On data-preprocessing for effective predictive maintenance on multi-purpose machines

Lukas Meitz, Michael Heider, Thorsten Schöler, Jörg Hähner

Angaben zur Veröffentlichung / Publication details:

Meitz, Lukas, Michael Heider, Thorsten Schöler, and Jörg Hähner. 2023. “On data-preprocessing for effective predictive maintenance on multi-purpose machines.” In *Proceedings of the 12th International Conference on Data Science, Technology and Applications, July 11-13, 2023, in Rome, Italy*, edited by Oleg Gusikhin, Slimane Hammoudi, and Alfredo Cuzzocrea, 606–12. Setúbal: SciTePress. https://doi.org/10.5220/0012146700003541.

On Data-Preprocessing for Effective Predictive Maintenance on Multi-Purpose Machines

Lukas Meitz[1][a], Michael Heider[2][b], Thorsten Schöler[1][c] and Jörg Hähner[2][d]

[1] *Hochschule Augsburg, An der Hochschule 1, Augsburg, Germany*

[2] *Universität Augsburg, Am Technologiezentrum 8, Augsburg, Germany*

Keywords: Predictive Maintenance, Data Preprocessing, Multi-Purpose Machines.

Abstract: Maintenance of complex machinery is time and resource intensive. Therefore, decreasing maintenance cycles by employing Predictive Maintenance (PdM) is sought after by many manufacturers of machines and can be a valuable selling point. However, currently PdM is a hard to solve problem getting increasingly harder with the complexity of the maintained system. One challenge is to adequately prepare data for model training and analysis. In this paper, we propose the use of expert knowledge–based preprocessing techniques to extend the standard data science–workflow. We define complex multi-purpose machinery as an application domain and test our proposed techniques on real-world data generated by numerous machines deployed in the wild. We find that our techniques enable and enhance model training.

1 INTRODUCTION

Using data-driven models, Predictive Maintenance (PdM)—a Machine Learning (ML) application domain made possible by the ML advances in recent years—detects and predicts machine failures based on collected data, which can lead to better efficiency and reliability while reducing maintenance cost. While there are many models that have been successfully applied (Serradilla et al., 2022), implementations of preprocessing and training are mostly based on a few openly-available or easy-to-collect data-sets.

However, for some real-world applications, data needs to be processed even further before model training. This is especially the case for multi-purpose machinery, which produces heterogeneous data not directly suited for learning. The goal of this article is to introduce three more considerations and steps in preprocessing for PdM in order to make heterogeneous data suitable for model training.

The following article will first give a brief overview of related publications on PdM preprocessing and different representative PdM datasets. The motivation for preprocessing is discussed in a foundation section, followed by a definition of the term complex machinery. Based on the properties and challenges of data collected from the described type of machines, three additional steps for data preprocessing in PdM applications are introduced. A short illustrative example is introduced, which is based on real-world data. The article ends with a short discussion and outlook.

2 RELATED WORK

Maintenance is a huge cost factor in industry and manufacturing, with up to 60% of production cost being spent on it in some cases (Mobley, 2002). Additionally, because of ineffective maintenance actions, about one third of the maintenance cost is estimated to be wasted (Mobley, 2002).

Using the predictive paradigm for maintenance, these cost factors can be reduced. Predictive Maintenance can be described as condition based preventive maintenance, with preventive maintenance meaning replacing parts before a breakdown occurs (Mobley, 2002). If the observed condition changes into a unhealthy state, maintenance is carried out before a machine breakdown occurs.

The condition of machinery is observed using monitoring systems, most of which are data-driven

[a] https://orcid.org/0000-0001-7409-2401
[b] https://orcid.org/0000-0003-3140-1993
[c] https://orcid.org/0000-0001-5487-1862
[d] https://orcid.org/0000-0003-0107-264X

Meitz, L., Heider, M., SchÃ¶ler, T. and HÃ¤hner, J.
On Data-Preprocessing for Effective Predictive Maintenance on Multi-Purpose Machines.
DOI: 10.5220/0012146700003541
In *Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA 2023)*, pages 606-612
ISBN: 978-989-758-664-4; ISSN: 2184-285X

and based on recorded sensor-values. In this section, we will highlight some research regarding the preprocessing of data in PdM applications and give an overview of different types of data-sets that are typically used.

