Condition monitoring of rolling element bearings: benchmarking of data-driven methods

E Iunusova¹, K Szipka¹ and A Archenti²
¹Sustainable Production Development Department, KTH Royal Institute of Technology Kvarnbergagatan 12, 15136 Södertälje, Sweden
²Production Engineering Department, KTK Royal Institute of Technology
Brinellvägen 68, 100 44 Stockholm, Sweden
E-mail: iunusova@kth.se

Abstract. Condition-based maintenance (CBM) is a maintenance strategy used to gain updated information about equipment condition and is today considered a natural part of the engineering field. The replacement of the traditional scheduled maintenance strategy in favor of CBM has the potential to significantly improve the safety of the system operating in harsh environments of the operation and increase in productivity by prolonging the life of an asset and preventing costly breakdowns. For many years CBM remained the subject of vigorous research and discussions. Increasing the automation level and the number of sensors in industries allowed obtaining and collecting data in large amounts. The current level of computational power allows us to process and analyse this massive amount of data, which has given a new leap in the development of industrial analytics. Rather than in the case of classical knowledge-based modelling tools, data-driven methods propose modelling and forecasting frameworks based on data analysis. Consequently, the transition to the data-driven modelling gave a leap in CBM research and has recently drawn increasing attention, providing new case studies, algorithms, and results. However, technical challenges remain. Despite great flexibility and good forecasting performances, there are several limitations of data-driven algorithms. This paper provides an overview of the data-driven failure algorithms for rolling element bearings monitoring. Bearings have played a pivotal role in industrial machinery to operate with high efficiency and safety. They are considered to be one of the most common machine elements of precision rotating machinery. A benchmarking of various predictive and descriptive algorithms was performed. The analysis was carried out on a dataset from the run-to-failure experiments on bearings from NASAs Data Repository. This paper also summarizes the current trends and highlights the limitations with respect to traditional knowledge-based modelling. Special attention is paid to identifying research gaps and promising research directions.

1. Introduction
The Fourth Industrial Revolution is strongly associated with the integration between physical and digital systems in production environments. The proliferation of sensing technologies exponentially increases the amount of data extracted from the manufacturing process, production system, and equipment. This data can bring out valuable information and knowledge that could enhance the effectiveness of the whole manufacturing process, as a valuable asset for decision support in various areas, particularly in health monitoring and condition-based maintenance. Driving by the aspiration to
increase equipment uptime and reduce unnecessary maintenance work, more and more attention has been put into predictive maintenance, which necessitates advanced prognosis tools to focus on performance degradation monitoring and assessment so failure can be predicted and prevented [1].

The various methods and tools regarding failure prognostic can be grouped into the knowledge-based and data-driven approaches. Knowledge-based approaches rely on the use of a set of algebraic and differential equations obtained by using traditional laws of physics. Data-driven approaches aim to transform raw data into relevant information and behavior models by using only the data provided by the monitoring system. In this case, the target is not to establish an analytical model of the system (physical and mathematical knowledge of the process) neither to provide a deep understanding of its physical parameters [2].

The design of knowledge-based process monitoring and fault diagnosis systems has been a remarkable research topic for several decades. The well-established knowledge-based techniques have been successfully applied to plenty of processes for industrial electronics, automatic control systems, etc. Recently, to ensure the reliability and safety of modern large-scale industrial processes, data-driven methods have been receiving considerably increasing attention, particularly for the purpose of process monitoring [3]. The malfunction of manufacturing machines can be generally attributed to various faults of different categories, including drive inverter failures, stator winding insulation breakdown, bearing faults, and air gap eccentricity. Several surveys regarding the likelihood of induction machine failures conducted by the IEEE Industry Application Society and the Japan Electrical Manufacturers’ Association reveal that bearing fault is one of the foremost causes of breakdowns and is responsible for 30 to 40 percent of all machine failures [4]. Such failure can be catastrophic, resulting in costly downtime.

Bearings have played a pivotal role in industrial machinery to operate with high efficiency and safety, and they are considered to be one of the most common machine elements of precision rotating machinery. Furthermore, bearings are critical to almost all forms of rotating machinery and are among the most common machine elements. Therefore, rolling element bearing condition monitoring and diagnostics has been a research frontier for engineers and scientists for many years. In order to prevent unexpected bearing failure, vibration analysis has been used extensively for various bearing condition monitoring techniques [5]. A rolling-element bearing is illustrated in figure 1a, which contains the outer race typically mounted on the motor cap, the inner race to hold the motor shaft, the balls or the rolling elements, and the cage for restraining the relative distances between adjacent rolling elements [4].

