Regression methods for detecting anomalies in flue gas desulphurization installations in coal-fired power plants based on sensor data⋆

Marek Moleda[1,2][0000−0002−6556−5083], Alina Momot[2][0000−0002−9308−8077], and Dariusz Mrozek[2][0000−0001−6764−6656]

[1] TAURON Wytwarzanie S.A. `marek.moleda@gmail.com`
[2] Silesian University of Technology
ul. Akademicka 16, 44-100 Gliwice, Poland
`dariusz.mrozek@polsl.pl`

Abstract. In the industrial world, the Internet of Things produces an enormous amount of data that we can use as a source for machine learning algorithms to optimize the production process. One area of application of this kind of advanced analytics is Predictive Maintenance, which involves early detection of faults based on existing metering. In this paper, we present the concept of a portable solution for a real-time condition monitoring system allowing for early detection of failures based on sensor data retrieved from SCADA systems. Although the data processed in systems, such as SCADA, are not initially intended for purposes other than controlling the production process, new technologies on the edge of big data and IoT remove these limitations and provide new possibilities of using advanced analytics. This paper shows how regression-based techniques can be adapted to fault detection based on actual process data from the oxygenating compressors in the flue gas desulphurization installation in a coal-fired power plant.

Keywords: Predictive maintenance · Power Plant · SCADA · Anomaly detection

1 Introduction

Operational Technology (OT) systems refer to hardware and software solutions that are capable of changing various industrial processes through the direct monitoring and controlling of physical devices, procedures, routes, and events in a

factory. Internet of Things (IoT) and integration of IT and OT systems provide new opportunities for intelligent use of advanced analytical solutions to increase the reliability of various production processes, including those performed in power plants and the entire energy sector [8, 14]. A significant risk factor in the energy sector is unplanned downtime, which causes production breaks and, thus, loss of revenue. Increasing the reliability of equipment operation is a critical factor in reducing the losses mentioned above, and one of the methods to achieve this goal is the use of predictive maintenance. Predictive maintenance reduces planned and unplanned downtime by early detection of faults and better planning of repairing actions. In this work, we present how the regression-based techniques can be used for real-time condition monitoring on the basis of sensor data retrieved from SCADA (supervisory control and data acquisition) systems [5] and failure logs. The research objective of this work is to adjust the data mining process to achieve satisfactory results with a minimum effort from analysts and system engineers. In order to prove the portability of the presented solution, we use the algorithm previously applied for the feed pumps [13] for the analysis of data coming from oxygenating compressors. We optimize the steps of the algorithm to achieve satisfactory prediction results. Experiments are carried out on real data from sensors monitoring the oxygenating compressor, which is a component of the flue gas desulphurization installation in a coal-fired power plant. Oxygenating compressors are indispensable in the process of producing gypsum from sulfur dioxide as a side effect of coal combustion. We present a model-based anomaly detection approach with the potential to be easily applied to other devices. We also show that general-purpose regression-based heuristic algorithms can enhance the benefits of the processing of data from OT repository, various data streams, and related big data sets, including various logs.

2   Background and Related works

Predictive maintenance is applied in various branches of industry, especially in those of them, in which failures may cause considerable economic consequences.

2.1   Predictive maintenance

Predictive maintenance is a concept that involves planning maintenance work based on the health of the equipment. This approach is possible when we have complete data on the operation of the device and expert knowledge on how to use it for analysis [1, 4, 22]. The task consists of the analysis of archival events, diagnostic results, and device statistics by the production engineer or production data analyst. OT systems, including SCADA repositories, can be used as sources of data for predictive maintenance. Then, we can codify the expert knowledge in the form of rules and classifiers to get essential indicators, monitor devices, and discover anomalies, as it is presented in [7, 18, 16, 17]. In this way, we can create a predictive maintenance system based on real-time condition monitoring. Predictive maintenance is a reasonable compromise between *preventive maintenance*
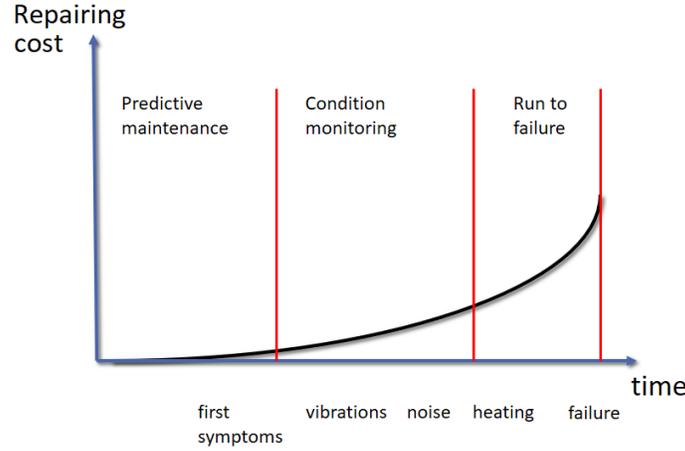
Fig. 1. Repairing costs depending on the time of detection of the fault.

