine, which can avoid the loss caused by the unexpected failure.

III. PROPOSED APPROACH

A. FRAMEWORK

This paper proposes a data-driven integrated framework for health prognostics of steam turbines based on XGBoost and DTW. Because the process monitoring data are used to detect the operating anomalies of a steam turbine, the temporal variables from the process monitoring data are used as the input of the framework. As health prognostics focuses on current health state assessment and RUL prediction [31], the proposed framework contains two modules: anomaly detection and RUL prediction. The schematic of the proposed data-driven health prognostics framework is shown in Fig. 1. The major steps of the proposed framework are listed as follows.

1. The temporal variables with good performance on anomaly detection are selected using an embedded method.
2. A decision tree model based on XGBoost is trained using the selected temporal variables and the labeled operating conditions.
3. The historical degradation data are decomposed into several windows and the RUL is calculated for each window.
4. Anomalies are detected based on the on-line process monitoring data using the decision tree model.
5. The RUL is predicted according to the period in the historical case in which the data characteristics are most similar to those in the current period.

Among these steps, the steps 1, 2 and 3 are realized offline and the steps 4 and 5 are constructed online.

B. ANOMALY DETECTION

The purpose of anomaly detection is to monitor the operating condition and detect whether the overall operation is normal or not based on the temporal variables. In this process, two main tasks are carried out: one is the temporal variables selection, and the other is the anomaly detection.

When the operating condition of the steam turbine is being monitored, several temporal variables can be obtained. However, not all temporal variables are needed by anomaly detection. The anomaly detection algorithm based on too many temporal variables will be with high complexity and large amount of computation. It is necessary to distinguish the temporal variables highly related to anomalies from needless ones. In this framework, an embedded method for temporal variables selection is proposed. At first, all the temporal variables are selected as input candidates to train the decision tree based on XGBoost. Then, after training the decision tree, the importance indexes of each temporal variable will be obtained. The higher the importance index score is, the greater the effect of the corresponding temporal variable

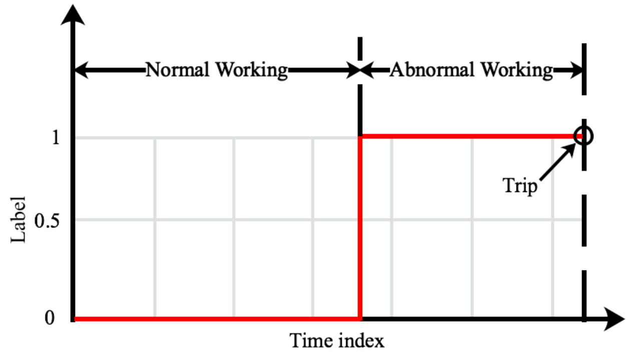


FIGURE 2. Scheme of operating condition label.

on the anomaly detection has. At last, the temporal variables with high scores are selected as the final inputs.

As the XGBoost is a supervised machine learning algorithm, the operating conditions should be labeled. When the steam turbine works properly, the operating condition is labeled as 0. When a fault occurs, the operating condition is labeled as 1. In this way, the useful life of a steam turbine is divided into normal working period and abnormal working period. In the abnormal working period, the steam turbine begins to degrade. A curve reflecting the operating conditions of the steam turbine is constructed, as shown in Fig. 2. Using such a labeling scheme, the profile of the curve in the whole life cycle is similar to that of a step function with the level jumping from 0 to 1. The labeled operating conditions are used as the outputs while training the decision tree.

The XGBoost algorithm is then used to train the decision tree that models the relationship between the temporal variables and the operating conditions of the steam turbine.

Given a data set $D = \{(\mathbf{x}_i, y_i)\}$, a tree ensemble model to predict the output is defined by

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^m$ is the input vector with m temporal variables, $y_i \in \{0, 1\}$ is the output, $\hat{y}_i$ is the predicted output, $i = 1, 2, \cdots, n$, K is the number of trees, f_k corresponds to an independent tree with the structure q_k and the leaf weight w_k.