This paper provides an overview of the most widely used data-driven failure algorithms for rolling element bearing monitoring. The results show that the improvement of existing Machine Learning (ML) algorithms and the creation of new ones are gathering more significant momentum from year to year. Modern algorithms allow reaching good prediction results even for those who don’t understand the fundamentals of the investigation subject. Complete disregard of the physical foundations in pursuit of a more accurate forecast leads to the creation of untraceable models that are adapted to the particular case with particular data from particular equipment. They do not rely on understanding physical and mathematical laws, and in many cases, trained models do not generalize well.

This paper aims to assess the current state and the development direction of failure prognostics approaches, investigate modern data-driven trends, and highlight their limitations. The remaining sections of this paper are organized as follows: In Section 2, the description of the run-to-failure test dataset and fault types are provided. In Section 3, the workflow of the used methodology is shown and briefly described. The comparison of the results of implementing the proposed methodology is presented in Section 4. Finally, Section 5 presents a summary of the work with the highlighting limitations and future investigation plans.
2. Overview

Data is the foundation for data-driven modelling. It is necessary to have a good collection of datasets to be able to develop effective ML algorithms for bearing fault detection. Natural bearing degradation is a gradual process and may develop throughout months or years. Bearing faults are investigated through three major approaches for monitoring purposes. First, with the application of simulations after the detailed modelling of bearing behavior, including the simulation of the sensor signals. Second, artificially induced faults, for instance by scratching or drilling the bearing surface or created by exerting a shaft current for accelerated life testing. Third, bearings are run under given conditions with performing run-to-failure tests. While the data collection (especially in the third case) can be expensive and time-consuming, fortunately, a few organizations have made an effort and published bearing fault datasets for investigations of various ML algorithms. Their widespread prevalence has another great advantage. Researchers from all over the world could compare the results of implementing different algorithms to the same dataset. The data contains a complete record of natural bearing defect evolution generated by the NSF I/UCR Center for Intelligent Maintenance Systems with support from Rexnord Corp. in Milwaukee was used in this paper [6].

Figure 1. The test rig: a) the main components of the rolling element bearing [7], b) the photo of the test rig [6], c) the test rig setup scheme [6].
2.1. Test Rig Setup
In the test rig, four Rexnord ZA-2115 double row bearings were installed on the shaft (figure 2). The test rig element’s characteristics are presented in table 1 [6].

Table 1. The test rig element’s characteristics.

| Element            | Characteristics                                                                 |
|--------------------|---------------------------------------------------------------------------------|
| Test Rig           | • The rotation speed was kept constant at 2000 RPM                                |
|                    | • A radial load of 6000 lbs is applied onto the shaft and bearing                 |
|                    | • All bearings are force lubricated                                              |
| Rolling elements   | • Bearings have 16 rollers in each row                                            |
| bearings           | • The pitch diameter of 2.815 in.                                                |
|                    | • The roller diameter of 0.331 in.                                               |
|                    | • The tapered contact angle of 15.17 degrees                                      |
| Accelerometers     | • PCB 353B33 High Sensitivity Quartz ICP accelerometers                           |
|                    | • Installed on the bearing housing (two for each bearing [x- and y-axes] for data set 1, one for each bearing for data sets 2 and 3) |
|                    | • All failures occurred after exceeding the designed lifetime of the bearing, which is more than 100 million revolutions |

2.2. Data Structure
During the three independent run-to-failure bearing experiments the acquisition of the vibration signals were collected from start to failure with explicit time stamps. Each test dataset consists of individual files that are 1-second signals recorded every 10 minutes. Each file consists of 20,480 points, with the sampling rate set at 20 kHz. This sampling frequency can assure that valuable higher frequency ranges are also captured, which are commonly used for early bearing degradation identification. Data was acquired by DAQCard-6062E NI. Inspections confirmed that test 1 ended up with an Inner Race (IR) failure in bearing 3 and a Rolling Elements (RE) failure in bearing 4. Test 2 and 3 ended with an Outer Race (OR) failure in bearing 1 and 3, respectively [6].

2.3. Equipment life cycle stages
The typical life cycle of the bearings from the beginning of operation until the defect condition could be divided into three stages: normal, degradation, and defect. At the beginning of the operation, the equipment is in a normal stage which can be understood as the operational condition within functional specifications. After any fault occurs, the equipment goes into the degradation stage. Depending on the type of fault, the duration and speed of captured health degradation will vary. This behavior, together with large amount of noise in the input data can result in potential challenges for abnormality identification.