2.1 Preprocessing in PdM

As in any data science project, recorded raw data has to be prepared in order to be used in an application. An extensive overview of the then state-of-the-art in preprocessing of data for PdM was given by (Cernuda, 2019). They focused primarily on standard statistical features and workflows common in other types of ML and their usage in PdM.

(Cofre-Martel et al., 2021) apply similar preprocessing steps and then propose the labelling of this more concise data with the use of expert knowledge. This facilitates the application of supervised learning techniques which is commonly very difficult in PdM due to the rarity of breakdowns but has advantages over unsupervised clustering as run-to-failure cases are highlighted directly (Yun et al., 2021). In their work they highlighted the need for differentiation between the use of real-world and benchmark data.

(Bekar et al., 2020) propose preprocessing to be based on the CRISP-DM cycle and K-Means clustering. In their accompanying case study, they validate their proposed method with data of a simple spindle application which consists of load and vibration observations.

2.2 Data-Sets in PdM Studies

In most cases, data-sets for predictive maintenance applications consist of time-series. They are sensor readings including one or more physical observations, like noise level, vibration, or power consumption. In most of these cases, only a single machine is used to generate the data, which leads to inherent homogeneity and comparability, thus, eliminating the need for excessive data preparation. Some examples of this type of application are centrifugal pumps (Chen et al., 2022), electrical load of washing machines (Casagrande et al., 2021), refrigerators (Kulkarni et al., 2018), or vibration data collected from bearings (Wang et al., 2020; Sugumaran and Ramachandran, 2011). In other scenarios, the data used for implementing PdM-applications are benchmark data sets such as NASA Turbofan (Bruneo and De Vita, 2019) that do not need further processing to effectively train models.

In some cases, only simple statistics of the data (e.g. mean, variance, skewness, etc.) are used for further processing rather than the data itself. This is especially common in vibration or power level monitoring, where, often, only a shift in the underlying pattern is used for anomaly detection. Examples have been published by (Ding et al., 2019) and (da Silva Arantes et al., 2021).

For Deep Learning applications, models handle feature extraction implicitly, therefore, the preprocessing is limited to standard steps such as data cleaning rather than complex manual feature engineering. Examples are the use of Autoencoders for anomaly detection (Sun et al., 2019; Bampoula et al., 2021; Kim et al., 2021).

(Züfle et al., 2022) implement an anomaly detection solution for the degradation of a CNC milling tool. They incorporate a preprocessing step to highlight phases of the same machine action in order to get comparable segments of machine operation, i.e. when the machine is actively milling material.

PdM literature does rarely focus on specific data preprocessing articles as many data-sets are rather simplistic not requiring special preparation or do only contain very limited sensor and actuator reading variety, i.e. focus on a singular part and its condition. In most research studies, the type of application does not create a need for dealing with data heterogeneity. In the following, we will describe some techniques for dealing with the type of complex data that is found in many real-world applications.

3 FOUNDATIONS

As a foundation for the remainder of this article, the term *complex machinery* and its characteristics in the domain of PdM are introduced. Furthermore, this section gives a brief overview of the motivation and importance behind data preparation specific to PdM.

3.1 Data Preprocessing

Preprocessing has multiple roles that are important for the further progress of a data science project. A general definition of preprocessing is given in (Cernuda, 2019): ”the set of actions performed to raw data [...] with the aim of improving the modelling capabilities”. However, there are more specific goals that all aim for better model performance, as described in (Luengo et al., 2020):

Reduction of Size and Complexity. Reduced size improves runtime performance, which speeds up model training and inference. Reduced complexity enables a model to achieve faster convergence and allows the use of models with less parameters which might help their explainability.

