Diagnostics and Prognostics of Boilers in Power Plant Based on Data-Driven and Machine Learning

Achmad Widodo[1], Toni Prahasto[1], Mochamad Soleh[2], and Herry Nugraha[3]

[1]*Department of Mechanical Engineering, Universitas Diponegoro, Semarang 50275, Indonesia*
achmadwidodo@lecturer.undip.ac.id
toniprahasto@lecturer.undip.ac.id

[2]*PT. PLN (Persero) Research Institute, Jl. Duren Tiga 102, Jakarta 12760, Indonesia*
mochamad.soleh@pln.co.id

[3]*Faculty of Technology and Energy Business, Institut Teknologi PLN, Jakarta 11750, Indonesia*
herry.nugraha@pln.co.id

Abstract

This paper reports diagnostics and prognostics study of boiler in power plant using actual boiler operating data. This study aims to early detect anomalies that occur in the boiler and to predict the remaining useful life (RUL) after anomalies are detected. The proposed method utilizes machine learning techniques through support vector machine (SVM) and random forest algorithm (RFA) for anomaly detection and similarity-based method of dynamic time warping (DTW) for RUL prediction. The developed method is validated by testing the prediction models using real operating data acquired from three boilers in power plant. The results show that some anomalies are successfully detected by prediction model even though there are anomalies that give low accuracies in predictions. RUL prediction also provides fair results given the limitations of the real data used in building prediction models. Overall, the results of this study have potential to be applied in real system as an auxiliary tool in the boiler condition monitoring to support boiler maintenance programs.

1. Introduction

Boiler is one of the important equipment in thermal power plant to generate electricity. Boiler converts chemical energy from fuel into thermal energy for steam generation (Dzikuć et al., 2020). The steam is then used to rotate steam turbine that drives electric generator to generate electricity that is sent to the network through the main transformer. Boiler consists of thousands of heat exchanger tubes that transfer heat from combustion fuel to feedwater in the tubes and change phase into superheated steam or steam to drive steam turbine. As important equipment, boiler must have high level of reliability and equipment availability. To keep the boiler system operating with good performance, maintenance strategy must be chosen properly. Preventive maintenance through periodic inspection is usually provided, but this will be expensive because of many technical interests and labor involved in preventive maintenance. Another method, namely condition-based maintenance (CBM), provides a maintenance strategy that involves installing many sensors including temperature and pressure sensors in many positions of the boiler system (Mushiri et al., 2018). In addition, the position of the water level in the drum is also an important parameter for boiler monitoring. Much data can be collected through sensors that represent the actual condition of the boiler. Further data analysis is then carried out on sensor variables including data interpretation to obtain parameters that are useful for boiler operation. In practice, boiler maintenance is usually carried out with a combination of preventive maintenance and condition-based maintenance to obtain a more beneficial and reliable maintenance system.

Furthermore, CBM applied in boiler maintenance strategy should have the function of fault diagnostic and boiler health prognostic. Fault diagnostic means detection of some anomalies and isolation of fault symptoms based on parameters from sensor data. When fault symptoms appear for the first time at the early stage, it means an anomaly occurs, the CBM monitoring system will notify the operator through an alert indicating that the boiler has some problems. Prognostic means predicting how much time is left after the fault is first detected through the fault diagnostic task in the

https://doi.org/10.36001/IJPHM.2025.v16i1.4222

CBM system. Prognostic is intended to predict the remaining useful life (RUL) based on historical data and actual degradation trends observed from condition monitoring information (Do et al., 2015). When the RUL is predicted, the decision maker can set a schedule for the boiler maintenance actions as well.

Machine fault diagnostics and prognostics are interesting research topics in the field of engineering maintenance. The development of machine fault diagnostics and prognostics methods has attracted researchers' interest for many years. Montero Jimenez et al. (2020) reported based on a literature survey on the developed fault diagnostics and prognostics methods applied in predictive maintenance (Montero Jimenez et al., 2020). In this survey, he showed that there has been a significant increase in the number of publications on machine diagnostics and prognostics studies over the 25 years to 2019. Other reviews also present similar results that a large number of research works related to prognostics have been carried out. And many literatures have provided a very good overview of the process in predicting machine RUL (Lei et al., 2018; Dalzochio et al., 2020; Diez-Olivan et al., 2019).

Basically, the methods applied in prognostics can be classified into two approaches: model-based and data-based. In model-based methods, prognostics is performed using the development of models of potential failure mechanisms and identification of locations in the system. Model-based prognostics involves the derivation of mathematical equations that represent the system in both normal and faulty conditions (Hong-feng, 2012; Lin & Ghoneim, 2016). Modeling of mechanical systems is sometimes not easy because of their very complex structures and different operating conditions of the system. In addition, determining a model that can represent the degradation status under various operating conditions and dynamic physical failure models is not an easy task.

Therefore, another data-driven approach is needed to predict machine health, which is capable of detecting anomalies and predicting future machine status (Madrigal-Espinosa et al., 2017; Duong et al., 2019; Sohaib & Kim, 2019). Data-driven approach is to utilize huge data to enable knowledge discovery in data and proper decision making with the help of certain methods such as statistical analysis or artificial intelligence. Statistical models in prediction build predictive models by fitting available observations into random variable models or stochastic models with probabilistic methods without relying on any physics and principles. Random variance is generally introduced into the variable model to describe the uncertainty caused by various types of source variability, such as temporal variability, unit-to-unit variability, and measurement variability. Therefore, statistical approaches are effective in describing the uncertainty of the degradation process and its impact on RUL prediction (Wang et al., 2016). Artificial intelligence methods are used for data-driven prediction by imitating the human brain that tries to learn the machine degradation pattern from historical data observations. The degradation process of mechanical systems is sometimes too difficult to be associated with physical or statistical approaches. Therefore, many researchers have proposed methods that treat and analyze such mechanical systems as black boxes. This means that the behavior of the system will be analyzed through the responses measured by sensors. The degradation patterns are then extracted from the sensor data where artificial intelligence can learn the dynamics of the patterns. In addition, using trained artificial intelligence, target parameters will be predicted to determine the RUL. The results of artificial intelligence approaches are actually difficult to explain due to the lack of transparency (Świercz & Mroczkowska, 2019). In addition, there are also many creative ideas proposed by some researchers that utilize hybrid methods that try to integrate the advantages of different approaches through some adjustments. Several papers report the use of hybrid methods in data-driven prognostics, for example by Liao and Köttig (2014), Sbarufatti et al. (2016) and (Acuña & Orchard, 2017).

Boiler fault diagnostic studies are still an interesting research field because boilers are critical equipment in industry. Quasi-linear parameter variants (quasi-LPV) representing the dynamics of more critical variables including turbine pressure, drum pressure, and electric power are studied for the detection and isolation of boiler-turbine system faults (Madrigal-Espinosa et al., 2017). In their study, the proposed method contributes a reliable fault diagnostic system to detect sensor faults in a wide operating range of boilers. Duong studied boiler leak detection based on acoustic emission (AE) signals and deep learning methods used as an intelligence tool for leak detection (Duong et al., 2019). The proposed system idea uses the capability of AE sensors that are able to capture high amplitude impulses generated by the interaction between coal fuel flow and boiler tubes through signal shape information. However, this work was tested in a laboratory testbed scale or not in a real industrial environment. The use of AE sensors for boiler condition monitoring is still rare in real industrial applications due to the high-cost requirements for AE sensors and data acquisition units. Another study using AE sensors for boiler diagnostics was reported by Sohaib and Kim (2019). The following researchers also contributed some techniques for boiler diagnostics using their own adapted methods such as Wang et al., (2016), Swiercz and Mroczkowska (2019), Cui et al. (2020) and Panday et al. (2021).

In addition to fault diagnostic methods, there are some studies reporting prognostics for boiler RUL. Khan studied the prognostics of steam generators (or boilers) in nuclear power plants based on Eddy current inspection data and particle filter (PF) method (Khan et al., 2011). The purpose of their study was to assess the condition of boiler tubes through RUL prediction to ensure corrective actions before tube leakage that may cause accidents. The data used in their work were inspection data which means the steam generator was out of service. Nguyen et al. (2018) also worked on a data-driven method to predict boiler RUL. They used a time series analysis method namely autoregressive integrated moving average (ARIMA) on chemical plugging data on tube support plates (TSP) to predict boiler RUL. Similar to the previous study by Khan et al. (2011), the boiler condition was also out of service.