[12, 3] and *run to failure* maintenance. The primary purpose of the methods used in predictive maintenance is the early detection of faults [15, 26, 23] based on the first symptoms to minimize the repairing costs (as shown in Fig. 1). Other potential benefits are:

- Changing the operational models to *Fix as required*, in which the service is done before failure is expected.
- Minimizing the probability of unplanned downtime by real-time health monitoring.
- Reducing unnecessary repairs of equipment being in good condition.

2.2   Model-based prediction

The concept of model-based prediction is to create a black-box model [23] that represents a specific process, as it is shown in Fig. 2.
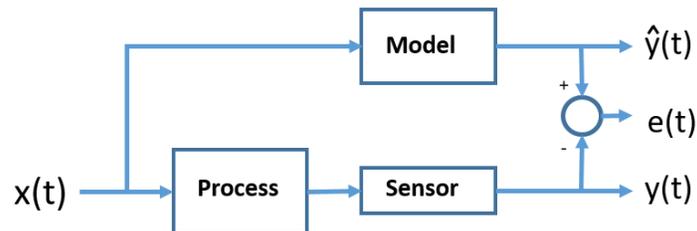

Fig. 2. Model-based prediction system.

The model created for the component (in this case it is a measurement) is designed to predict the indication of sensor $y(t)$ based on known input $x(t)$. The difference between the actual value $y(t)$ and the predicted $\hat{y}(t)$ is a measure of the process deviation from the model value. The deviation from the expected value is assumed to be a symptom of a changing health condition of the device, i.e., an imminent failure or an improvement in performance after repair. The essence of the proposed method is to create a model based on data from the failure-free period and to monitor the $e(t)$ deviation value.

2.3   Modeling based on regression

Machine learning (ML) is one of the techniques frequently applied in predictive maintenance performed in the industrial environment. However, the spectrum of used ML models varies and is specific for the application area. Machine learning-based algorithms generally can be divided into two main classes:

– supervised - where information on the occurrence of failures is present in the training data set;
– unsupervised - where process information is available, but no maintenance related data exists.

Supervised approaches require the availability of a data set $S = \{x_i, y_i\}_{i=1}^n$, where a couple $\{x_i, y_i\}$ contains the information related to the i-th process iteration. Vector $x_i \in \mathbb{R}^{1 \times p}$ contains information related to the p variables associated with available process information [21]. Depending on the type of y we distinguish: classification models (if categorical labels are predicted) and regression models (if the results are continuous values)

Supervised learning is successfully used in the area of predictive maintenance to classify faults by building fault detectors. In the literature, these detectors rely on various Artificial Intelligence (AI) techniques, such as Artificial Neural Networks [2, 6], k-Nearest Neighbours [24], Support Vector Machines [10, 20], or Bayesian networks [19, 25], frequently using some methods for reducing dimensionality of data, such as Principle Component Analysis [11, 27].

2.4   The coding of electrical components

To obtain a uniform system of marking devices and installations in the power plant, the KKS coding system (Kraftwerk Kennzeichensystem) is used. The KKS marking system had been used since the early 1980s by power plant constructors and power plant operators to name and identify all components of a power plant. The code structure is shown in Fig. 3 with the example of a temperature meter.