Format Conversion of the Data. Depending on the application and model, different sizes or chunks of data are needed. Most models need fixed-length input or smaller samples of the data. Some work with aggregations or statistical features, that have to be extracted from the data.

Retention Only of Important Information. Most importantly, preprocessing is used to filter data to only the important information used for model training. Model performance is greatly aided by eliminating noise and distracting signals beforehand.

3.2 Heterogeneous Machine Data

Data-driven Predictive Maintenance aims to extract information about a machine's condition based on collected data. The aforementioned preprocessing of recorded data is a necessary step to extracting this information.

As shown in Section 2, applications have been implemented based on a variety of sensor values that are collected from runtime data. Most of the time however, the data observes a singular component that performs a specialised task like rotation, pumping, or pressing. Although this is an important foundation for applications in industry and commerce, there are many remaining application domains that have not been subject to research because of their machine complexity.

Complex machines that can be used for multiple applications, when observed by sensors, create heterogeneous data. To clarify the type of complexity referred to in this article, the following properties of complex multi-purpose machinery are introduced:

1. It can be used for more than one *application*:
   The machine is able to perform a specific parameterised process, i.e. milling, but is automated to a level where it can produce results of high variety which is leading to vastly different data, e.g. because of different lengths, rotations of the drill, motor movements of varying speeds, etc. We call one such parametrisation of the process an application.
2. It consists of *multiple components*:
   Such components can be different actuators and sensors—not accounting for the sensors used purely for the PdM application or other monitoring tasks. For our definition, we assume that the number of components is at least five with at least three actuators. A prototypical example would be a multi-pump system where distributed pressure sensors trigger pump activation cycles.

Challenges. When collecting operation data from complex machines, some challenges emerge for the further processing of this heterogeneous data. These challenges are closely related to the domain of big data and are the result of the properties volume, variety and variance (Sagiroglu and Sinanc, 2013).

There is a high amount of data collected from the different components, not all of it useful for model training. This data volume requires space to be stored and computing performance to be efficiently processed.

Another problem is the handling of complexity, or variety, from the many different components and applications of complex machines. Each component can be of a different kind and therefore produce a different type of data. For each application of the machine, the underlying process can be different and subject to parameters leading to different observable states.

The collected data contains significant heterogeneity in its patterns because of the different circumstances of machine operation. This results in high variance and noise in physical observations (i.e. process lengths, starting conditions, or material properties) and can lead to poor model performance.

Example. To give an example of a complex machine, we will look at the case of a *CNC-mill*—a commonplace machine in modern manufacturing plants. Its actuators consist of multiple motors, one or more for each axis, and a spindle for material removal. To sense the physical state, there are limit switches and spindle load measurements. Additionally, vibration and noise sensors are often attached to the machine, to enable PdM scenarios. Using this set of actuators and sensors, the machine is able to process different materials to produce many different parts. When recording the operation data, this application variety leads to raw data with a high variance, which is not directly comparable amongst different processes. For this or similar types of machinery, it is therefore necessary to further break down and prepare raw machine data to create usable data sets.

4 PROPOSED METHODS FOR PREPROCESSING OF HETEROGENEOUS MACHINE DATA

To overcome the challenges mentioned in the last section, state-of-the-art data preprocessing, as seen in related PdM applications, is not sufficient. In this section, three additional important considerations are introduced, which have to be taken into account during the data preparation process for PdM applications of complex machinery.

These considerations can be implemented as independent steps and are suited for the preparation of real-world complex machine data, whose characteristics and challenges have been explained beforehand. Most of the steps rely on the application of expert knowledge and are not intended as fully-automatic implementations.