This paper presents a study on boiler fault diagnostics and prognostics based on real industrial data. The data used in this study inline with the data used for boiler performance forecasting and monitoring reported by Jia et al. (2022), Hong et al. (2022) and Xu et al. (2024). The contribution of this study is that the data used for fault diagnostics and prognostics are acquired while the boiler is still in operation. There is no interruption to the steam generation process at all while the data are acquired. The data are collected by a supervisory control and data acquisition (SCADA) system streamed online from several sensors installed on the boiler and stored into a hard disk for further analysis. This is in stark contrast to some published papers that only use data from laboratory test benches or inspection data when the boiler is shut down. A machine leaning approach through support vector machine (SVM) and random forest algorithms (RFA) are applied as tools for automatic boiler detection and RUL prediction is performed using dynamics time warping (DTW).

SVM is a very popular method in machine learning, introduced by Vapnik (2013), which is implemented in classification and regression for various fields of study in industrial applications. Alegeh et al. (2019) used SVM in the area of product service systems to monitor the degradation of a 5-axis gantry machine and used the results to offer maintenance services. Li et al. (2019) developed tool wear detection based on audio signal processing and data compression using PCA. In their work, SVM was used to detect tool conditions based on classification techniques. A recent paper presents a comprehensive review of machine learning, including SVM, in industrial applications has been reported by Bertolini et al. (2021). However, observing the publications related to boiler diagnostics, the use of SVM in this field is relatively few, for example the papers published by Chen et al. (2011), Berahman et al. (2013) and Khalid et al. (2020). In addition, RFA applied in boiler diagnostics was contributed by Shohet et al. (2019, 2020) but their work was for boilers in building systems rather than in power plants.

The rest of this paper is organized as follows: Section 2 describes an overview of boilers in power plants, Section 3 presents the proposed framework for boiler fault diagnostics and prognostics including the presentation of the algorithms used in this study. Section 4 illustrates the experimental work including the results and discussion of the approach. Finally, the paper concludes with the research conclusions and some perspectives in Section 5.

2. OVERVIEW BOILER IN POWER PLANT

The target of the proposed method is a boiler that supplies steam to drive a 600 MW steam turbine-generator system.

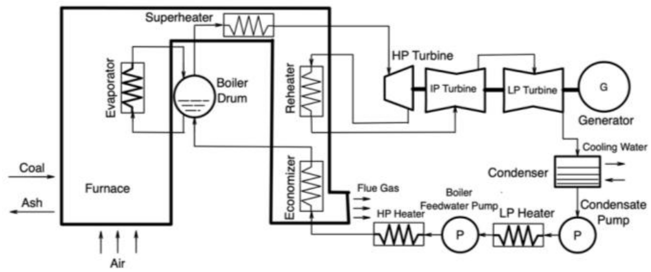


Figure 1. A simplified schematic diagram of boiler

Figure 1 shows a simplified schematic diagram of the target boiler system. Basically, the boiler generates steam by boiling feedwater using thermal energy converted from fossil fuels. In the process of generating steam, first, the feedwater is preheated by extracting steam from the turbine in a device called feedwater heater. The preheated feedwater is pumped to the economizer for reheating with flue gas and then fed to the boiler drum. The feedwater and saturated water in the drum are transported to the evaporator through the downcomer and become saturated steam by absorbing radiant heat from the furnace. The saturated water and steam are separated in the drum. The primary steam from the drum is converted into high-purity superheated steam by the superheater and fed to the high-pressure (HP) turbine. After being fed to the HP turbine, the primary steam is reheated by the reheater and fed to the intermediate-pressure (IP) turbine and the low-pressure (LP) turbine. The primary steam leaving the LP turbine is then condensed into condensate water and pushed back as feedwater by the boiler feed pump to the feedwater heater for preheating before being fed back to the boiler again. For more details on steam generation, see Kitto and Stultz (Babcock & Company, 1923) and Sarkar (2015).

Due to the continuous operation of boilers in power plants, some faults are inevitable to occur in their components. With the variation of operating conditions, the causes, types, and mechanisms of faults to boilers are also different. Therefore, boiler condition monitoring is needed which allows a real-time monitoring system. Process monitoring data which is historical data from many variables operating with different output power, can be used to assess the boiler operating conditions and detect whether the boiler is faults or still normal (Indrawan et al., 2021).

3. PROPOSED APPROACH

3.1. Framework

This paper proposes a method for boiler diagnostics and prognostics based on condition monitoring data using a framework as shown in Figure 2. The proposed framework mainly consists of the following:

1. Data acquisition and processing

   The boiler has been equipped with a SCADA system for remote monitoring which includes a data acquisition unit and data storage. Predictive models for diagnostics are developed using offline data, namely historical data available in the database system. The data used in developing the predictive model is operational data that includes normal and anomalous conditions. Data processing is carried out to improve data quality and to adapt the machine learning work environment used. The selection of sensor variables is intended to select sensor variables that are sensitive to anomalies (Khan et al., 2023). When the sensor variables are selected, the process continues with the extraction of several statistical features, and the remaining sensor variables will not be used. Due to the varying working conditions of the boiler, data normalization is also included in the data processing to eliminate the influence of the order of magnitude (Sola & Sevilla, 1997; Orrù et al., 2020). After normalizing the data, labels are added to the data according to the data source originating from normal or fault conditions. In this work, labels are intended for multiclassification algorithms in supervised learning to detect more than two types of faults in the boiler.

2. Module for boiler diagnostics

   This module is developed by training SVM and RFA with a portion of the input data set of sensor variables obtained in data processing. The prediction model is then tested by another portion of the input data set where this data has never been used for the training process. When good testing accuracy is achieved, the tested model will be considered as a prediction model for boiler diagnostics. The process of building a prediction model as described above is done offline. In addition, boiler fault diagnostics are performed based on an online process by recognizing some anomalous patterns contained in the sensor variable data. Once these patterns match the prediction model, the boiler condition is determined as well as anomaly detection.

3. Module for RUL boiler prognostics

   The sensor variable containing patterns anomalies are monitored continuously and tend to follow degradation of boiler condition. In fact, there is a degradation period of the boiler from the time an anomaly is detected to system failure. RUL can be predicted using a method based on the similarity between online degradation of sensor variable and historical degradation using DTW. The historical degradation data is decomposed into several windows of a certain length and the RUL is estimated for each window. Finally, RUL of system is predicted according to the period in the historical case in which the data characteristics are the most similar to those in the current period.

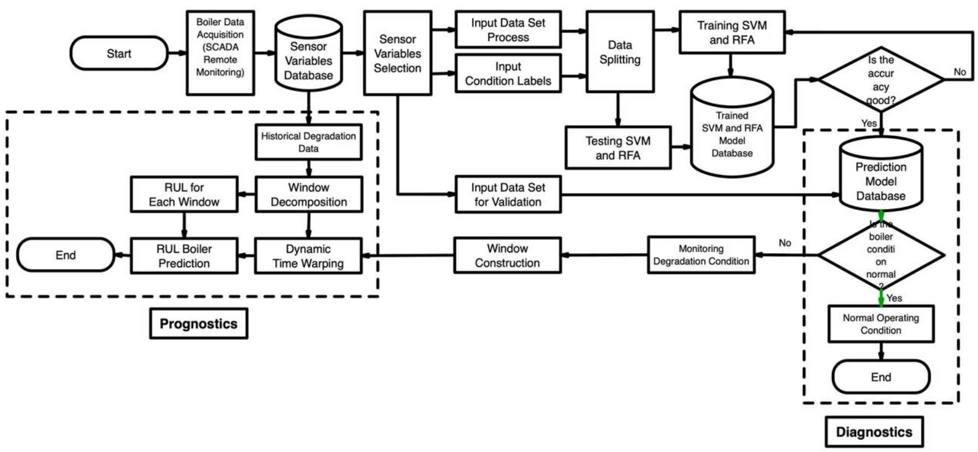


Figure 2. The framework of the proposed method

3.2. Anomaly Detection

The purpose of fault diagnostics is to monitor the operating conditions and isolate the sensor variables that represent the actual condition of the monitored equipment, whether it is normal or faulty. Due to the large number of sensor variables that need to be assessed, the main tasks of fault diagnostics include the selection of sensor variables and anomaly detection methods. Sensor variables are also known as temporal variables. The selection of sensor variables must be done because not all variables contain information that is closely related to anomalies. Dealing with all sensor variables is subject to high complexity and a lot of computation time. Therefore, the proposed method involves the selection of sensor variables that produce important features for SVM and RFA training.