Usually, data collection is not interrupted immediately after a breakdown. Due to this reason, data is still captured from the defected part. In this research, the defective part will be excluded from the investigation after its identification. The target of the first part of the proposed workflow is to identify the fault and failure points and divide the data into phases, which are shown in figure 2.
3. Methodology

In order to create prediction models for root cause identification, supervised learning algorithms will be used. A labelling process, including clustering approaches, will support the creation of this prediction model. This is necessary since the run-to-failure data collection does not contain labels for training a model based on various supervised learning. Thus, the overall workflow was divided into two parts: labelling and modelling. In this research, the unsupervised learning algorithms were used in the labelling part to support the supervised algorithms in the modelling part. The most important parts of the workflow are demonstrated in figure 3.

Figure 2. The typical life cycle of the equipment.

Figure 3. The workflow used for data investigation.
3.1. Labelling workflow
The first step in the workflow is the time-domain feature generation. More than 390 features were created by TSFEL library [8]. Furthermore, several statistic features as crest factor, shape factor were added. In total 406 features were created for each sensor. The examples of several different features are shown in figure 4a. It can be seen that the variation of the values for each feature becomes bigger after crossing some timestamp point. Figure 4b shows a 2-dimensional representation of all the features after implementing the Principal Component Analysis (PCA). It has to be highlighted that all calculations were made in higher dimensions, and the visualizations are only provided for two dimensions, after the application of PCA. However, even in this simplified case, cases that potentially can be abnormal seem to be detectable. No filtering is implemented at this stage to reduce the effect of noise or focus on certain bandwidths of the input signal.

![Figure 4. a) the examples of the created features, b) the two-dimentional representation of all the features after implementing PCA.](image)

After performing data preprocessing, the failure point was identified in order to distinguish the combined normal and degradation stages from the defect stages for each test case. This was implemented using unsupervised learning to divide into two clusters using four different clustering algorithms, such as K-Means, Mean Shift, DBSCAN, and Birch. The final clusters were determined by the implementation of the voting technique. To find a particular split point, the logistic regressions sigmoid were used (figure 5).

![Figure 5. The clustering process.](image)
A similar approach was implemented in the next step to determine the point of splitting the normal and degradation stages (fault point). The first cluster accepted as a normal stage, and the second was accepted as a degradation stage. The data above 0.5 of probability were accepted as degradation cluster, under 0.5-as a normal cluster.

3.2. Modelling for abnormality and root cause identification

The labeled data were used as input for the supervised learning algorithms to create separate abnormality detection models for each type of failure and create a general predictive model for all failures to identify the Root Cause Analysis (RCA). In the modelling part were used such algorithms as Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), XGBoost (XGB). To improve results, the model hyperparameters were optimized through cross-validation. In figure 6 the results from the RF model are presented.

![Figure 6. Modelling results of RF algorithm for: a) IR failure, b) RE failure, C) OR failure.](image-url)
The first model (figure 6a) detected the IR failure in bearing 3 from test 1. It can be seen that the results show a high model performance. All calculations were made in higher dimensions. This figure is only a two-dimensional representation of this higher-order solution. However, in this case, the decision boundaries can be already easily determined. The same is true for the second model (figure 6b) that detects the failure of the RE in bearing 4 from test 1 and, in particular, for model three (figure 6c) that detect the OR failure in bearing 1 from test 2. In OR determining case, the failure resulted in a significant sudden change of system behavior, which made it easy to detect an abnormality.

The confusion matrix of a general model trained for RCA also shows high performance (figure 7a). The separate results of predictive models for each failure are shown on the three subfigures on the left side of the figure including IR, RE, and OR failure detection. In the middle, all various conditions and failures are represented.

4. Results and benchmarking

The results for the rest of the algorithms are shown in table 2. Splitting of the sensor data was carried out in order to implement training and testing. For testing, the one-third part of the data was used, including each third timestamp. The rest of the data was used for training purposes. Cross-validation technique was applied to enhance training performance. In order to evaluate if the trained models generalize well, the OR failure was verified on a bearing dataset, which was acquired from a completely independent test. The IR and RE failures were verified using data from the duplicated sensors from the same bearings. A limitation in those verifications was that for the IR and RE cases, the test conditions were not changed as the setup was the same. Only different sensors were used for the same setup.