3   Regression-based Anomaly Detection

Our failure prediction model applies regression algorithms in predictive maintenance tasks. For the set of all input data, where $X = [x_1, x_2, x_3, \ldots, x_m]^T$ is
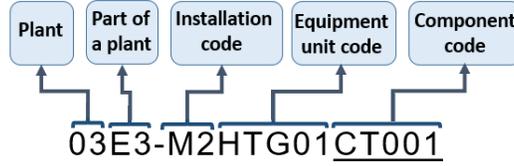
Fig. 3. Coding of temperature meter using KKS notation

the vector of individual measurements from m sensors, we estimate our response variable as a polynomial function of other variables, i.e.:

$$\hat{y}_i = \sum_{d=1}^{d_{max}} \sum_{\substack{j=1 \\ j \neq i}}^{m} a_{jd}^{(i)} x_j^d + a_0^{(i)} + \varepsilon^{(i)} \qquad \forall i \in \{1, 2, \ldots, k\}, k \leq m, \tag{1}$$

where $\varepsilon^{(i)}$ is the i-th independent identically distributed normal error and coefficients $a_j^{(i)}$ are calculated using the method of least squares [9]. In such a way, we build a bag of predictive models for the investigated (monitored) variables (signals) and predict the values of the variables based on other signals that are correlated with the investigated one.

The difference between the observed and the predicted values is not an absolute measure that can be used when comparing the values with other signals. To normalize the results, we introduced the NRE (normalized relative error) coefficient [13] (which is a multiple of the mean standard deviation) to measure the degree of deviation for the i-th variable in a data set:

$$NRE_i = \frac{|y_i - \hat{y}_i| - MAE_i}{RMSE_i}, \tag{2}$$

where y_i represents the observed values, $\hat{y}_i$ represents the predicted values, $RMSE_i$ is the root mean square error, and MAE_i is the mean absolute error.

By selecting a variable with the maximum value of the NRE_{max} (called *maximum normalized relative error*) we can identify the signal which is probably the cause of the upcoming fault:

$$NRE_{max} = \max(NRE_1, NRE_2, .., NRE_k). \tag{3}$$

This makes it possible to diagnose a source of the anomaly quickly.

Withe the use of regression algorithms and the NRE metrics, we create a bag of models for detecting anomalies. In our case, we focus on creating a reproducible process that is easy to use to monitor various equipment in the power plant. In order to achieve this goal, we automate data processing tasks and reduce the necessary analytical efforts.

3.1   Description of input data and correlation between signals

The data used in the experiment comes from monitoring two oxygenating compressors from flue gas desulphurization installations. Each of the devices is de-

scribed by ten parameters, such as power, bearing temperature, vibrations, oil pressure. These parameters are measurements collected by sensors and are stored in the SCADA system. Available process data covers the period from January 2017 to May 2019, in which the compressors were in operation for about one-third of that time. Data was obtained off-line for two compressors labeled as 1HTG01 and 2HTG01. Analyzing Pearson's correlation coefficient for individual signals, we can distinguish groups of strongly correlated signals. The existence of strong correlations (i.e., values close to 1 or -1) means significant relationships between different variables and justifies the use of the regression model for anomaly prediction. A graphical matrix of correlations for a period of six months is shown in Fig. 4.
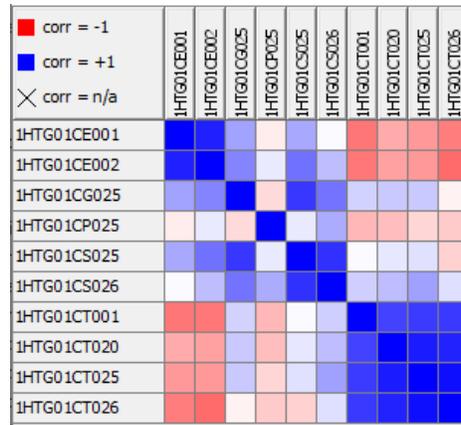

Fig. 4. Correlation matrix for the training set.

3.2   Data preprocesing and integration

Data collection Data preprocessing covers the selection of a group of measurements that will be part of the input set. In the case of a single prediction task, we can easily determine the group of measurement units related to the device (e.g., using technical documentation). However, when we want to use the prediction model to cover the whole unit containing tens of thousands of measurement points, the work is no longer trivial. The coding method described in section 2.4 is helpful in this case. The notation method allows for easy separation of all subcomponents associated with the master equipment, as it is shown in Fig. 5. In this way, we can significantly reduce the set of potential input signals for further analysis (e.g., correlation analysis, constant value filtering, etc.).