The regularized objective function is defined as

$$L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{2}$$

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \|w_k\|^2, \tag{3}$$

where $l(\cdot)$ is a differentiable convex loss function that measures the difference between the predicted output $\hat{y}_i$ and the target output y_i. The commonly used functions include the mean square loss function and the Logistic loss function [32]. The function $\Omega(\cdot)$ penalizes the complexity of the model. T_k is the number of leaves in the tree. $\gamma \in [0, 1]$ is the learning rate. γT_k stands for the spanning tree pruning, which is used to prevent the overfitting. λ is a parameter in the regularization term that controls the weight of the model's complexity. The target of the XGBoost algorithm is to minimize equation (2).

It can be found from equation (2) that the tree ensemble model includes the functions of functions. As a result, equation (2) cannot be optimized by using the traditional optimization methods in the Euclidean space. Under this circumstance, let $\hat{y}_i^{(p)}$ be the estimate of the i-th sample at the p-th iteration, then equation (2) can be rewritten as

$$L^{(p)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(p-1)} + S_p(\mathbf{x}_i)\right) + \Omega(S_p), \tag{4}$$

where S_p represents the tree generated by the i-th sample in the p-th iteration that is added to minimize the objective function.

Equation (4) is optimal when the loss function is the square types, but it becomes complicated when other types of loss functions are used. Therefore, a second-order approximation is used to optimize the objective function for other kinds of loss functions

$$L^{(p)} \approx \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}^{(p-1)}\right) + g_i S_p(\mathbf{x}_i) + \frac{1}{2} h_i S_p^2(\mathbf{x}_i)\right] + \Omega(S_p), \tag{5}$$

where $g_i = \partial_{\hat{y}^{(p-1)}} l\left(y_i, \hat{y}^{(p-1)}\right)$ and $h_i = \partial^2_{\hat{y}^{(p-1)}} l\left(y_i, \hat{y}^{(p-1)}\right)$ are the first and second order gradient statistics on the loss function $l(\cdot)$.

In equation (5), $l\left(y_i, \hat{y}^{(p-1)}\right)$ is a constant term, so it can be removed to simplify equation (5). Define $I_j = \{i \mid q(x_i) = j\}$ as the sample set of the leaf j, $j = 1, 2, \cdots, T_k$. Equation (5) can be rewritten by expanding $\Omega(\cdot)$ as

$$\tilde{L}^{(p)} = \sum_{j=1}^{T_k}\left[\left(\sum_{i \in I_j} g_i\right) w\_l_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w\_l_j^2\right] + \gamma T, \tag{6}$$

where $\tilde{L}^{(p)}$ is the simplification of $L^{(p)}$ by removing the constant terms. For a fixed structure $q(\mathbf{x})$, the optimal weight $w\_l_j^*$ of the leaf j can be computed as

$$w\_l_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \tag{7}$$

And the corresponding optimal value can be calculated by

$$\tilde{L}^{(p)}(q) = -\frac{1}{2}\sum_{j=1}^{T_k}\frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T, \tag{8}$$

Normally, it is impossible to enumerate all the possible tree structures as the number of tree structures is infinite. A greedy algorithm that starts from a single leaf and iteratively adds branches to the tree is used instead to find the optimal splitting point. Let I_L and I_R be the samples of the left and right nodes after the splitting and $I = I_L \cup I_R$, then the loss reduction after the splitting is given by