SVM is included in the supervised learning method, namely a discriminative algorithm that separates examples of different class labels using a hyperplane. The solution is to find the optimal hyperplane that separates data that lies in opposite class labels so that it produces the maximum separation margin. SVM tries to place a linear boundary between two different classes, and orients it in such a way that the margin is maximized. In other words, SVM tries to find a boundary such that the distance between the boundary and the nearest data point in each class is maximized. Figure 3 shows the SVM hyperplane placed in the middle of the margin between two points.

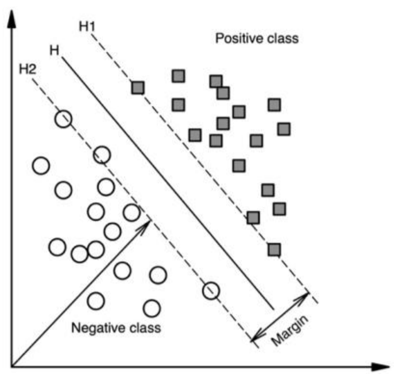


Figure 3. Classification in SVM using hyperplane

Given the data $D = \{(X_i, y_i)\}$, $y_i \in \{-1, +1\}$, $i = 1,\ldots l$ where X_i is input samples and y_i is output labels. These samples are assumed have two classes that is positive class and negative class. SVM aims to separate D by defining a hyperplane such that all data inputs in same class are on the same sides while maximizing the distance between two classes and separating hyperplane. The optimal separating hyperplane is presented by linear classifier as follows:

$$f(x) = sign\{\sum_{i=1}^{l} \lambda_i\, y_i\, X^T X_i + b\},\ i = 1, \ldots, l \qquad (1)$$

where $sign(*)$ is the $sign$ function and the Lagrange coefficients λ_i is the solution of the following quadratic programming problem

Maximize:

$$W(\lambda) = -\sum_{i=1}^{l} \lambda_i + \frac{1}{2}\sum_{i=1}^{l} \lambda_i\, \lambda_j\, y_i\, y_j\, X_i\, X_j \qquad (2)$$

Subject to:

$$\sum_{i=1}^{l} \lambda_i y_i = 0 \qquad (3)$$

$$\lambda_i \geq 0 \qquad (4)$$

In real industrial applications, classification problems usually involve data that can only be separated by nonlinear decision solutions. Therefore, the input data needs to be transformed into a high-dimensional space using a kernel function so that the nonlinear problem becomes linearly separable for the SVM classifier solution as presented in the following equation.

$$f(x) = sign\left\{\sum_{i=1}^{l} \lambda_i y_i K(X, X_i) + b\right\} \qquad (5)$$

The kernel functions applied to high-dimensional feature space is presented in Table 1 where, d is the degree of the polynomial, r is constant parameter and γ is the kernel width parameter.

Kernel	$K(X, X_i)$
Linear	$X^T X_i$
Polynomial	$(\gamma X^T X_i + r)^d$
Gaussian RBF	$\exp(-\|\| X - X_i\|\|^2 / 2\gamma^2)$
Sigmoid	$\tanh(\gamma X^T X_i + r)$

Table 1. Formulation of the kernel functions: $K(X,X_i)$

Another classification technique used for the proposed method is the random forest algorithm (RFA). RFA is an ensemble learning method that can be used for classification tasks based on the construction of a multitude of decision trees. RFA was created by Ho (Tin Kam, 1995) and an extension of this algorithm was developed by Breiman (Breiman, 1996, Breiman, 2001) which combines bagging and random feature selection methods to build a collection of decision trees controlled by variance.

Given k random vectors $\boldsymbol{\theta}_k$ independent of the previous random vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{k-1}$ but with the same distribution to build a tree between RF. The corresponding individual classifiers are denoted by $C(\mathbf{X}, \boldsymbol{\theta}_k)$, where X is the input vector. In the bagging process, random vectors $\boldsymbol{\theta}_k$ for observation number N are randomly drawn proportionally from the entire training data. This is referred to as random forest (RF) by Breiman (1996, 2001).

RF deals with an ensemble of classifiers series $C_1(\mathbf{X})$, $C_2(\mathbf{X})$, …, $C_k(\mathbf{X})$ and with the training process of a random data set from the distribution of random vectors Y, X, the margin is defined as

$$mg(\mathbf{X}, Y) = av_k I(C_k(\mathbf{X}) = Y) - max_{j \neq Y} av_k I(C_k(\mathbf{X}) = j) \qquad (6)$$

where Y, av_k and $I(\cdot)$ are the corresponding vector for class, the averages number of votes at $\mathbf{X}$ and the indicator function, respectively. In this case, the larger the margin, the more accurate the classification. In addition, RF generalization errors are given by

$$PE' = P_{X,Y}(mg(X,Y) < 0) \tag{7}$$

where PE' is the probability that X, Y space is exceeded.

In addition, one of the advantages of RF is that it is less likely to overfit when the number of trees increases. As the number of trees increases for almost all orders of θ_k, PE' converges to the following form:

$$P_{X,Y}\big(P_{\theta}(C(X,\theta)=Y) - max_{j\neq Y}P_{\theta}(C(X,\theta)=j) < 0\big) \tag{8}$$

RF does not experience overfitting when more trees are added while it produces limited generalization error values.

In this paper, SVM and RFA are trained using input vectors selected from available sensor variables in the boiler monitoring process. The trained SVM and RFA are then used to predict class labels from new sensor variable data obtained from online process monitoring to detect the actual boiler condition. The predicted label outputs mean diagnostics of the actual condition as well as detection of boiler anomalies.

3.3. Prediction of RUL

RUL prediction is performed using a similarity-based method through two-variable DTW, namely historical degradation and online sensor variable degradation. In the RUL prediction study, DTW was reported by Barr' who applied health prognostics to electric vehicle batteries (Barr et al., 2014). Tao, et al. applied the similarity recognition method of online charging and discharging data curves using spatial DTW to estimate the capacity of lithium-ion batteries (Tao et al., 2015). Que and Xu (2019) used DTW for steam turbine RUL prediction based on generator output power data because it has a good degradation tendency.

Given two data sequences of time series $P = \{p_1, p_2, \ldots, p_N\}$ and $Q = \{q_1, q_2, \ldots, q_M\}$ the time warping distance between P and Q is defined recursively as follows

$$D_{dtw}(P,Q) = d(M,N) \tag{9}$$

$$d(i,j) = \left(q_i - p_j\right)^2 + min\begin{cases} d(i,j-1) \\ d(i-1,j) \\ d(i-1,j-1) \end{cases} \tag{10}$$

$$d(0,0) = 0, d(i,0) = 0, d(0,j) = \infty, (i = 1,\ldots,N;\ j = 1,\ldots,M) \tag{11}$$

where $d(i,j)$ is the optimal distance between the first i and the first j elements of two time series P and Q.

Assume that P is historian time series of sensor variables in the database with length L_P, and Q is a new time series of same sensor variables coming from online with length L_Q. The similarity of both time series is measured by introducing a size of L_P-by-L_Q matrix called **SM**. The $SM_{i,j}$ stands for the distance between points p_a and q_b in the series as follows

$$SM_{a,b} = (q_b - p_a)^2 \tag{12}$$

where a = 1,2, …, L_Q and b = 1, 2, …, L_P. Figure 4(a) shows the example of matrix **SM** that measures distances between P and Q as formulated in Eq. (10) with P = {1, 3, 5, 6, 9, 11, 12, 13} and Q = {2, 4, 6, 8, 10, 12}. The optimal distance which is regarded as warping time path is calculated using Eq. (10) as shown in Figure 4(b).

P_j							
	13	121	81	49	25	9	1
	12	100	64	36	16	4	0
	11	81	49	25	9	1	1
	9	49	25	9	1	1	9
	6	16	4	0	4	16	36
	5	9	1	1	9	25	49
	3	1	1	9	25	49	81
	1	1	9	25	49	81	121
		2	4	6	8	10	12
				Q_i			

(a)

P_j							
	13	221	145	85	41	13	1
	12	181	113	61	25	5	1
	11	130	74	34	10	2	2
	9	65	29	9	1	2	10
	6	25	5	1	4	20	52
	5	10	2	2	10	34	74
	3	2	2	10	34	74	130
	1	1	10	34	74	130	202
		2	4	6	8	10	12
				Q_i			

(b)

Figure 4. Illustration of DTW with L_P = 8 and L_Q = 6: (a) Matrix **SM** for distances between p_a and p_b; (b) optimal distance of warping time

The RUL prediction method starts with the decomposition of the historical degradation time series P into m closed windows of fixed length L_m such that P = {P_1, P_2, …, P_m}. This decomposition allows some overlap between the decomposed windows as depicted in Figure 5. Let L_n be the length of the time series during the degradation period, then the RUL for each decomposed window is determined as follows.

$$RUL_{P_k} = t(L_n - kL_m) \tag{13}$$

where k = 1, 2, …, m and t is sampling interval of time series.