Automatic data labelling To learn the model of how the proper signal looks like, we have to provide the data describing the state when the investigated device
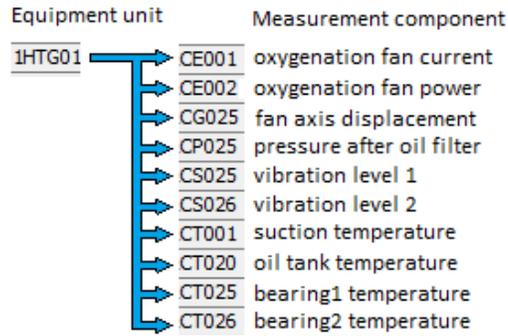
Fig. 5. Grouping measurement units for the equipment with the use of KKS code.

is healthy. To filter out the periods in which failures occurred from the input data set and train the prediction model on correct data, it is necessary to label the records with data from a service log. Information about the occurrence of faults is needed to extract a training set for the training process and to further evaluate and visualize the results. Analysis of the impact of faults is time-consuming and requires engineering expertise to categorize events accordingly. Therefore, the labeling process was simplified to automatically mark the data set based on fault registration dates and repairing service completion. A column containing true/false data is added to the training set to indicate whether the device was in a fault condition at that time, as shown in Fig. 6. The input data is filtered based on the priority to isolate less significant events such as service works, including oil change or periodic inspections. Automation of the labeling process requires a standardized way of recording service works and good quality of data. However, possible errors can be compensated for by the size of the training set, and their impact on the quality of the algorithm can be marginalized.

3.3   Model optimization

Optimisation step is employed in our process to adjust the parameters of the model to give the best results, and at the same time to avoid overfitting situation. The parameters we optimize are the polynomial degree and a set of input features. The optimization was verified in the k-fold cross-validation process, where we divided the input set into five equal parts. By minimizing the mean squared error value while optimizing the polynomial degree on a 3-month test set, we obtained the results shown in Table 3.3.

Polynomial fitting For most variables, the best prediction results were obtained for a low polynomial degree. The higher the polynomial degree, the more sensitive the changes in the input set are. On the other hand, for the higher polynomial degrees, the prediction model exposes more abnormal situations, but it operates very unstable and is susceptible to noise.
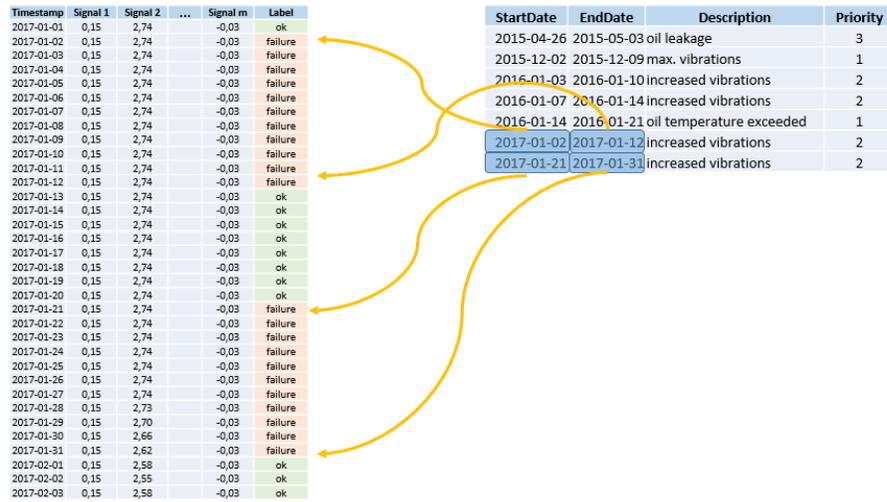
Fig. 6. Method for automatic labeling.

**Table 1.** Mean square error (MSE) depending on the degree of polynomial.

Signal name	polynomial degree				
	1	2	3	4	5
1HTG01CP025	0,0001	0,0002	0,0002	0,0002	0,0011
1HTG01CS025	2,7962	1,2322	1,8621	29,408	264,76
1HTG01CS026	2,6363	2,1789	2,9565	16,536	34,689
1HTG01CG025	0,00001	0,00001	0,00002	0,0001	0,0002
1HTG01CT025	0,0087	0,0111	0,0196	0,0286	0,1813
1HTG01CT026	0,0067	0,008	0,0173	0,0106	0,0281
1HTG01CT001	12,433	12,643	14,953	19,346	30,349
1HTG01CT020	0,0818	0,0745	0,0791	0,1373	1,1435
1HTG01CE001	0,1006	0,1084	0,1286	0,1659	3,1217
1HTG01CE002	8,0852	8,4233	8,0673	30,904	154,71

Feature selection In the next optimization step, we select the input features for the polynomial degree calculated earlier for all k signal models in a bag. By choosing the right parameters it is possible to increase the accuracy of the model by about 20% (i.e., reduce the mean squared error by 20% on average) as it is shown in Table 3.3. For this purpose, we tested two algorithms - Backward elimination and Genetic algorithm. The use of both methods gave the same results in terms of the number of features, the mean square error, and improvement in prediction accuracy. The time of the Backward elimination (29,785 combinations) compared to the Genetic algorithm (27,535 combinations) is also similar.