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma \tag{9}$$

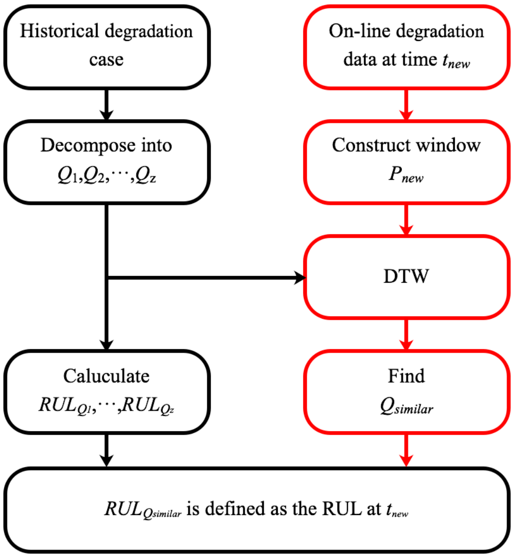


FIGURE 3. **Scheme of RUL prediction.**

More details on XGBoost can be found in [20], [32]. After training, the tree structure and leaf weights in the decision tree model can be obtained. The selected temporal variables obtained from the on-line process monitoring data are input into the decision tree model. The input vector is discriminated from the root node to the leaf node and the prediction score can be obtained by leaf weight calculation. The finial prediction score is mapped to the output. The output of the decision tree model can be used to detect the anomalies. If the output is 0, it is concerned that the steam turbine is judged to be working properly. If the output is 1, it is concerned that overall operation of steam turbine is abnormal.

C. RUL PREDICTION

After anomaly detection, the time when the steam turbine starts to degrade can be determined. There is a degradation period from fault occurrence to system failure. To predict the potential failure, a similarity-based RUL prediction algorithm is proposed during the degradation period, as shown in Fig. 3. In the process of RUL prediction, the generator power is selected as the life indicator as the generator power can reflect the operating state of the steam turbine system in an intuitive way and it has good degradation tendency.

At first, one or several historical degradation cases of some steam turbines are learned. As shown in Fig. 4. the length of the time series in the degradation period is defined as Len_{time}. Then this time series is decomposed into several windows $Q = \{Q_1, Q_2, \cdots, Q_z\}$, where z is the number of windows. There are overlaps between close windows. The length of each window is given Len_Q. The RUL of each window is defined as

$$RUL_{Q_v} = \ell \cdot (Len_{time} - v \cdot Len_Q) \tag{10}$$

where ℓ is the sampling interval of the process monitoring data, $v = 1, 2, \cdots, z$.

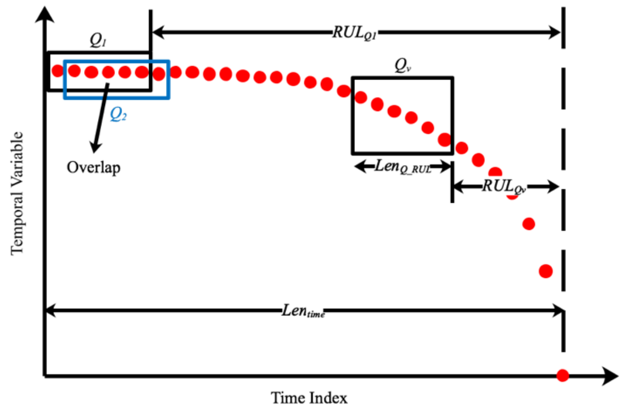


FIGURE 4. **Scheme of the historical cases learning.**

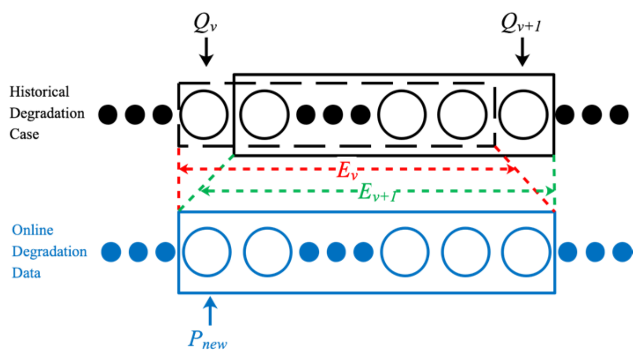


FIGURE 5. **Scheme of the moving window strategy.**

The relationship between the decomposed windows and RUL is used to predict the RUL based on on-line process monitoring data. At the time t_{new}, new process data are sampled. Then a new window P_{new} is constructed, as shown in Fig. 5. The length of P_{new} is Len_P, which can be different from Len_Q. DTW is used to calculate the similarity between P_{new} and Q_v. $Q_{similar}$, in which the data are the most similar to those in P_{new}, is found by using a moving window strategy, as shown in Fig. 5.