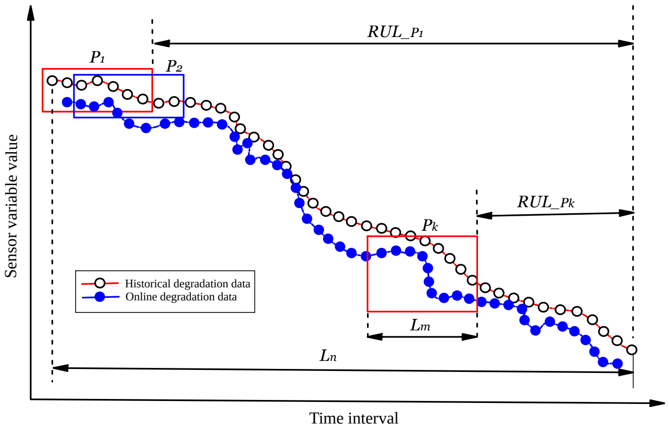


Figure 5. Decomposition windows of historical time series degradation for case learning

In the online monitoring process, new data is sampled at time t_{new} and a new window is constructed as Q_{new}. The window length of Q_{new} is L_q which can be different from L_m. DTW is used to calculate the similarity of P_k and Q_{new}. The shortest distance between P_k and Q_{new} is considered as P_{Sim} which means the data series that is most similar to P_k in the new series Q_{new}. After P_{Sim} is determined, the boiler RUL is calculated using Eq. (13).

4. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed method is applied for boiler diagnostics in a power plant using a real data set collected from a SCADA system. The data set consists of 37 sensor variables recorded every minute using a SCADA system. The data set was obtained from three boilers namely SLA5, SLA6, and SLA7 which are considered as similar boilers in the power plant.

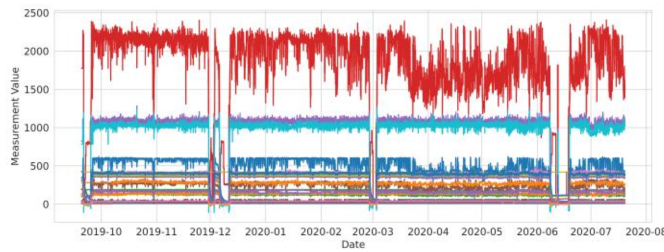


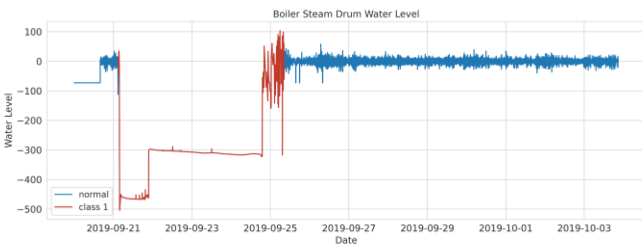
Figure 6. Presentation of 37 sensor variables of boiler SLA5 in single frame line plot

Table 2 shows the sensor variables used in the experimental work and the data presentation is presented in Figure 6. Machine learning (ML) training was conducted using ground truth data obtained from faulty boilers. The ground truth data includes 8 (eight) types of faults recorded with varying durations from September 2019 to May 2020. In addition, a set of boiler data under normal operating conditions was also added to the training data. As an example of ground truth data, the presentation of data when a boiler leak occurs in the superheater and in the header drain is depicted in Figure 7.

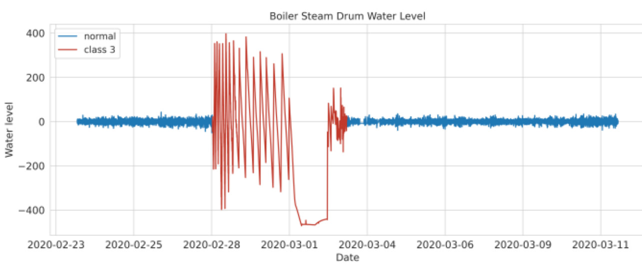
No	Sensor variable data	No	Sensor variable data
1	Auxiliary Steam Header Pressure	20	Primary Air Duct Pressure
2	Boiler Feed Pump T A: Feedwater Outlet Temperature	21	PAH B: inlet Air Temperature
3	Boiler Feed Pump-Turbine (BFPT) B: Feedwater Outlet Temperature	22	PAH B: Outlet Air Temperature
4	Boiler Total Air Flow	23	PAH B: Outlet Flue Gas Temperature
5	Boiler Steam Drum Pressure	24	Riser To Steam Drum Side A Water Temperature
6	Boiler Total Coal Flow	25	Riser To Steam Drum Side B Water Temperature
7	Booster BFPT A: Outlet Pressure	26	Secondary Air Heater (SAH) A Differential Pressure
8	Booster BFPT B: Outlet Pressure	27	SAH A: Outlet Pressure
9	Economizer Outlet Feedwater Side B Temperature	28	SAH B: Differential Pressure
10	Economizer Outlet Flue Gas Temperature	29	SAH B Outlet Pressure
11	Generator Gross Capacity	30	Secondary Super Heater (SSH) Inlet Steam Side A: Temperature
12	Primary Air Heater (PAH) A: Differential Pressure	31	SSH Inlet Steam Side B: Temperature
13	PAH A: Outlet Pressure	32	Secondary Air Duct Pressure
14	PAH B: Differential Pressure	33	SAH A: Inlet Air Temperature
15	PAH B: Outlet Pressure	34	SAH B: Inlet Air Temperature
16	Primary Super Heater (PSH) Inlet Steam Side A Temperature	35	SAH B: Outlet Flue Gas Temperature
17	PSH Inlet Steam Side B Temperature	36	Soot Blower Steam Side A Pressure
18	PSH Outlet Steam Side A Temperature	37	Soot Blower Steam Side B Pressure
19	PSH Outlet Steam Side B Temperature		

Table 2. The sensor variables acquired by SCADA system

The figure shows a comparison of sensor variable data when the boiler is in normal condition and leaking. Table 3 shows the class and composition of training data to build a prediction model. Considering Table 3, the training process for ML must be carried out using a multiclassification strategy so that learning can capture all features of the boiler condition. Before the training process is carried out, the data needs to be preprocessed to get the best training performance. First, exploratory data analysis (EDA) is conducted to assess data quality and extract meaningful insights. This step involves removing noise and outliers, as well as imputing missing values in the dataset. Second, data normalization is applied using a scaling method to rescale the sensor data values to a range of 0-1. This is necessary due to the significant differences in the magnitude of the sensor data values.



(a)



(b)

Figure 7. Sensor variable data of boiler: (a) Leak at superheater (c1); (b) Leak on header drain (c3)

Class	Condition remarks	# of data
c0	Normal condition	1508
c1	Leak in superheater	419
c2	Leak at convection pass wall (CPW)	244
c3	Leak on header drain HCP-PSH	207
c4	Noise in secondary superheater	176
c5	Leaks (unspecified)	164
c6	Leak in secondary superheater	105
c7	Leaks at sootblower	103
c8	Leak in primary superheater	82

Table 3 Class and composition of training data

Next, feature extraction is performed to transform the data from the original time series into a certain form that has minimum noise and can represent the data trend. In this study, five statistical features, namely mean, median, variance, skewness and kurtosis, were extracted from the sensor variable data. In total, there were 185 features extracted from 37 sensor variables. The extracted features were then normalized using the min-max formula to eliminate the order of magnitude between the data.

4.1. Diagnostics Prediction Model

4.1.1. SVM Prediction Model

The SVM model predictions are generated by training the SVM using 70% of the data features and the remaining 30% of the data features are used to validate the trained model. The training process includes the selection of the kernel function and hyperparameters of the kernel function used in the SVM. The selection of hyperparameters (also known as tuning) is very important in SVM learning because it affects the accuracy of the prediction model. In training, four kernel functions as presented in Table 1 are used and several ranges of hyperparameter values are investigated regarding the training accuracy. The target hyperparameter ranges are shown in Table 4.