| Model          | Parameter | LR test | LR verif. | RF test | RF verif. | NB test | NB verif. | XGB test | XGB verif. |
|----------------|-----------|---------|-----------|---------|-----------|---------|-----------|----------|------------|
| IR failure     | Accuracy  | 0,9975  | 0,9950    | 0,9975  | 0,9959    | 0,9508  | 0,9779    | 0,9950   | 0,9959     |
| detection      | Recall    | 0,9975  | 0,9950    | 0,9975  | 0,9959    | 0,9508  | 0,9779    | 0,9950   | 0,9959     |
|                | Precision | 0,9975  | 0,9951    | 0,9975  | 0,9959    | 0,9675  | 0,9813    | 0,9950   | 0,9959     |
| RE failure     | Accuracy  | 0,9975  | 0,9926    | 0,9926  | 0,9975    | 0,9606  | 0,9991    | 1,0000   | 0,9991     |
| detection      | Recall    | 0,9975  | 0,9926    | 0,9926  | 0,9975    | 0,9606  | 0,9991    | 1,0000   | 0,9991     |
|                | Precision | 0,9975  | 0,9927    | 0,9926  | 0,9975    | 0,9644  | 0,9991    | 1,0000   | 0,9991     |
| OR failure     | Accuracy  | 1,0000  | 0,9122    | 1,0000  | 0,9200    | 0,9970  | 0,8566    | 0,9985   | 0,8977     |
| detection      | Recall    | 1,0000  | 0,9122    | 1,0000  | 0,9200    | 0,9970  | 0,8252    | 0,9985   | 0,8977     |
|                | Precision | 1,0000  | 0,9106    | 1,0000  | 0,9183    | 0,9973  | 0,8728    | 0,9986   | 0,8956     |

This verification process was applied to the general model, where the same failures were aimed to be reproduced. The results of model verification can be seen in figure 7b through a confusion matrix. Even with the same type of bearings, at the same test rig, the results worsened by repeating the same test. This highlights that the models are over fitted on the given dataset and show overconfidence. This over fit is already the result of an optimization, highlighting significant limitations of using datasets without proper considerations of bearing frequencies and filtering of the input data. The generalization is also affected by the implemented labelling, using unsupervised learning, which is not investigated during the benchmarking.
5. Conclusion

During the last decades, a data-driven approach seemed like a powerful tool for condition monitoring. Widespread use has led to its rapid growth. However, it has significant limitations. Using the data-driven approaches results in illusive reliability on the trained model, which can happen due to limitations in the conditions mapped into the studied dataset. Even though the testing results show high performance, the models do not generalize well. It is crucial to keep in mind that in the case of using a data-driven approach, the quality of the model cannot get better than the quality of the used data.

After the discussion of the main limitations of the data-based approach and the performed benchmarking of algorithms, an important conclusion can be drawn. To improve the quality of the input data and the results of trained models, more elements of the knowledge-based approach have to be involved. A hybrid approach, which is the combination of knowledge-based and data-driven approaches, can effectively address the limitations. In such a hybrid approach, knowledge of the physical behaviour of the test case will be involved. The effect of filtering, taking into account the bearing frequencies, will be considered. Furthermore, a detailed study of feature generation tools will
be performed, which uses the physical parameters of the experimental setup. In the examination of the feature selection algorithms, the focus must be on keeping the traceability of the acquired data. The investigation of labelling techniques based on domain knowledge should have a significant impact on the results. The aim of that research will be to create a well-generalized model for different types of machine components, equipment and other industrial machinery.

6. References

[1] Lughofer E and Sayed-Mouchaweh M 2019 Predictive Maintenance in Dynamic Systems Advanced Methods, Decision Support Tools and Real-World Applications (1st ed. 2019, ed.)

[2] Tobon-Mejia D A, Medjaher K, Zerhouni N and Tripot G 2012 A Data-Driven Failure Prognostics Method Based on Mixture of Gaussians Hidden Markov Models IEEE Transactions on Reliability 61(2) pp. 491-503

[3] Carvalho T P, Soares F A A M N, Vita R, Francisco R da P, Basto J P and Alcalá S G S 2019 A systematic literature review of machine learning methods applied to predictive maintenance Computers and Industrial Engineering 137 p. 106024

[4] Zhang S, Zhang S, Wang B and Habetler T G 2020 Deep Learning Algorithms for Bearing Fault Diagnostics—A Comprehensive Review IEEE Access 8 pp. 29857-29881

[5] Qiu H, Lee J, Lin J and Yu G 2003 Robust performance degradation assessment methods for enhanced rolling element bearing prognostics Advanced Engineering Informatics 17(3-4) pp. 127-140

[6] Qiu H, Lee J, Lin J and Yu G 2006 Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. Journal of Sound and Vibration 289(4-5) pp. 1066-1090

[7] Amin Almasi 2017 Vibration, noise and failure in rolling element bearings Turbomachinery International Magazine

[8] Barandas M, Folgado D, Fernandes L, Santos S, Abreu M, Bota P, Liu H, Schultz T and Gamboa H 2020 TSFEL: Time Series Feature Extraction Library. SoftwareX 11 p. 100456