DTW is one of the best distance measure-based approaches [33]. It is widely applied to execute similarity searching and detection in time series [34]. DTW is a time series analysis approach which can measure the divergence between two time series with different phases and lengths [35]. By calculating an optimal warping path, the time series can be warped nonlinearly in the time dimension. After that, two extended time series will be placed in one-to-one correspondence, and their similarity can be evaluated.

Time series P_{new}, with the length Len_P, is selected as the testing set. The other series Q_v, with the length Len_Q, is selected as the training set.

$$P_{new} = P_{new_1}, P_{new_2}, \cdots, P_{nwe_{Len_P}}, \tag{11}$$

$$Q_v = Q_{v_1}, Q_{v_2}, \cdots, Q_{v_{Len_Q}}, \tag{12}$$

To measure the similarity between these two time series, an $Len_P - by - Len_Q$ matrix MT is created. The value of the $MT_{a,b}$ stands for the distance between the points P_{new_a}

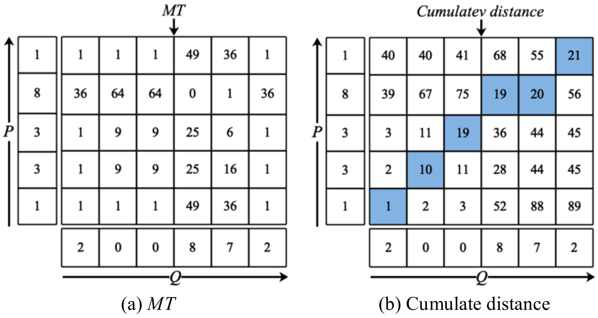


FIGURE 6. An example of DTW.

Temporal Variables						Label
F0	F1	F2	F3	F4	F5	
Generator Power	Condenser Vacuum	Inlet Temperature	Outlet Temperature	Inlet Pressure	Outlet Pressure	
......	......	......	......	......	......	......
0.9831	0.0606	0.7075	0.1892	0.3700	0.5133	0
0.9873	0.0613	0.7094	0.1908	0.3704	0.5140	0
0.9842	0.0619	0.7113	0.1922	0.3704	0.5148	1
0.9841	0.0626	0.7132	0.1937	0.3710	0.5155	1
......	......	......	......	......	......	......

FIGURE 7. Data marking scheme.

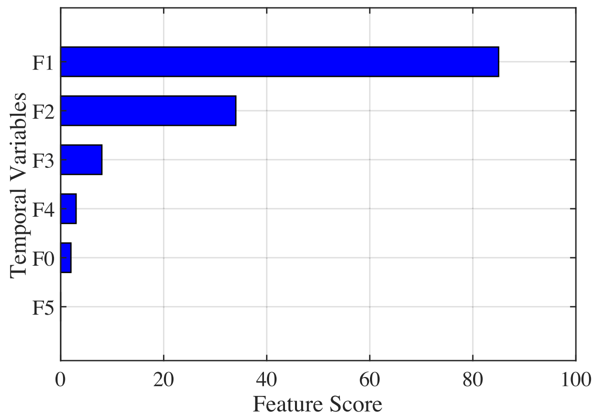


FIGURE 8. Temporal variables importance indexes.

and Q_{v_b}

$$MT_{a,b} = \left(P_{new_a} - Q_{v_b}\right)^2, \tag{13}$$

where $a = 1, 2, \cdots, Len_p$, $b = 1, 2, \cdots, Len_Q$.

To show the algorithm more clearly, Fig. 6 provides an example. In Fig. 6, the testing set is set as [1, 3, 3, 8, 1], and the training set is set as [2, 0, 0, 8, 7, 2]. The MT is a 5-by-6 matrix.

With the distance matrix MT, the warping path can be derived. The shortest warping path through the matrix is derived by,

$$E_{a,b} = r \cdot \left(P_{new_a} - Q_{v_b}\right)^2 + \min\{E_{a,b-1}, E_{a-1,b}, E_{a-1,b-1}\}, \tag{14}$$

$$E_{1,1} = \left(P_{new_1} - Q_{v_1}\right)^2, \tag{15}$$

where $E_{a,b}$ is the cumulated distance in the warping path, r is the weight of DTW, which is set to 1 for usual DTW computation. This procedure is called warping path calculation. The blue panes in Fig. 6 stand for the warping path.