If gamma_range = 'scale' is passed then it uses 1/(n_features*X.var) as the gamma value, where n_features and X.var are the number of features and their variances respectively. The selection of appropriate hyperparameters involved 48 combinations performed using the random search cross-validation method with 5 folds and 100 iterations.

Hyperparameters	Options
C	0.1, 0.2, 0.5, 0.8, 1, 2, 5, 10
Gamma	0.00001, 0.0001, 0.001, 0.01, 0,1, 'scale'
Kernel	linear, sigmoid, rbf, poly

Table 4. Hyperparameter options for SVM training

The results of the SVM training are summarized in Table 5. The best SVM model obtained from this training used the RBF kernel function with hyperparameters *C* and *Gamma* of 0.8 and 0.1, respectively.

Kernel functions	*C*	*Gamma*	Validation accuracy (%)	Test accuracy (%)
Linear	0.2	0.00001	92.10	91.93
Sigmoid	5	0.01	90.40	88.85
Gaussian RBF	**0.8**	**0.1**	**95.50**	**94.51**
Polynomial	0.5	0.1	93.50	94.17

Table 5. Results of SVM training

In addition, this study also utilizes principal component analysis (PCA) to reduce features from the original dataset while retaining as much information as possible from the original data (Świercz & Mroczkowska, 2019; Khan et al., 2023). Another benefit of using PCA is that it is very reliable in small datasets (Martinez & Kak, 2001). The feature reduction scenario is included in the SVM training to find the best feature set shown in Table 6. There are 384 combination pairs that have been calculated in finding the proper hyperparameters using random search with 5-fold cross validation and 100 iterations. The results of SVM training with PCA feature reduction are presented in Table 7. The SVM model using the RBF kernel function with *C* = 0.8 and *Gamma* = ‘scale’ is the best model obtained through training with 150 principal components (PC).

Hyperparameters	Options
C	0.1, 0.2, 0.5, 0.8, 1, 2, 5, 10
Gamma	0.00001, 0.0001, 0.001, 0.01, 0,1, ‘scale’
Kernel	linear, sigmoid, rbf, poly
#PC	5, 10, 15, 20, 25, 50, 100, 150

Table 6. Hyperparameter options for SVM and PCA

Kernel functions	# PC	*C*	*Gamma*	Validation accuracy (%)	Test accuracy (%)
Linear	150	0.2	0.01	94.70	95.69
Sigmoid	150	10	0.01	94.30	95.69
Gaussian RBF	150	**0.8**	**‘scale’**	**96.40**	**95.81**
Polynomial	150	10	0.1	95.90	95.14

Table 7. Results of SVM training with PCA

4.1.2. RFA Prediction Model

RFA training involves fine-tuning the hyperparameters to obtain the best RFA model. The hyperparameters in RFA are tuned within the following ranges as shown in Table 8.

Hyperparameters	Options
Estimators	500, 666, 833, 1000, 1166, 1333, 1500, 1666, 1833, 2000
max_features	log2, sqrt
max_depths	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
min_samples_split	2, 7, 12, 18, 23, 28, 34, 39, 44, 50
min_samples_leaf	2, 7, 12, 18, 23, 28, 34, 39, 44, 50

Table 8. Hyperparameter options for RFA

Hyperparameter tuning was performed using the random search cross-validation method with 5 folds and 100 iterations. There are 30,000 pairs of hyperparameter combinations that have the potential to be the best parameter candidates. The selected ‘sqrt’ means the number of features used for splitting nodes is the square root of the total number of features. The rationale behind this setting is to introduce randomness in the decision-making process of each tree, helping to prevent overfitting and ensuring that the trees in the forest are diverse. This randomness makes the model more robust and often leads to better generalization on unseen data. After tuning, the best hyperparameter pairs are obtained as summarized in Table 9. These parameters yield excellent training performance, achieving 97.20% accuracy on the validation set and 98.30% on the testing set for the best RFA model.

Hyperparameters	Options
Estimators	500
max_features	sqrt
max_depths	9
min_samples_split	7
min_samples_leaf	2

Table 9. Selected hyperparameter options for RFA training

4.1.3. Boiler Diagnostics

Boiler anomaly detection and diagnostics are performed by testing the trained SVM and RFA models using new sensor variable data collected from boilers SLA5, SLA6, and SLA7 from July 20 to September 10, 2020. These data are processed using a method similar to the training data as described previously. The anomaly detection prediction accuracy for all boilers is summarized in Table 10.

Table 10 shows that the prediction of SLA5 boiler conditions is dominated by normal conditions (c0) with prediction accuracies of 81.30%, 83.42% and 99.92% for SVM, SVM-PCA and RFA, respectively. Another condition predicted by the model is noise in the secondary superheater (c4). However, the prediction accuracy of c4 is relatively low, only reaching a maximum of 18.70%, 16.58% and 0.08% for the SVM, SVM-PCA and RFA models, respectively. This means that noise in the secondary superheater does not occur in the SLA5 boiler. In addition, other class patterns such as c1, c2, c3, c6 and c8 are not captured by the model because there are no features related to these conditions in the sensor variables.

Boiler	Class predicted (accuracy in %)		
	SVM	SVM-PCA	RFA
SLA5	c0 (**81.30**)	c0 (**83.42**)	c0 (**99.92**)
	c4 (18.70)	c4 (16.58)	c4 (0.08)
SLA6	c0 (**77.77**)	c0 (**80.60**)	c0 (**100.00**)
	c4 (22.23)	c4 (19.40)	
SLA7	c0 (**90.49**)	c0 (**88.61**)	c0 (**93.56**)
	c2 (0.08)	c2 (0.08)	c4 (0.24)
	c3 (0.08)	c4 (1.96)	c5 (1.41)
	c4 (1.65)	c5 (4.32)	c6 (0.55)
	c5 (4.24)	c7 (4.79)	c7 (0.08)
	c7 (3.22)	c8 (0.24)	c8 (4.16)
	c8 (0.24)	-	-

Table 10. Anomaly detection prediction accuracy for all boilers

The prediction of the SLA6 boiler condition gives a normal condition (c0) of 77.77% and 80.60% for the SVM and SVM-PCA models, respectively. The noise condition in the secondary superheater is also captured by the SVM and SVM-PCA models with prediction accuracies of 22.23% and 19.40%, respectively. This accuracy is relatively low and it is difficult to consider the actual condition of the SLA6 boiler. The RFA prediction model only captures the normal condition (c0) with an accuracy of 100%. This experiment concludes that the condition of the SLA6 boiler is normal.

The condition of the SLA7 boiler is normal (c0) based on predictions from all models with accuracies of 90.49%, 88.61% and 93.56% for the SVM, SVM-PCA and RFA models, respectively. All prediction models agree that the condition of the SLA7 boiler is normal even though there are other conditions that are also captured but the accuracy is relatively low. The highest prediction accuracy for abnormal conditions is only 4.79% for leaks in the sootblower (c7) and the others are lower.

By observing Table 10, the prediction results are consistent for all model predictions. The boiler condition is predicted under normal operating conditions based on the test data. In this work, the use of PCA for dimensionality reduction does not provide significant improvement in SVM accuracy. As presented in the training, the best selection of the number of principal components results in 150 principal components for the best SVM prediction model. SVM with PCA provides better accuracy than the SVM prediction model in detecting normal conditions for SLA5 and SLA6 boilers although only slightly different. In the prediction of anomalies such as noise in the secondary superheater (c4) SVM is better than SVM-PCA in prediction accuracy but also only slightly different. The difference in prediction accuracy is not more than 3% concluding that the two models are similar. In addition, the prediction of the SLA7 boiler condition provides normal conditions with 90.49% and 88.61% for the SVM and SVM-PCA prediction models, respectively. There is a slight difference in prediction accuracy which is less than 2% indicating that the models are similar. In addition, the prediction model can capture and recognize features derived from other anomalies as summarized in Table 10. At least five anomalies were successfully detected by the prediction model but the accuracy was very low. This happened to both the SVM and SVM-PCA prediction models and both gave little difference.

The performance of RFA prediction model outperforms SVM and SVM-PCA prediction models as shown in Table 10. In the case of RFA performance, Han et al. (2018) also reported similar results that RFA outperforms SVM in terms of recognition accuracy in a study of intelligent diagnosis of rotating machinery. Another paper reviewing the application of ML in predictive maintenance has mentioned the performance of RFA outperforming other ML techniques (Çınar et al., 2020). In the condition prediction of SLA5 and SLA6 boilers, the RFA prediction model has almost 100% accuracy in predicting the normal conditions of these boilers. Meanwhile, the normal condition prediction of SLA7 boiler using RFA also gives better accuracy than SVM. The RFA prediction model can recognize other anomalies in SLA7 boiler, but its accuracy is low.