4.1 Use-Case-Specific Data Selection

There are two possible scenarios for implementing PdM in an industrial environment. With multiple machines available, two use-cases can emerge for the collected data: creating statements relative to a single machine and creating statements about a group of machines. In the case of PdM, the modelling scope, i.e. one or many supposedly identical machines' data, can decide on how to partition the data during preparation. This partitioning can help reducing the variety and volume of data.

The decision that is implicitly made is which kind of data variance will be selected in the course of preprocessing. For single machine applications, time-variance will be detected, as this is the main feature that changes over available data recordings. This is useful for finding trends or anomalies in the data created by the machine in different points in time. A degradation or failure can be found this way, by comparing the past data points to new ones.

When trying to establish time-variance for multiple machines, the slight difference in hardware of each device will be enough to introduce enough noise to mask the time-variance. Therefore, in a multi-machine setting the suitable application is anomaly detection based on the bulk of the hardware. When using multiple machines, one can create statements about the majority behaviour of such machines. Outliers can be found this way, which are machines that differ from the mean in the data they generate.

To conclude, depending on the target application and the type of model, a selection of only the relevant parts of the available data is necessary. Single-machine models can learn from historical data and establish degradation trends. Multi-machine models are useful for finding anomalies or outliers in a set of different machines. Depending on the sought out information, further selection of only relevant subcomponents may be helpful.

4.2 Discerning Actuators and Sensors

As stated in the introduction of heterogeneous data, time-series data generated by complex machines can be split into two distinct categories: actuator control signals and sensor observations. By distinguishing these two types of data, the data handling can be improved.

Actuators. Actuator control signals are generated by the machine's internal controller. They are often of binary form, take discrete values from a fixed set of possibilities, or are real-valued. Because they are generated by a controller, they encode information about the machine's inner state. This is a useful asset for reducing variety in the data, as each observation can be annotated with a supposed state as assumed by the machine controller. Examples for binary signals are relay-controlled heating, valve opening, or pump operation. More high-level control signals could convey the current machine status, target temperature, or target water quantity.

Sensors. Sensors are observers of the physical world, they measure a value associated with a physical property, and their values are typically continuous. Because of the nature of sensing, values can be subjected to noise and outliers. The information encoded into sensor measurements is useful for observing the actual physical state (in contrast to the controller's 'set'-state) of a machine. Examples are temperature, pressure, water flow, or electrical energy.

Division and Preparation. Dividing data into the aforementioned groups can help in creating models by reducing variance to the desired scope. To illustrate: Actuator control signals are useful for discerning the machine state. For a specific machine state, models can be trained using only sensor measurements, which represent the physical state of a machine, thereby, automatically discarding multiple input signals of varying noise that would not aid predictions at all. Because actuator control signals are generated by the machine controller itself, they do not change for the same machine state. Physical observations, however, can change over time based on the underlying hardware and make useful features for degradation modelling.

4.3 Data Segmentation

Complex (multi-purpose) machines are used for more than one application (cf. Section 3.2). This means that processes, e.g. manufacturing a part, are not directly comparable to each other because every process is either different due to its parametrisation or executed under different circumstances. This difference in circumstances leads to data variance in collected observations that is much higher than the possible degradation effect (or rapid shift) which would warrant maintenance. To overcome this issue, data segments have to be created, that contain only data gathered under similar circumstances.

Using the separation of control signals and sensor measurements from the previous section, segments of similar machine state can be found by comparing control signals. For similar patterns in actuator signals, the machine is likely in a similar state and the sensor measurements are therefore comparable. There are multiple possibilities for creating such segments of similar circumstances, some of which are explained in the following paragraphs.

Rule-Based Segmentation. Using existing expert knowledge, a simple solution for creating time-series clusters is creating segmentation rules. This assumes that knowledge about the machine's processes is present an can be formulated as simple if-then-rules.

Pattern Matching. Using a recurring pattern in actuator signals, common circumstances in the machine behaviour can be found. This is useful for fixed-pattern processes that stay the same for every occurrence of the process.