Usually, the combination of the distance matrix calculation and the warping path calculation is regarded as the complete DTW matrix calculation. E_{Len_p,Len_Q} can be used to assess the similarity between P_{new} and Q_v. The smaller the value of E_{Len_p,Len_Q}, the more similar the two time series.

After $Q_{similar}$ is found, equation (10) is used to calculate the $RUL_{Q_{similar}}$ that is defined as the RUL of the steam turbine at time t_{new}.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed framework is to be validated by using the real case data from an industrial steam turbine. The process monitoring data record the whole process from the normal working phase to the trip caused by some fault. Several temporal variables are monitored in real time, including generator power (F0), condenser vacuum (F1), condenser inlet temperature of circulating water (F2), condenser outlet temperature of circulating water (F3), condenser inlet pressure of circulating water (F4), and condenser outlet pressure of circulating water (F5). The sampling interval of the process monitoring data is 1s. Due to the variations in the working environment and the operating conditions of different steam turbines, the temporal variables are normalized to eliminate their influence on the orders of magnitude. After normalization, the temporal variables are labeled according to the labeling method mentioned in Section III. The scheme of the process monitoring data is shown in Fig. 7.

When the proposed framework is validated, the historical data from one steam turbine are selected as the training group, and the data from another steam turbine are used for testing. The two steam turbines have the similar operation conditions.

The importance indexes of each temporal variable are shown in Fig. 8. It can be found that only a few temporal variables have high scores, such as the condenser vacuum and the inlet temperature, which means that these temporal variables have great influence on anomaly detection. Some other temporal variables, such as the inlet pressure and the outlet pressure, have very low score. Even some temporal variables, such as the outlet pressure, have no scores. In other words, these temporal variables have little or even no influence on anomaly detection. This fact indicates that the considered anomaly can be detected by training a few, rather than all, temporal variables, and including too many temporal variables may not improve the anomaly detection performance. Therefore, the temporal variables with high scores in Fig. 8 are selected to form a new temporal variable set used for training. The three selected temporal variables are the condenser vacuum, the inlet temperature and the outlet temperature.

Fig. 9 shows the anomaly detection results. From the real label curve, it can be found that the steam turbine system

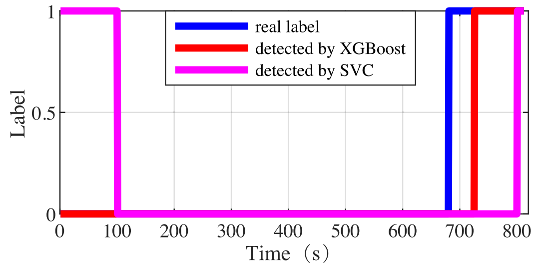


FIGURE 9. Results of anomaly detection.

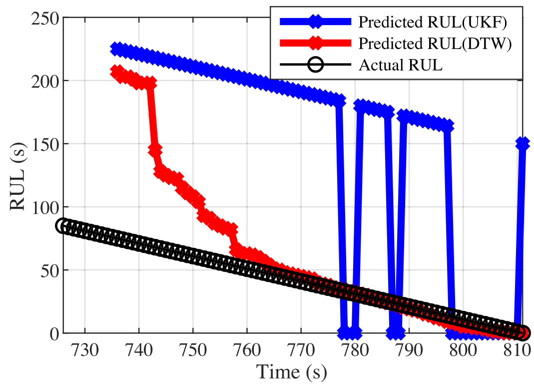


FIGURE 10. Results of RUL prediction.