Table 2. Results obtained by optimizing the number of analyzed features.

Signal name	All features		Backward elimination		Genetic algorithm		Improvement
	No. feat.	MSE	No. feat.	MSE	No. feat.	MSE	%
1HTG01CP025	9	0,0001	2	0,0001	2	0,0001	9,76%
1HTG01CS025	9	1,2322	7	1,0971	7	1,0971	12,3%
1HTG01CS026	9	2,1789	5	1,7033	5	1,7033	27,92%
1HTG01CG025	9	0,00001	3	0,00001	3	0,00001	57,77%
1HTG01CT025	9	0,0087	7	0,0073	7	0,0073	19,15%
1HTG01CT026	9	0,0067	6	0,0053	6	0,0053	26,18%
1HTG01CT001	9	12,433	4	10,352	4	10,352	20,09%
1HTG01CT020	9	0,0745	3	0,0499	3	0,0499	49%
1HTG01CE001	9	0,1006	7	0,099	7	0,099	1,58%
1HTG01CE002	9	8,067	3	4,8747	3	4,8747	65,49%

Time window length When designing a predictive system, an essential parameter of the built predictive model is a time window, which provides a training set for the model. The time window may also influence the predictive capabilities of the model. Therefore, we also investigated this parameter in our research. Visual differences in the results depending on the length of the time window are shown in Fig. 7. Longer time windows covering wide data range (3 months in our experiments) adapt more slowly to sudden changes such as renovation (in January 2019 in Fig. 7), but clearly shows a disparity between failure periods and proper operation. In the case of short time windows (1 month), the model adapts quickly to the variability of the environment, i.e., it can correct itself soon after a renovation. On the other hand, in the case of a long-term malfunction, it can treat the fault as the correct state.
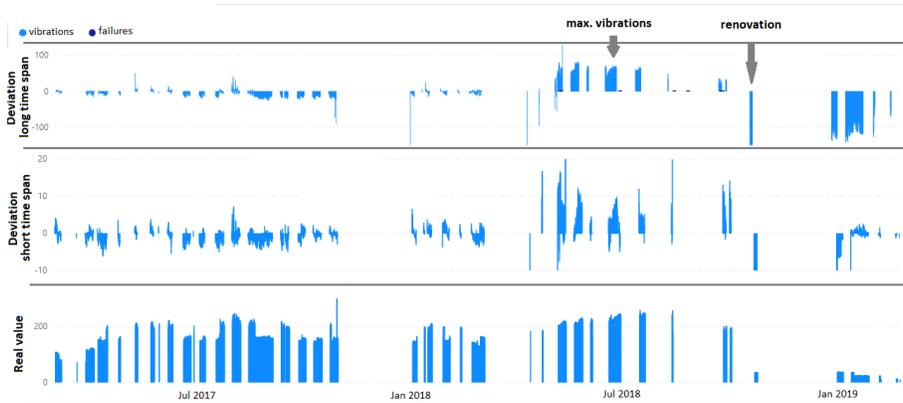
Fig. 7. Deviations from normal state obtained for long- and short time window.