Dynamic Time Warping. For processes that do not have the necessary actuator signals for clustering, Dynamic Time Warping (DTW) (Müller, 2007) can be used as another approach. Initially, a domain expert has to establish a reference process and create segment labels for later segmentation. DTW can now be used to establish a mapping from an observed time-series to the already segmented reference process. The labels can now be transferred from reference to observation and a new comparable segment is created.

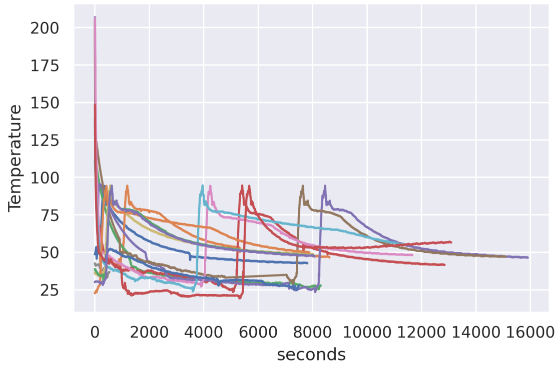

Figure 1: Temperature Readings from the same type of process.

5 EXAMPLE IMPLEMENTATION

To illustrate a possible implementation, we will apply the aforementioned techniques on a small data set sampled from real-world machine operation. The data was sampled from a single machine and process type, however this process is dependent on different starting conditions and parameters. This leads to high variance in the process length and sensor values, which is not suitable for information extraction in its raw state.

Figure 1 shows a plot of the recorded processes and illustrates their heterogeneity. Standard data preparation, as introduced by Cernuda (Cernuda, 2019), was conducted on the data beforehand.

Data Selection. The example data set contains multiple recordings of the same type of process observed on a single machine. This machine consists among others of a heater, a pump, and a temperature sensor. A first step to implementing a PdM model is to decide on the model scope. In this case, single-machine trend detection is the desired application. The goal is to compare a temperature degradation trend in this type of process, therefore only relevant observations to this phenomenon are selected, which are *Pump*, *Heating* and *Temperature*. Note that while the type of process is the same for each observation they vary greatly in length, which is one of the challenges in this setting.

Actuator and Sensor Separation. The example data set consists of multiple readings: actuators signals for a pump and a heating element as well as sensor readings of temperature. By employing domain expertise, the temperature sensor is selected for observation and heating and pump signals are used for machine state representation. More complex scenarios

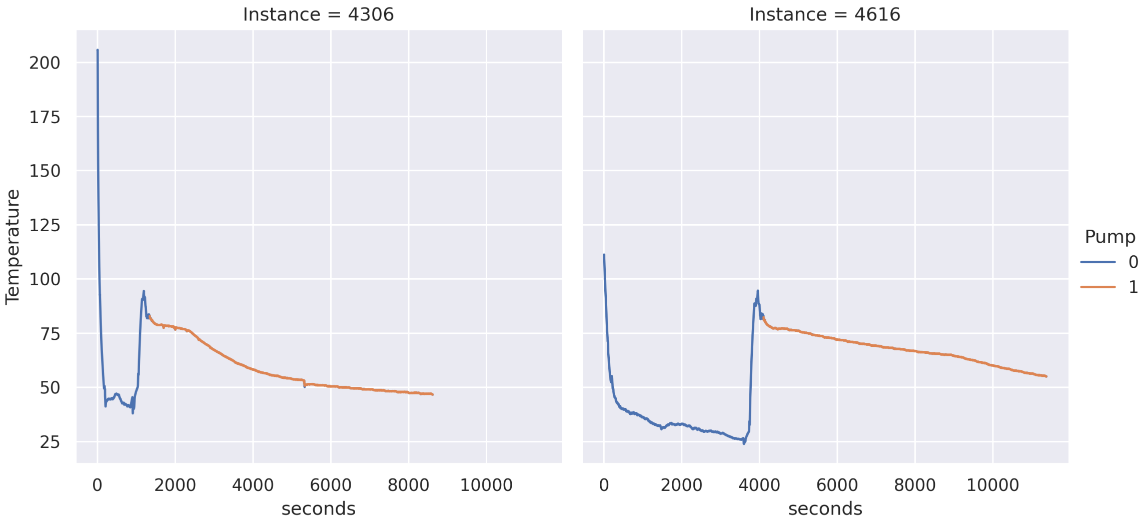

Figure 2: Two process instances with highlighted active pump segments.