The performance of the prediction model in this work certainly cannot be applied generally to all boilers. Machine learning has special characteristics in terms of building predictive models and depends on the data used in training. The prediction target must come from the system where the prediction model is built. In addition, the architecture of the prediction model is also an important phase as well as the design of the classification rules. Basically, the more examples for ML training, the better the performance of the prediction model will be obtained.

4.2. RUL Prediction Model

The use of machine learning techniques, specifically SVM and RFA methods, for boiler diagnos**tics** has been thoroughly explored with the goal of identifying the root cause of malfunctions. In this section, boiler prognostics based on DTW will be discussed, with the aim of predicting the **RUL** of the boiler before failure occurs**.** Similarity-based DTW is constructed to build a prediction model for RUL **prediction** as discussed in Section 3.3. The selection of sensor variables for prediction input purpose is done by training SVM and RFA independently to find the most influential sensor variables in the prediction model for anomaly detection. The top five selected sensor variables with the highest scores are as follows: Generator gross capacity, Booster BFPT A: Outlet pressure, Booster BFPT B: Outlet pressure, Riser to steam drum side B: Water temperature, and Boiler steam drum pressure.

The online data used to test the RUL prediction was recorded from July 30-31, 2020. Figure 8 shows the anomaly detection prediction based on the generator gross capacity sensor variable degradation. The left and right axes represent the generator gross capacity and the class (or anomaly) condition of the SLA7 boiler, respectively. Initially, the boiler condition was still in normal operation until condition c5 was detected. However, the detection of c5 was in a very short

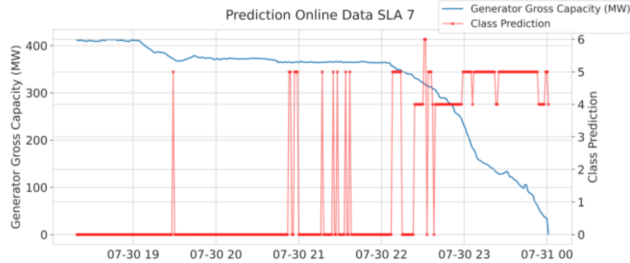


Figure 8. Anomaly detection of SLA7 boilers based on generator gross capacity degradation

duration at the initial stage and then changed to condition c0 otherwise but still in a short time. Condition c4 was also detected but only for a few minutes before the final degradation. Finally, condition c5 was detected again until the final degradation based on the generator gross capacity sensor variable.

Once the anomaly is detected, the prediction system starts working to execute the RUL prediction model based on the similarity between the degradation patterns contained in the online data and in the historical database through DTW. The RUL prediction results are presented in Figure 9. In the early stages, the RUL prediction using DTW for time data 0 - 70 minutes remains relatively constant, as the data closely resembles normal conditions, with the RUL around 191 minutes.. Then for time data 71 - 140 minutes, the RUL prediction results vary and are unstable because there is only a little degradation in the data sensor. The RUL prediction produces a relatively constant value for time data 141 - 225 minutes because the sensor variables also tend to be constant values and do not show degradation. After the 226th minute, the RUL sharply declines from 170 to approximately 80 minutes, following a substantial reduction in the generator gross capacity. Finally, the RUL decreases from around 80 minutes to the end of life for time data 226 - 344 minutes and the generator's gross capacity has reached zero.

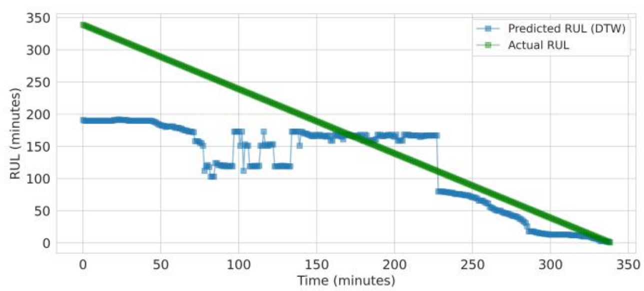


Figure 9. RUL prediction of boiler

It can be found that from Figure 9, the predicted RUL is relatively far from the actual RUL at the beginning prediction stage until reaches time data around 140 minutes. This is because the decline rate of degradation of the sensor generator gross capacity also very small. At above time

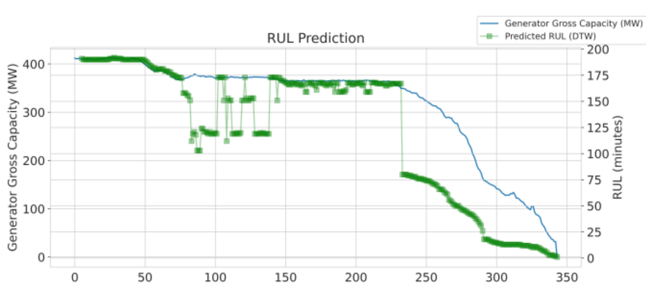


Figure 10. Predicted RUL according to generator gross capacity

range, the historical degradation data P_{sim} cannot truly approach the path of degradation tendency of Q_{new} from online data. The predicted RUL is improved and get near to the actual RUL for time data 140 – 200 minutes but after that tends to away again from actual RUL until reaches time data around 226 minutes. The RUL prediction improves after the 226th minute and continues to closely align with the actual RUL in the final stages, as shown in Figure 9. Overall, the prediction errors are 73.47 for RMSE and 56.36 for MAE.

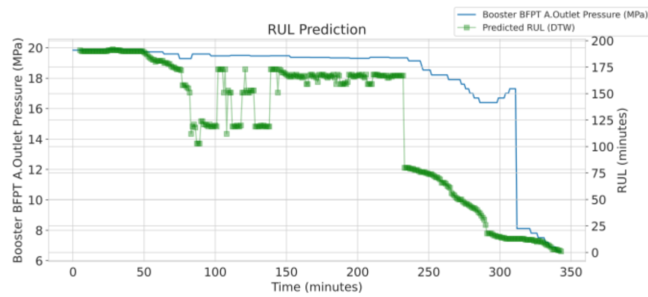


Figure 11. Predicted RUL according to booster BFPT A: outlet pressure

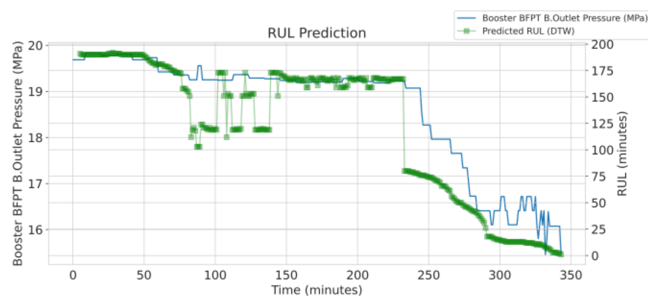


Figure 12. Predicted RUL according to booster BFPT B: outlet pressure

In the RUL prediction, the degradation information from the sensor variables is important parameter as well as degradation trend that produce an obvious gradually trend. When the decline rate of degradation is gradually accelerated as depicted in the final period of RUL, the result of prediction can be improved. This means that the historical degradation data P_{sim} can well approach the path of the degradation tendency of Q_{new} obtained from online system. Besides that, Figure 10-12 show the predicted RUL based on selected sensor variables used in the RUL prediction training. The selected sensor variables report similar trends when the boiler experiences anomalies and suffers degradation conditions. In this work, the results of RUL prediction around 191 minutes based on real system data seems unreliable but this fact cannot be avoided because of the real system can only provide very limited degradation data from boiler operating data for training. The proposed system can capture the degradation presented in the sensor variable with respect to predicted RUL especially starting from 226th minutes and showing consistency for top five selected sensor variables.

5. CONCLUSIONS

This paper presents a study on boiler diagnostics and prognostics based on real data obtained from an industrial power plant. Boiler diagnostics are performed using multiclass classification methods through SVM and RFA. The proposed method demonstrates that anomaly detection is effective and can be used for the early identification of boiler conditions, particularly those associated with common fault types such as leakage. Predictive models for diagnostics and anomaly detection are developed using ground truth data obtained when the boiler experiences some anomalies. The quality of ground truth data is very important and determines the quality of the prediction model obtained from ML training. Ground truth data with prominent fault features and sufficient in size and time duration will produce good prediction models for diagnostics and prognostics. In this work, the proposed method successfully detects the boiler condition when new sensor variables are introduced into the system. Therefore, it can be applied in boiler condition monitoring and can help operators or operation managers to know the boiler status early or to make decisions in case of undesirable conditions. In addition, the prognostic task is also confirmed through boiler RUL prediction using similarity-based DTW utilizing historical and current degradation data. In this work, historical degradation data serves as a reference to determine new paths of degradation conditions related to RUL prediction. Once historical degradation data is recorded at a gradual decline rate, it can serve as a good reference for RUL prediction based on the similarity method.