started to degrade at the 680th second. Before that time, the steam turbine was under the normal working condition. The detected curve by XGBoost has a similar profile to that of the real curve. The time when the steam turbine system started to degrade detected by XGBoost is the 726th second, which indicates that XGBoost can detect the anomaly after the occurrence of the actual anomaly with the delay of 46 seconds. This comes from the fact that the temporal variables of the system did not change significantly in the initial period of degradation. The behaviors of the temporal variables during the initial degradation period are similar to those during the normal working period, which will lead to the misjudgment in anomaly detection. Overall, the detected curve provided by XGBoost is close to the real curve, indicating good performance in anomaly detection. To show the effectiveness of XGBoost, the support vector classifier (SVC), which is widely used in fault detection, is chosen for comparison. From Fig. 9, it can be found that the detected curve provided by SVC does not coincide with the real curve in the initial stage of normal working. The healthy states that were in normal working period in reality are judged as abnormal states. The time when the steam turbine started to degrade detected by SVC is the 800th second, which is 74 seconds later than the time detected by XGBoost. The comparatively big errors from SVC comes from the fact that SVC needs to map the data to an appropriate space through some kernel function when dealing with the nonlinear problems. However, it is difficult to select an appropriate kernel function for a specific application, which will affect the performance of SVC on anomaly detection.

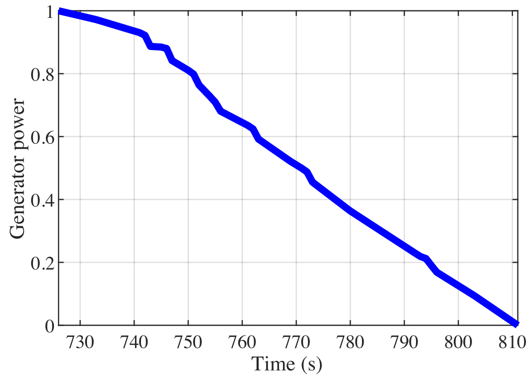


FIGURE 11. Schematic of degradation trend.

Fig. 10. shows the RUL prediction results. It can be found from Fig. 10 that there is a significant difference between the predicted life and the actual life at the beginning of the RUL prediction. This is caused by the fact that the degradation path changed greatly after a certain instant. As shown in Fig. 11, in the initial degradation period, the decline rate of the degradation path is small and the degradation trend shows a tendency of concussion descent. The $Q_{similar}$ from the historical case cannot truly represent the degradation tendency of P_{new} in the early degradation period. Therefore, there is a large error between the predicted and the true RULs. From Fig. 10, it can be found that the predicted curve is close to the real one in the middle and late periods of RUL prediction, indicating that the RUL prediction is accurate. With the increase of the degradation information, the degradation trend of the steam turbine is gradually obvious in the later degradation process. The decline rate gradually accelerated, and the generator power quickly dropped to 0 at the end of the degradation process. At this time, the $Q_{similar}$ from the historical case can well describe the degradation tendency of P_{new}. The UKF-based approach is used to compare the RUL prediction performance. From Fig. 10, it can be concluded that the proposed approach has better performance than the UKF for RUL prediction.

V. CONCLUSIONS

This paper has proposed a data-driven integrated framework for health prognostics of steam turbines. The proposed framework includes two modules: anomaly detection and RUL prediction. The temporal variables that have great influence on anomaly detection are selected for anomaly detection in an embedded method. The anomaly detection results using XGBoost are close to the realities, and it has better performance than SVC. The similarity-based method can track the trend of the degradation process and predict RUL accurately when the trend has been learned. The proposed framework is verified by using the real case data from an industrial steam turbine.

In the current study, the misjudgment in anomaly detection will be occur in the initial degradation period. It needs to be improved by optimizing the algorithm. In RUL prediction module, to find the most similar window, the strategy of traversing all windows in the historical case is adopted. This strategy is time-consuming, so it needs to be optimized to improve the searching efficiency.

REFERENCES

[1] A. Dhini, B. Kusumoputro, and I. Surjandari, "Neural network based system for detecting and diagnosing faults in steam turbine of thermal power plant," in *Proc. Int. Conf. Awareness Sci. Technol.*, Taichung, Taiwan, Nov. 2017, pp. 149–154.

[2] J. F. Wan, S. L. Tang, D. Li, S. Y. Wang, C. L. Liu, H. Abbas, and A. V. Vasilakos, "A manufacturing big data solution for active preventive maintenance," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2039–2048, Aug. 2017.

