An experimental setup for learning models for the RUL estimation of bearings in rotary machine tools: towards the learning of more interpretable models

Charles-Maxime Gauriat, Yannick Pencolé, Pauline Ribot, Gregory Brouillet

An experimental setup for learning models for the RUL estimation of bearings in rotary machine tools: towards the learning of more interpretable models

Charles-Maxime Gauriat[1], Yannick Pencolé[2], Pauline Ribot[3], and Gregory Brouillet[4]

[1,4] *BOSCH, Rodez, France.*
CharlesMaxime.Gauriat@fr.bosch.com
Gregory.Brouillet@fr.bosch.com

[1,2,3] *LAAS-CNRS, Université de Toulouse, CNRS, INSA, UPS, Toulouse, France.*
ypencole@laas.fr pribot@laas.fr

ABSTRACT

This technical brief reports about an experimental setup whose objective is the acquisition of relevant data for learning a RUL estimation model for spindles in rotary machine tools that are available in BOSCH-Rodez manufacturing site. Two classical machine learning methods have then been used to evaluate the quality of the acquired data. First results indicate that such a setup could be effective for building interpretable models to be embedded in the global strategy for the predictive maintenance of machine tools on site.

1. INTRODUCTION

BOSCH-Rodez plans to improve the maintenance of its fleets of machine tools by instrumenting them in order to acquire data to implement a predictive maintenance strategy over the site. Machine tools are complex systems composed of motors, spindles, chucks, oil pumps. Predictive maintenance consists in optimally deciding when to replace a component in the machine tool so that the machine tool is always operating properly. It also helps to prevent manufacturing waste.

To get such a predictive maintenance strategy, the objective is to add relevant sensors in the machine tools, to measure data at operating time and use a prognostic model to check the current health and predict the Remaining Useful Life (RUL) of every component of the machine at any time based on the current set of measured data. To reach that goal, a first step is to investigate how to design such a prognostic model for a specific component of the machine tools: a spindle. Indeed, the most critical parts of such a system are the spindles, as there are the ones that are the most frequently replaced. This technical brief reports about an experimental setup to acquire data and use machine learning techniques to learn a prognostic model on a specific type of spindles called UVA6011 based on vibration signals (Sassi, Badri, & Thomas, 2007).

Some Machine Learning (ML) methods have been developed that learn the degradation behavior of bearings (Riley, Debray, Moons, Schaar, & Hemingway, 2019) and synthesize predictive models (Rao, Pai, & Nagabhushana, 2012). For instance, ML methods like K-Nearest-Neighbours and Support Vector Machine have been used for RUL prediction through classification models (Chelmiah, McLoone, & Kavanagh, 2022). Amongst these ML methods, Neural Network approaches are more and more used as they can handle large and complex computations to produce efficient models. However, even if already proposed ML methods greatly improve RUL prediction and more generally solutions for Prognostic and Health Management (PHM), further investigations about the effective quality of the models are required, especially their interpretability. At present, most of the models used for prediction are set up based on accuracy score and computational speed rather than on the human ability to understand them (Rudin, 2019). Models obtained by deep learning techniques are known to be black-boxes, meaning that they cannot be open to understand their decisions as interpretable. Some previous works have reached a certain level of explainability like (Lundberg & Lee, 2017) but these results do not give the full insight about how interpretable the effective model's choices are. However, it is important for a human operator to understand how an algorithmic model determines a maintenance decision with respect to human-interpretable physical laws and quantities: *how and why* such a model plans the decision. In fact, interpretability give operators the confidence needed towards trained models (Marcinkevics & Vogt, 2020).

This paper is organized as follows. Section 2 describes the experimental setup and details the type of dataset that is acquired. Section 3 details two machine learning techniques

that have been applied on this dataset. The first one is a Neural Network (NN) approach, detailed in Section 3.1. By implementing this approach, the objective is to firstly ensure that acquired data have good quality and the proposed experimental setup is relevant: indeed a classification method like neural networks, that is known for having good performance with complex problems and non-linearity, should confirm that the acquired data are relevant as soon as it synthesizes a model that is highly accurate. The second reason is to confront its performances to an interpretable approach, the Decision Tree (DT) approach, detailed in Section 3.2. DT are known to be simple to implement and easy to interpret. However, these models reach high performance only if the relations between input features and the output are simple. The aim of this approach is to provide a prognostic model that is more interpretable than the NN and to check whether such a simpler model can reach the same performance as the NN while remaining interpretable with the proposed dataset. A comparative analysis of both methods applied on the dataset is then proposed in Section 4.

2. System and Data Acquisition

This section reports about the experimental setup to acquire and process real data in near-real conditions for learning a prognostic model for a set of UVA6011 spindles that are available at BOSCH-Rodez manufacturing site.

2.1. Prognosis and predictive maintenance of a spindle

A spindle itself is a complex system and, by experience, the critical element of the spindle is the bearing as it is the first component of a spindle to become defective and therefore the first component to make the overall spindle failing. In this context, predictive maintenance then consists in deciding when to replace such bearings in the spindles so that the spindles are always operating properly. The following experimental setup proposes to acquire vibration data from the spindle's bearings and to learn from them a prognostic model that can be used afterward at operating time.