4   Testing Prediction Capabilities

4.1   Evaluation function and error distribution

While detecting anomalies, one of the quality factors assessing the performance of the built predictive model is to obtain a high error ratio for the failure period comparing to the normal operation period. This is achieved in our model. For example, the error distribution (deviation from the actual value) for the 2GHG01 compressor is shown in Fig. 8. The red color is used to indicate normal operation state, and the blue color is used to indicate emergency conditions (the stated period was extended by an earlier two weeks to assess the ability to detect pre-emergency conditions). While the red histogram is close to a normal distribution with an average of around 0, the blue histogram is characterized by a significant diagonality and a shift in the average value.
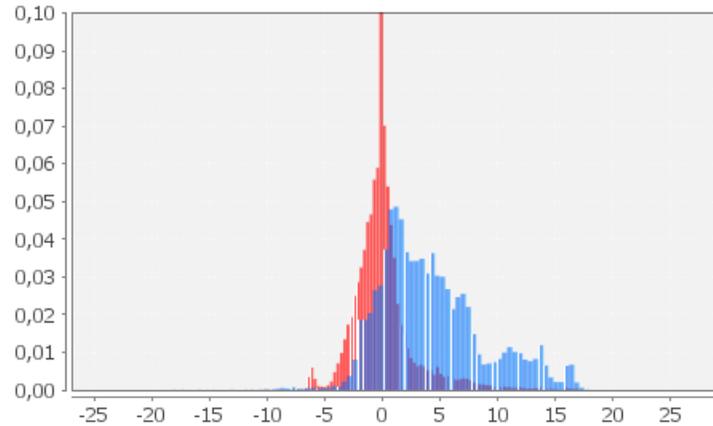

Fig. 8. Distribution of deviations for the fault and normal states.

4.2   Detection capability of the model

Short term prediction   One of the failures, which was recorded on 23 February for 1GTG01 compressor, was bearing damage. The damage occurred at a time when increased vibration levels were recorded, resulting in a significant overhaul. However, before the failure itself, we can notice a significant increase in the deviation of the estimated values. About an hour before the device was switched off on 20 February, a considerable increase in the deviation is visible (more than 100% of the average value in normal operation). The physical effect of the failure observed had a reflection on the actual pressure drop behind the oil filter. The fault occurred three days later on the compressor startup, as it is visible in Fig. 9.
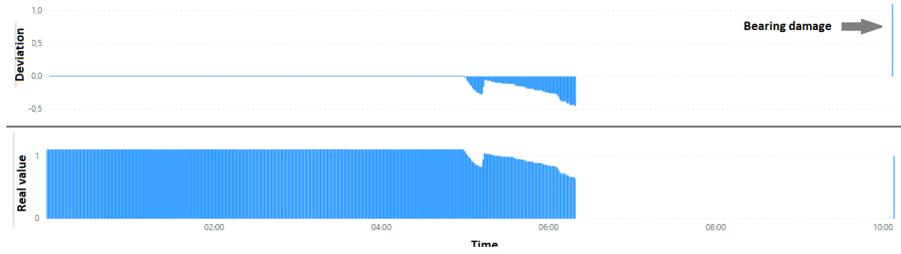
Fig. 9. Deviation and actual oil pressure behind the filter before bearing failure.

Long term prediction During the period considered, several significant events were recorded in the fault logs. For the compressor 1HTG01, the bearing was damaged in February 2017, followed by reports of loud operation and bearing vibrations, as a result of which the device was overhauled in April 2017. After the renovation, the prediction model did not show any deviations despite the reports of increased vibrations around July-August. However, the analysis of actual vibration measurements did not indicate any abnormal operation. The compressor 2HTG01 was characterized by failure-free operation for an extended period until May 2018, when increased oil temperature was recorded, and in the following weeks, several vibration defects appeared. The compressor was overhauled in October 2018, the impact of the overhaul is visible in the chart in Fig. 10 as a significant decrease in the error (to a considerable negative value). For both compressors, the failure points visible in Fig. 10 as registered events coincide with the deviations indicated by the predictive model for the investigated variable, which shows that the predictive model works well.
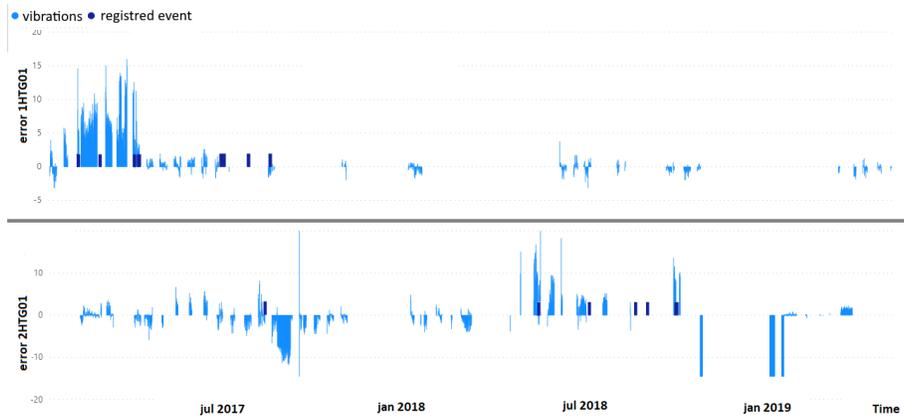

Fig. 10. Vibration deviation with the indication of faults and service actions.