It should be noted that in the presented study, all characteristics in the data and operating conditions cannot be involved, which causes difficulties and some limitations as follows:

1. The exact time when the boiler starts to experience anomalies is difficult to measure precisely. Training data for anomaly detection is collected from the database in the power plant based on operator reports.
2. In real systems, the dynamics of degradation that change over time cannot be captured in the overall situation, therefore the prediction model for RUL prediction still has limitations and produces relatively short predictions.
3. Operating conditions in power plants can change in a relatively short period, i.e., the load so that operating data becomes temporary while anomaly detection is developed based on training steady-state data, so this is a real challenge in fault detection studies.

ACKNOWLEDGEMENT

Authors would like to thank to PT PLN Indonesia Power for supporting this study through permission in use real boiler operating data. Thanks, are also extended to Yasir Abdur Rahman and Eko Abdul Aziz for their assistance with computer programming.

REFERENCES

Acuña, D. E. & Orchard, M. E. 2017. Particle-Filtering-Based Failure Prognosis Via Sigma-Points: Application To Lithium-Ion Battery State-Of-Charge Monitoring. *Mechanical Systems And Signal Processing,* 85**,** 827-848. https://doi.org/10.1016/j.ymssp.2016.08.029

Alegeh, N., Shagluf, A., Longstaff, A. P. & Fletcher, S. 2019. Accuracy In Detecting Failure In Ballscrew Assessment Towards Machine Tool Servitization. *International Journal Of Mechanical Engineering And Robotics Research,* 8**,** 667-673. http://dx.doi.org/10.18178/ijmerr.8.5.667-673

Babcock & Company, W. 1923. *Steam: Its Generation And Use*, Babcock & Wilcox Company.

Barr, A., Suard, F., Gérard, M. & Riu, D. A Real-Time Data-Driven Method For Battery Health Prognostics In Electric Vehicle Use. Phm Society European Conference, 2014. https://doi.org/10.36001/phme.2014.v2i1.1514

Berahman, M., Safavi, A. & Shahrbabaki, M. R. Fault Detection In Kerman Combined Cycle Power Plant Boilers By Means Of Support Vector Machine Classifier Algorithms and PCA. The 3rd International Conference On Control, Instrumentation, And Automation, 2013. IEEE, 290-295. https://doi.org/10.1109/ICCIAutom.2013.6912851

Bertolini, M., Mezzogori, D., Neroni, M. & Zammori, F. 2021. Machine Learning For Industrial Applications: A Comprehensive Literature Review. *Expert Systems With Applications,* 175**,** 114820. https://doi.org/10.1016/j.eswa.2021.114820

Breiman, L. 1996. Bagging Predictors. *Machine Learning,* 24**,** 123-140. https://doi.org/10.1007/BF00058655

Breiman, L. 2001. Random Forests. *Machine Learning,* 45**,** 5-32. https://doi.org/10.1023/A:1010933404324

Chen, K.-Y., Chen, L.-S., Chen, M.-C. & Lee, C.-L. 2011. Using Svm Based Method For Equipment Fault Detection In A Thermal Power Plant. *Computers In Industry,* 62**,** 42-50. https://doi.org/10.1016/j.compind.2010.05.013

Çınar, Z. M., Abdussalam Nuhu, A., Zeeshan, Q., Korhan, O., Asmael, M. & Safaei, B. 2020. Machine Learning In Predictive Maintenance Towards Sustainable Smart Manufacturing In Industry 4.0. *Sustainability,* 12**,** 8211. https://doi.org/10.3390/su12198211

Cui, Y. J., Xia, L. W., Huang, Y. & Ma, X. C. Research On Fault Diagnosis And Early Warning Of Power Plant Boiler Reheater Temperature Deviation Based On Machine Learning Algorithm. 2020 IEEE 6th

International Conference On Control Science And Systems Engineering (ICCSSE), 17-19 July 2020 2020. 212-216. https://doi.org/10.1109/ICCSSE50399.2020.9171950

Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J. & Barbosa, J. 2020. Machine Learning And Reasoning For Predictive Maintenance In Industry 4.0: Current Status And Challenges. *Computers In Industry,* 123**,** 103298. https://doi.org/10.1016/j.compind.2020.103298

Diez-Olivan, A., Del Ser, J., Galar, D. & Sierra, B. 2019. Data Fusion And Machine Learning For Industrial Prognosis: Trends And Perspectives Towards Industry 4.0. *Information Fusion,* 50**,** 92-111. https://doi.org/10.1016/j.inffus.2018.10.005

Do, P., Voisin, A., Levrat, E. & Iung, B. 2015. A Proactive Condition-Based Maintenance Strategy With Both Perfect And Imperfect Maintenance Actions. *Reliability Engineering & System Safety,* 133**,** 22-32. https://doi.org/10.1016/j.ress.2014.08.011

Duong, B. P., Kim, J., Kim, C.-H. & Kim, J.-M. 2019. Deep Learning Object-Impulse Detection For Enhancing Leakage Detection Of A Boiler Tube Using Acoustic Emission Signal. *Applied Sciences* [Online], 9. https://doi.org/10.3390/app9204368

Dzikuć, M., Kuryło, P., Dudziak, R., Szufa, S., Dzikuć, M. & Godzisz, K. 2020. Selected Aspects Of Combustion Optimization Of Coal In Power Plants. *Energies,* 13**,** 2208. https://doi.org/10.3390/en13092208

Han, T., Jiang, D., Zhao, Q., Wang, L. & Yin, K. 2018. Comparison Of Random Forest, Artificial Neural Networks And Support Vector Machine For Intelligent Diagnosis Of Rotating Machinery. *Transactions Of The Institute Of Measurement And Control,* 40**,** 2681-2693. https://doi.org/10.1177/0142331217708242

Hong-Feng, W. 2012. Prognostics And Health Management For Complex System Based On Fusion Of Model-Based Approach And Data-Driven Approach. *Physics Procedia,* 24**,** 828-831. https://doi.org/10.1016/j.phpro.2012.02.124

Hong, F., Rui W., Jie S., Mingming, G., Jizhen, L., & Dongteng, L.. 2022. A performance evaluation framework for deep peak shaving of the CFB boiler unit based on the DBN-LSSVM algorithm, *Energy* 238, 121659. https://doi.org/10.1016/j.energy.2021.121659

Indrawan, N., Shadle, L. J., Breault, R. W., Panday, R. & Chitnis, U. K. 2021. Data Analytics For Leak Detection In A Subcritical Boiler. *Energy,* 220**,** 119667. https://doi.org/10.1016/j.energy.2020.119667

Jia, X., Yichen, S., Yanjun, L., Wei, D., & Guolei, Z.. 2022. Short-term forecasting for supercharged boiler safety performance based on advanced data-driven modelling framework, *Energy* 239,122449. https://doi.org/10.1016/j.energy.2021.122449

Khalid, S., Lim, W., Kim, H. S., Oh, Y. T., Youn, B. D., Kim, H.-S. & Bae, Y.-C. 2020. Intelligent Steam Power Plant Boiler Waterwall Tube Leakage Detection Via Machine Learning-Based Optimal Sensor Selection. *Sensors,* 20**,** 6356. https://doi.org/10.3390/s20216356

Khan, P. W., Yeun, C. Y. & Byun, Y. C. 2023. Fault Detection Of Wind Turbines Using Scada Data And Genetic Algorithm-Based Ensemble Learning. *Engineering Failure Analysis,* 148**,** 107209. https://doi.org/10.1016/j.engfailanal.2023.107209

Khan, T., Udpa, L. & Udpa, S. Particle Filter Based Prognosis Study For Predicting Remaining Useful Life Of Steam Generator Tubing. 2011 IEEE Conference On Prognostics And Health Management, 20-23 June 2011 2011. 1-6. https://doi.org/10.1109/ICPHM.2011.6024323

Lei, Y., Li, N., Guo, L., Li, N., Yan, T. & Lin, J. 2018. Machinery Health Prognostics: A Systematic Review From Data Acquisition To Rul Prediction. *Mechanical Systems And Signal Processing,* 104**,** 799-834. https://doi.org/10.1016/j.ymssp.2017.11.016