can be handled by creating sets of the unique values of each variable as actuators take only a small number of specific values, sensors produce a great amount of distinct values, as the physical observation is often bound to be continuous.

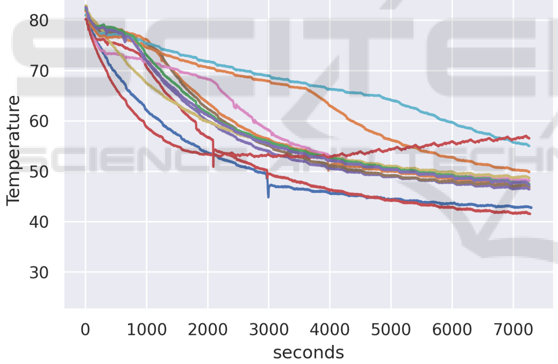

Figure 3: Extracted segments of the same process for comparison.

Process Segmentation. Using the set of actuator values, the selected process can be further segmented to synchronise the sensor readings and aid observation comparability. Only a part of the process recording is relevant, therefore a simple rule-based approach for process segmentation is used. In the simple case of this pump and heating system, the segmentation is achieved by selecting the process segments with an active pump signal. Figure 2 highlights the selected and relevant process segment of two specific instances. After extracting only the relevant segments, observations are already intuitively comparable, as shown in Figure 3. Using these previously selected sensor readings, that have afterwards been segmented into comparable chunks of similar length, further implementation of PdM has been enabled.

6 CONCLUSION

Data preprocessing for PdM applications is notoriously hard when data of systems with many interacting components was collected in the wild. This article described additional considerations to be made for data preprocessing in PdM applications and gives an example of their effects. As many systems for which PdM has successfully been developed are on the simpler side, i.e. few moving parts, few components, or few control signals, we gave a definition for complex multi-purpose machines (multiple components; multiple applications) and used real-world data from one such machine for our examples. Three steps for creating data sets useful for model training and further processing are proposed in the main section of this paper. These are data selection based on the desired use-case, actuator and sensor clustering, and finally data segmentation.

To highlight the effects of the at-first rather abstract techniques, an example data set of a real-world machine has been processed by applying the proposed methods. Using the additional constraints described in the article, the heterogeneous recorded data was processed into segments of comparable information.

There are some limitations to our approach and presented preliminary results, which can be improved upon in the future. The considerations described in the main section are dependent upon domain knowledge and the manual application of data science expertise. This means the process of preprocessing incorporating the proposed steps has to be implemented for each application separately. However, the proposed techniques enable a further processing of data in different use cases. Model training and other applications can be implemented after creating homogeneous data sets, which would not be possible or well-performing in the case of raw and heterogeneous machine data.

For future articles, a formalisation of the proposed process is a necessary step. Additionally, by generalising the steps and implementing them in different application scenarios, a comprehensive evaluation will be an important next step.

REFERENCES

Bampoula, X., Siaterlis, G., Nikolakis, N., and Alexopoulos, K. (2021). A Deep Learning Model for Predictive Maintenance in Cyber-Physical Production Systems Using LSTM Autoencoders. *Sensors*, 21:972.

Bekar, E. T., Nyqvist, P., and Skoogh, A. (2020). An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. *Advances in Mechanical Engineering*, 12.

Bruneo, D. and De Vita, F. (2019). On the Use of LSTM Networks for Predictive Maintenance in Smart Industries. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 241–248.