[3] C. Karlsson, J. Arriagada, and M. Genrup, "Detection and interactive isolation of faults in steam turbines to support maintenance decisions," *Simul. Model. Pract. Theory*, vol. 16, no. 10, pp. 1689–1703, Nov. 2008.

[4] W. Choi, B. D. Youn, H. Oh, and N. H. Kim, "A Bayesian approach for a damage growth model using sporadically measured and heterogeneous on-site data from a steam turbine," *Rel. Eng. Syst. Saf.*, vol. 184, pp. 137–150, Mar. 2018.

[5] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, "A data-driven design for fault detection of wind turbines using random forests and XGboost," *IEEE Access*, vol. 6, pp. 21020–21031, May 2018.

[6] W.-C. Lin and Y. A. Ghoneim, "Model-based fault diagnosis and prognosis for electric power steering systems," in *Proc. IEEE Int. Conf. PHM*, Ottawa, ON, Canada, Jun. 2016, pp. 1–8.

[7] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, May 2018.

[8] H.-F. Wang, "Prognostics and health management for complex system based on fusion of model-based approach and data-driven approach," *Phys. Procedia*, vol. 24, pp. 823–831, Jan. 2012.

[9] S. Gentil, J. Montmain, and C. Combastel, "Combining FDI and AI approaches within causal-model-based diagnosis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 5, pp. 2207–2221, Oct. 2004.

[10] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018.

[11] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, Jun. 2017.

[12] A. K. Waljee, B. Liu, K. Sauder, J. Zhu, S. M. Govani, R. W. Stidham, and P. D. R. Higgins, "Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis," *Alimentary Pharmacol. Therapeutics*, vol. 47, no. 6, pp. 763–772, Jan. 2018.

[13] Z. Huang, Z. Xu, W. Wang, and Y. Sun, "Remaining useful life prediction for a nonlinear heterogeneous Wiener process model with an adaptive drift," *IEEE Trans. Rel.*, vol. 64, no. 2, pp. 687–700, Jun. 2015.

[14] 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," *Appl. Acoust.*, vol. 120, pp. 1–8, May 2017.

[15] G. Dong, Z. Chen, J. Wei, and Q. Ling, "Battery health prognosis using Brownian motion modeling and particle filtering," *IEEE Trans. Ind. Electron.*, vol. 65, no. 11, pp. 8646–8655, Nov. 2018.

[16] J. Deutsch and D. He, "Using deep learning-based approach to predict remaining useful life of rotating components," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 48, no. 1, pp. 11–20, Jan. 2018.

[17] D. Liu, J. Pang, G. Song, W. Xie, Y. Peng, and X. Peng, "Fragment anomaly detection with prediction and statistical analysis for satellite telemetry," *IEEE Access*, vol. 5, pp. 19269–19281, Sep. 2017.

[18] K. Salahshoor, M. Kordestani, and M. S. Khoshro, "Fault detection and diagnosis of an industrial steam turbine using fusion of SVM (support vector machine) and ANFIS (adaptive neuro-fuzzy inference system) classifiers," *Energy*, vol. 35, pp. 5472–5482, Dec. 2010.

[19] R. Khelif, B. Chebel-Morello, S. Malinowski, E. Laajili, F. Fnaiech, and N. Zerhouni, "Direct remaining useful life estimation based on support vector regression," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2276–2285, Mar. 2017.

[20] T. Chen and C. Gustrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[21] Z. Wang, Q. Yu, C. Shen, and W. Hu, "Feature selection in click-through rate prediction based on gradient boosting," in *Proc. Int. Conf. Intell. Data Eng. Automat. Learn.*, Yangzhou, China, 2016, pp. 134–142.

[22] X.-S. Si, W. Wang, C.-H. Hu, and D.-H. Zhou, "Remaining useful life estimation—A review on the statistical data driven approaches," *Eur. J. Oper. Res.*, vol. 213, no. 1, pp. 1–14, Aug. 2011.