Li, Z., Liu, R. & Wu, D. 2019. Data-Driven Smart Manufacturing: Tool Wear Monitoring With Audio Signals And Machine Learning. *Journal Of Manufacturing Processes,* 48**,** 66-76. https://doi.org/10.1016/j.jmapro.2019.10.020

Liao, L. & Köttig, F. 2014. Review Of Hybrid Prognostics Approaches For Remaining Useful Life Prediction Of Engineered Systems, And An Application To Battery Life Prediction. *IEEE Transactions On Reliability,* 63**,** 191-207. https://doi.org/10.1109/TR.2014.2299152

Lin, W. C. & Ghoneim, Y. A. Model-Based Fault Diagnosis And Prognosis For Electric Power Steering Systems. 2016 IEEE International Conference On Prognostics And Health Management (Icphm), 20-22 June 2016 2016. 1-8. https://doi.org/10.1109/ICPHM.2016.7542840

Madrigal-Espinosa, G., Osorio-Gordillo, G. L., Astorga-Zaragoza, C. M., Vázquez-Román, M. & Adam-Medina, M. 2017. Fault Detection And Isolation System For Boiler-Turbine Unit Of A Thermal Power Plant. *Electric Power Systems Research,* 148**,** 237-244. https://doi.org/10.1016/j.epsr.2017.03.021

Martinez, A. M. & Kak, A. C. 2001. Pca Versus Lda. *IEEE Transactions On Pattern Analysis And Machine Intelligence,* 23**,** 228-233. https://doi.org/10.1109/34.908974

Montero Jimenez, J. J., Schwartz, S., Vingerhoeds, R., Grabot, B. & Salaün, M. 2020. Towards Multi-Model Approaches To Predictive Maintenance: A Systematic Literature Survey On Diagnostics And Prognostics. *Journal Of Manufacturing Systems,* 56**,** 539-557. https://doi.org/10.1016/j.jmsy.2020.07.008

Mushiri, T., Mhazo, T. K. & Mbohwa, C. 2018. Condition Based Monitoring Of Boiler Parameters In A Thermal Power Station (Case Of Anonymous Company). *Procedia Manufacturing*, 21**,** 369-375. https://doi.org/10.1016/j.promfg.2018.02.133

Nguyen, H.-P., Fauriat, W., Zio, E. & Liu, J. A Data-Driven Approach For Predicting The Remaining Useful Life Of Steam Generators. 2018 3rd International Conference On System Reliability And Safety (Icsrs), 2018. IEEE, 255-260. https://doi.org/10.1109/ICSRS.2018.8688716

Orrù, P. F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R. & Arena, S. 2020. Machine Learning Approach Using Mlp And Svm Algorithms For The Fault Prediction Of A Centrifugal Pump In The Oil And Gas Industry. *Sustainability*, 12**,** 4776. https://doi.org/10.3390/su12114776

Panday, R., Indrawan, N., Shadle, L. J. & Vesel, R. W. 2021. Leak Detection In A Subcritical Boiler. *Applied Thermal Engineering*, 185**,** 116371. https://doi.org/10.1016/j.applthermaleng.2020.116371

Que, Z. & Xu, Z. 2019. A Data-Driven Health Prognostics Approach For Steam Turbines Based On Xgboost And Dtw. *IEEE Access*, 7**,** 93131-93138. https://doi.org/10.1109/ACCESS.2019.2927488

Sarkar, D. 2015. *Thermal Power Plant: Design And Operation*, Elsevier. https://doi.org/10.1016/C2014-0-00536-9

Sbarufatti, C., Corbetta, M., Manes, A. & Giglio, M. 2016. Sequential Monte-Carlo Sampling Based On A Committee Of Artificial Neural Networks For Posterior State Estimation And Residual Lifetime Prediction. *International Journal Of Fatigue*, 83**,** 10-23. https://doi.org/10.1016/j.ijfatigue.2015.05.017

Shohet, R., Kandil, M. S. & Mcarthur, J. Machine Learning Algorithms For Classification Of Boiler Faults Using A Simulated Dataset. Iop Conference Series: Materials Science And Engineering, 2019. Iop Publishing, 062007. https://dx.doi.org/10.1088/1757-899X/609/6/062007

Shohet, R., Kandil, M. S., Wang, Y. & Mcarthur, J. J. 2020. Fault Detection For Non-Condensing Boilers Using Simulated Building Automation System Sensor Data. *Advanced Engineering Informatics*, 46**,** 101176. https://doi.org/10.1016/j.aei.2020.101176

Sohaib, M. & Kim, J.-M. 2019. Data Driven Leakage Detection And Classification Of A Boiler Tube. *Applied Sciences* [Online], 9. https://doi.org/10.3390/app9122450

Sola, J. & Sevilla, J. 1997. Importance Of Input Data Normalization For The Application Of Neural Networks To Complex Industrial Problems. *IEEE Transactions On Nuclear Science*, 44**,** 1464-1468. https://doi.org/10.1109/23.589532

Świercz, M. & Mroczkowska, H. 2019. Application Of Pca For Early Leak Detection In A Pipeline System Of A Steam Boiler. *Przegląd Elektrotechniczny*. https://doi.org/10.15199/48.2019.10.43

Tao, L., Lu, C. & Noktehdan, A. 2015. Similarity Recognition Of Online Data Curves Based On Dynamic Spatial Time Warping For The Estimation Of Lithium-Ion Battery Capacity. *Journal Of Power Sources*, 293**,** 751-759. https://doi.org/10.1016/j.jpowsour.2015.05.120

Tin Kam, H. Random Decision Forests. Proceedings Of 3rd International Conference On Document Analysis And Recognition, 14-16 Aug. 1995 1995. 278-282 Vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Vapnik, V. 2013. *The Nature Of Statistical Learning Theory*, Springer Science & Business Media. https://doi.org/10.1007/978-1-4757-3264-1

Wang, Y., Yin, X. & Wang, B. A Method Of Diagnosing Leakage Of Boiler Steam And Water Pipes Based On Genetic Neural Network. 2016 12th World Congress On Intelligent Control And Automation (Wcica), 2016. IEEE, 2829-2833. https://doi.org/10.1109/WCICA.2016.7578778

Xu, J., Zhipeng, C., Suxia, M., Xiaowei, W., Zhiyao, Z., & Guoxia, Z.. 2024. Data based digital twin for operational performance optimization in CFB boilers, *Energy* 306, 132532. https://doi.org/10.1016/j.energy.2024.132532

BIOGRAPHIES

Achmad Widodo received his Bachelor of Engineering and Master of Engineering degrees in Mechanical Engineering from Universitas Diponegoro and the Bandung Institute of Technology, Indonesia, in 1998 and 2002, respectively. He received his Doctorate in Mechanical Engineering from Pukyong National University, South Korea, in 2007. He is currently a lecturer in the Department of Mechanical Engineering, Universitas Diponegoro, Indonesia. His research interests include machine condition monitoring, and applied machine learning for fault diagnostics and prognostics.

Toni Prahasto is a lecturer in Mechanical Engineering of Universitas Diponegoro, Semarang, Indonesia. He holds a Doctor of Philosophy and Master of Applied Science in Mechanical Engineering from the University of Waterloo, Canada in 1998 and 1992 respectively and a Bachelor of Mechanical Engineering from the Bandung Institute of Technology, Indonesia in 1985. His research interests are in the fields of optimization of maintenance scheduling and strategy, forecasting and anomaly detection of time series data, and multi-objective optimization for operation and maintenance.

Mochamad Soleh is currently a General Manager of PT PLN (Persero) Research Institute Indonesia. He received Bachelor and Master degrees from Department of Mechanical Engineering Universitas Gadjah Mada and Bandung Institute of Technology Indonesia in 2002 and 2012, respectively. His research interest is technology of biomass coal co-firing for power plant.

Herry Nugraha received the B.S. degree in mechanical engineering and the M.S. degree in electrical engineering from the Bandung Institute of Technology, Bandung, Indonesia, in 1993 and 2010, respectively. He received the Ph.D. degree in electrical engineering from Bandung Institute of Technology, Bandung, Indonesia, in 2016. He is currently a lecturer at Faculty of Technology and Energy Business, Institut Teknologi PLN, Indonesia. His research interests include maintenance and operation engineering of power plants, economics of electric energy systems, statistics and optimization theory and its applications, and risk analysis.