4.3 Comparison of model results

The results obtained (the values of NRE) were calculated by dividing the deviation value by the local root mean square error (equivalent to standard deviation) as in eq. 2, and then the values for which the standard deviation exceeded the threshold value 6 (determined experimentally) for less than 15 minutes were filtered out. That means that the alarm for the predicted incoming fault is raised only when the $NRE > 6$ for at least 15 minutes. The same threshold was used for the condition monitoring of boiler feed pumps described in [13], bringing good quality prediction results. In Table 3, we show the effectiveness of the
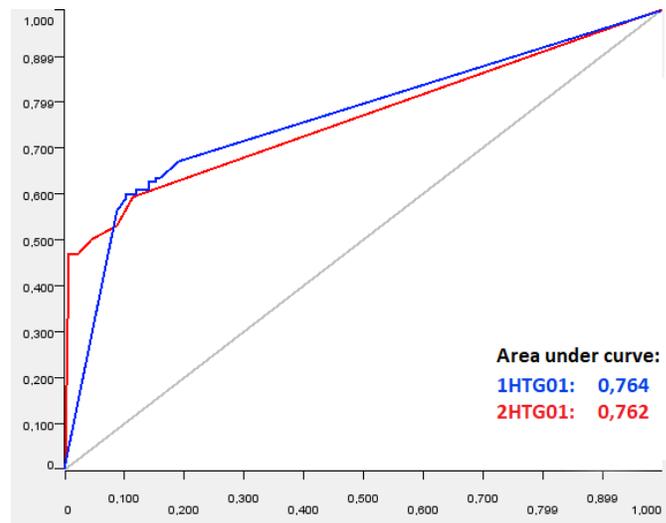

Fig. 11. ROC curves for anomaly detection predictive models built for compressors 1HTG01 and 2HTG02.

predictive models built for compressors 1HTG01 and 2HTG02, and for boiler feed pumps investigated in our previous work [13]. The results confirm that the presented approach to anomaly detection through regression-based predictive analytics achieved a good level of specificity, which translates into a small number of false positives (type I errors). Sensitivity can be improved by developing an evaluation method or improving optimization steps.

5 Discussion and Conclusions

The obtained results confirm good portability of the presented approach, which after some adjustments, can be used for detecting anomalies and early fault prediction for various types of equipment located in power plants. The regression-based algorithm, previously used to predict faults in the boiler feed pumps, was

Table 3. Performance of the predictive models built for various devices in power plant.

	Accuracy	AUC	Sensitivity	Specificity
Boiler feed pump	0,86	0,89	0,67	0,95
Compressor 1HTG01	0,79	0,764	0,58	0,91
Compressor 2HTG01	0,93	0,762	0,47	0,98

successfully applied to the oxygenating compressors in the flue gas desulphurization installation. The results obtained for different devices are comparable. The fact that they were achieved using simple data mining methods and little effort encourage further works on improving model quality. Comparing to our previous work, we proposed methods of automation of feature selection, data collection, and grouping based on specific coding of devices and installations in the power industry. The unquestionable advantage of the algorithm is the absence of data labeling compared to most classification-based methods mentioned in the related literature. Data labeling in the current work is not a necessary element and was done only to assess the quality of the model.