[23] Z. Huang, Z. Xu, X. Ke, W. Wang, and Y. Sun, "Remaining useful life prediction for an adaptive skew-Wiener process model," *Mech. Syst. Signal Process.*, vol. 87, pp. 294–306, Mar. 2017.

[24] R. Khelif, S. Malinowski, B. Chebel-Morello, and N. Zerhouni, "RUL prediction based on a new similarity-instance based approach," in *Proc. IEEE Int. Symp. Ind. Electron.*, Istanbul, Turkey, Jun. 2014, pp. 2463–2468.

[25] F. Jia, Y. Lei, L. Guo, J. Lin, and S. Xing, "A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines," *Neurocomputing*, vol. 272, pp. 619–628, Jan. 2018.

[26] K. Venkatesh, P. V. Raju, and T. J. Kumar, "Residual life assessment of 60 MW steam turbine rotor," *Int. J. Sci. Res.*, vol. 8, no. 2, pp. 1–11, Aug. 2012.

[27] H. C. Sun, C. M. Huang, and Y. C. Huang, "Fault diagnosis of steam turbine-generator sets using an EPSO-based support vector classifier," *IEEE Trans. Energy Convers.*, vol. 28, no. 1, pp. 164–171, Mar. 2013.

[28] W. Huang, J. Yu, X. Zhao, and X. Lu, "Fault diagnosis for steam turbine based on flow graphs and Naïve Bayesian classifier," in *Proc. IEEE Int. Conf. Mech. Autom.*, Tianjin, China, Aug. 2014, pp. 396–401.

[29] E. Tsoutsanis and N. Meskin, "Derivative-driven window-based regression method for gas turbine performance prognostics," *Energy*, vol. 128, pp. 302–311, Jun. 2017.

[30] J. A. Rodrigues, J. C. Gracia, E. Aloinso, Y. El Hamzaoui, J. M. Rodríguez, and G. Urquiza, "Failure probability estimation of steam turbine blades by enhanced Monte Carlo method," *Eng. Failure Anal.*, vol. 56, pp. 80–88, Oct. 2015.

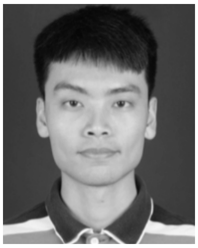
[31] J. Lee, F. J. Wu, W. Y. Zhao, M. Ghaffari, L. X. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems-reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, nos. 1–2, pp. 314–334, Jan. 2014.

[32] Z. Chen, F. Jiang, Y. J. Chen, X. Gu, W. Liu, and J. Peng, "XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud," in *Proc. IEEE Big Data Smart Comput.*, Shanghai, China, Jan. 2018, pp. 251–256.

[33] X. Xu, F. Lin, A. Wang, X. Yao, Q. Lu, W. Xu, Y. Shi, and Y. Hu, "Accelerating dynamic time warping with memristor-based customized fabrics," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 4, pp. 729–741, Apr. 2018.

[34] M. M. Zhang and D. Pi, "A new time series representation model and corresponding similarity measure for fast and accurate similarity detection," *IEEE Access*, vol. 5, pp. 24503–24519, Oct. 2017.

[35] J. Mei, M. Liu, Y.-F. Wang, and H. Gao, "Learning a Mahalanobis distance-based dynamic time warping measure for multivariate time series classification," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1363–1374, Jun. 2016.



ZIJUN QUE was born in China, in 1992. He received the B.Eng. and M.Eng. degrees in mechanical engineering from the Zhejiang University of Technology, Hangzhou, China, in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree with the College of Control Science and Engineering, Zhejiang University, Hangzhou.

His current research interests include reliability evaluation, and fault diagnosis and prognosis.

ZHENGGUO XU received the B.S. degree in electrical engineering from Changsha Electric Power University, Changsha, China, in 2002, and the M.S. and Ph.D. degrees in control science and engineering from Tsinghua University, Beijing, China, in 2009.

He is currently an Associate Professor with the College of Control Science and Engineering, Zhejiang University, Zhejiang, China. His research interests include data analysis and artificial intelligence applications in energy systems, fault diagnosis and prediction, and reliability analysis.