Casagrande, V., Fenu, G., Pellegrino, F. A., Pin, G., Salvato, E., and Zorzenon, D. (2021). Machine learning for computationally efficient electrical loads estimation in consumer washing machines. *Neural Computing and Applications*, 33:15159–15170.

Cernuda, C. (2019). On the Relevance of Preprocessing in Predictive Maintenance for Dynamic Systems. In Lughofer, E. and Sayed-Mouchaweh, M., editors, *Predictive Maintenance in Dynamic Systems: Advanced Methods, Decision Support Tools and Real-World Applications*, pages 53–93. Springer International Publishing, Cham.

Chen, L., Wei, L., Wang, Y., Wang, J., and Li, W. (2022). Monitoring and Predictive Maintenance of Centrifugal Pumps Based on Smart Sensors. *Sensors*, 22:2106.

Cofre-Martel, S., Lopez Droguett, E., and Modarres, M. (2021). Big Machinery Data Preprocessing Methodology for Data-Driven Models in Prognostics and Health Management. *Sensors*, 21:6841.

da Silva Arantes, J., da Silva Arantes, M., Fröhlich, H. B., Siret, L., and Bonnard, R. (2021). A novel unsupervised method for anomaly detection in time series based on statistical features for industrial predictive maintenance. *International Journal of Data Science and Analytics*, 12:383–404.

Ding, H., Yang, L., and Yang, Z. (2019). A Predictive Maintenance Method for Shearer Key Parts Based on Qualitative and Quantitative Analysis of Monitoring Data. *IEEE Access*, 7:108684–108702.

Kim, D., Lee, S., and Kim, D. (2021). An Applicable Predictive Maintenance Framework for the Absence of Run-to-Failure Data. *Applied Sciences*, 11:5180.

Kulkarni, K., Devi, U., Sirighee, A., Hazra, J., and Rao, P. (2018). Predictive Maintenance for Supermarket Refrigeration Systems Using Only Case Temperature Data. In *2018 Annual American Control Conference (ACC)*, pages 4640–4645.

Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., and Herrera, F. (2020). *Big Data Preprocessing: Enabling Smart Data*. Springer International Publishing, Cham.

Mobley, R. (2002). *An Introduction to Predictive Maintenance*. Plant Engineering. Elsevier Science.

Müller, M. (2007). Dynamic Time Warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer, Berlin, Heidelberg.

Sagiroglu, S. and Sinanc, D. (2013). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 42–47.

Serradilla, O., Zugasti, E., Rodriguez, J., and Zurutuza, U. (2022). Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. *Applied Intelligence*.

Sugumaran, V. and Ramachandran, K. I. (2011). Fault diagnosis of roller bearing using fuzzy classifier and histogram features with focus on automatic rule learning. *Expert Systems with Applications*, 38:4901–4907.

Sun, C., Ma, M., Zhao, Z., Tian, S., Yan, R., and Chen, X. (2019). Deep Transfer Learning Based on Sparse Autoencoder for Remaining Useful Life Prediction of Tool in Manufacturing. *IEEE Transactions on Industrial Informatics*, 15:2416–2425. Conference Name: IEEE Transactions on Industrial Informatics.

Wang, J., Liang, Y., Zheng, Y., Gao, R. X., and Zhang, F. (2020). An integrated fault diagnosis and prognosis approach for predictive maintenance of wind turbine bearing with limited samples. *Renewable Energy*, 145:642–650.

Yun, H., Kim, H., Jeong, Y. H., and Jun, M. B. G. (2021). Autoencoder-based anomaly detection of industrial robot arm using stethoscope based internal sound sensor. *Journal of Intelligent Manufacturing*.

Züfle, M., Moog, F., Lesch, V., Krupitzer, C., and Kounev, S. (2022). A machine learning-based workflow for automatic detection of anomalies in machine tools. *ISA Transactions*, 125:445–458.