References

1. Afefy, I.H.: Reliability-centered maintenance methodology and application: a case study. Engineering **2**(11), 863 (2010)
2. Awadallah, M.A., Morcos, M.M.: Application of ai tools in fault diagnosis of electrical machines and drives-an overview. IEEE Transactions on energy conversion **18**(2), 245–251 (2003)
3. Bachus, L., Custodio, A.: Know and understand centrifugal pumps. Elsevier Science (2003)
4. Bevilacqua, M., Braglia, M.: The analytic hierarchy process applied to maintenance strategy selection. Reliability Engineering & System Safety **70**(1), 71–83 (2000)
5. Boyer, S.A.: SCADA: supervisory control and data acquisition. International Society of Automation (2009)
6. Cococcioni, M., Lazzerini, B., Volpi, S.L.: Robust diagnosis of rolling element bearings based on classification techniques. IEEE Transactions on Industrial Informatics **9**(4), 2256–2263 (2012)
7. Feng, Y., Qiu, Y., Crabtree, C.J., Long, H., Tavner, P.J.: Monitoring wind turbine gearboxes. Wind Energy **16**(5), 728–740 (2013)
8. Fortino, G., Savaglio, C., Zhou, M.: Toward opportunistic services for the industrial internet of things. In: 2017 13th IEEE Conference on Automation Science and Engineering (CASE). pp. 825–830. IEEE (2017)
9. Freedman, D.A.: Statistical Models: Theory and Practice. Revised Edition. Cambridge University Press, New York (2009)
10. Konar, P., Chattopadhyay, P.: Bearing fault detection of induction motor using wavelet and support vector machines (svms). Applied Soft Computing **11**(6), 4203–4211 (2011)
11. Malhi, A., Gao, R.X.: Pca-based feature selection scheme for machine defect classification. IEEE Transactions on Instrumentation and Measurement **53**(6), 1517–1525 (2004)

12. Marscher, W.D., et al.: Avoiding failures in centrifugal pumps. In: Proceedings of the 19th International Pump Users Symposium. Texas A&M University. Turbomachinery Laboratories (2002)
13. Moleda, M., Momot, A., Mrozek, D.: Predictive maintenance of boiler feed water pumps using scada data. Sensors **20**(2), 571 (2020)
14. Moleda, M., Mrozek, D.: Big data in power generation. In: International Conference: Beyond Databases, Architectures and Structures. pp. 15–29. Springer (2019)
15. Nouri, M., Fussell, B.K., Ziniti, B.L., Linder, E.: Real-time tool wear monitoring in milling using a cutting condition independent method. International Journal of Machine Tools and Manufacture **89**, 1–13 (2015)
16. Qiu, Y., Chen, L., Feng, Y., Xu, Y.: An approach of quantifying gear fatigue life for wind turbine gearboxes using supervisory control and data acquisition data. Energies **10**(8), 1084 (2017)
17. Qiu, Y., Feng, Y., Sun, J., Zhang, W., Infield, D.: Applying thermophysics for wind turbine drivetrain fault diagnosis using scada data. IET Renewable Power Generation **10**(5), 661–668 (2016)
18. Qiu, Y., Feng, Y., Tavner, P., Richardson, P., Erdos, G., Chen, B.: Wind turbine scada alarm analysis for improving reliability. Wind Energy **15**(8), 951–966 (2012)
19. Qu, J., Zhang, Z., Gong, T.: A novel intelligent method for mechanical fault diagnosis based on dual-tree complex wavelet packet transform and multiple classifier fusion. Neurocomputing **171**, 837–853 (2016)
20. Rojas, A., Nandi, A.K.: Detection and classification of rolling-element bearing faults using support vector machines. In: 2005 IEEE Workshop on Machine Learning for Signal Processing. pp. 153–158. IEEE (2005)
21. Susto, G.A., Schirru, A., Pampuri, S., McLoone, S., Beghi, A.: Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics **11**(3), 812–820 (2014)
22. Swanson, L.: Linking maintenance strategies to performance. International journal of production economics **70**(3), 237–244 (2001)
23. Tautz-Weinert, J., Watson, S.J.: Using scada data for wind turbine condition monitoring–a review. IET Renewable Power Generation **11**(4), 382–394 (2016)
24. Tian, J., Morillo, C., Azarian, M.H., Pecht, M.: Motor bearing fault detection using spectral kurtosis-based feature extraction coupled with k-nearest neighbor distance analysis. IEEE Transactions on Industrial Electronics **63**(3), 1793–1803 (2015)
25. Wu, J., Wu, C., Cao, S., Or, S.W., Deng, C., Shao, X.: Degradation data-driven time-to-failure prognostics approach for rolling element bearings in electrical machines. IEEE Transactions on Industrial Electronics **66**(1), 529–539 (2018)
26. Yang, C., Liu, J., Zeng, Y., Xie, G.: Real-time condition monitoring and fault detection of components based on machine-learning reconstruction model. Renewable energy **133**, 433–441 (2019)
27. Zhang, X., Xu, R., Kwan, C., Liang, S.Y., Xie, Q., Haynes, L.: An integrated approach to bearing fault diagnostics and prognostics. In: Proceedings of the 2005, American Control Conference, 2005. pp. 2750–2755. IEEE (2005)