2.2. Experimental setup

To initiate the experiment and get first training measured data, an available set of five spindles has been installed on a test-bed in a closed chamber that simulates the real operating conditions (see Figure 1). Each spindle is connected with cooling oil inlet and outlet for cooling down the spindle (lubrication of the bearings). Then a VSA 005 sensor (the accelerometer) has been screwed on the front part of the spindle as it would be inside a machine tool using this type of spindle at operating time. This vibration sensor is thus located between the pair of bearings of the spindle. It measures the vertical acceleration just on the top of these bearings. As the space between the pair of spindle bearings is too close, the sensor is placed between them and acquires a single vibratory signal from the pair of bearings. This is not a problem as bearings are always both replaced if a maintenance action is required. The cable of the VSA sensor is connected to an IFM VSE 100 module that contains the software allowing to record the vibratory signals, a module that is used at operating time. The objective is to acquire data at 9K RPM so that frequency measurements are within the range of the available VSA sensor.

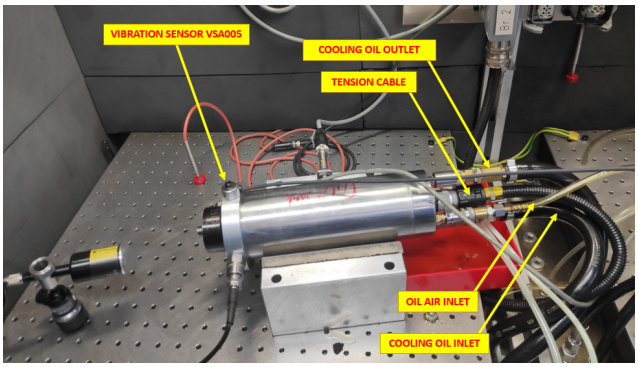

Figure 1. The experimental test-bed

The five spindles that have been used in this experiment have been selected by a maintenance operator. Based on the operator's expertise, these five spindles are in five distinct degradation modes (or classes), denoted C_m with $m \in \{0, 3, 4, 7, 9\}$. Mode C_m means that the spindle'RUL is m months. When $m = 9$, the spindle's bearings are new, when $m = 0$ they are worn out (see Table 1).

Degradation Class $\mathcal{C}_{RUL}$	RUL (months)	RUL(%)
C_9	9	100
C_7	7	77
C_4	4	44
C_3	3	33
C_0	0	0

Table 1. Labeling of the degradation class of a spindle associated with the RUL

2.3. Data acquisition and processing

The IFM VSE 100 module is the interface between the sensor and the computer and is able to record raw signals. The sampling is set to the frequency 100kHz (100K values per seconds). No signal filters have been applied. For each spindle, raw signals have been recorded during 16 minutes. From each raw signal, 10000 signals of one second each have been extracted (as a set of sliding time windows with a sliding step of 0.1 second). Then, for each of the 10000 signals, it has been converted as a frequency spectrum by applying a Fast Fourier Transform (FFT). The bandwidth set of the full spectrum is from 0 to 9985Hz to remain in accordance with the range of the vibration sensor, which is fixed to a maximum

of 10kHz. This spectrum is then discretized to get $F = 409$ frequency amplitude values for each FFT signal with a frequency resolution of 24.414Hz. Finally, the available training dataset X is a set of $N = 50000$ individuals, an individual is composed of $F = 409$ features that are frequency amplitude values. Formally, the dataset is:

$$X = \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_N} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,F} \\ \vdots & & \vdots \\ x_{N,1} & \cdots & x_{N,F} \end{bmatrix}$$

Each individual $\mathbf{x_i}$ is associated with a label $l_i = C_m \in \mathcal{C}_{RUL}$, where C_m corresponds to the degradation class of the spindle that produces $\mathbf{x_i}$. Let $L = [l_1 \dots l_N]^T$ denote the labels associated with individuals X, by construction, each degradation class is then represented by 10000 individuals in X.

3. Prognostic model learning: first results

This section now reports on the application of two classical machine learning algorithms on the labeled dataset (X, L). The objective is to learn the function:

$$f_{RUL} : \mathbf{x} \rightarrow C_t \tag{1}$$

Given any new individual $\mathbf{x}$ (frequency spectrum as defined in Section 2.3) that is acquired on UVA6011 spindles at operating time, f_{RUL} maps the individual $\mathbf{x}$ to one of the classes C_t defined by the maintenance expert. If $f_{RUL}(\mathbf{x}) = C_t$ then it means that the corresponding spindle has a RUL of t months.

The first stage of the learning process is the construction of a training set and a testing set out of (X, L). As a lot of data are available (about 50K individuals), the training set consists of 70% of the individuals from the dataset and the testing set is composed of all the individuals of (X, L) not selected by the training set (30% of the individuals). The training set will be used to train the prognostic model to recognize a label l_i in function of $\mathbf{x}_i$. The testing set is used to validate the accuracy of the prognostic model and detect potential overfitting. Indeed, if accuracy metrics used to evaluate the prognostic model show on the testing set a lesser score than the training set, then the prognostic model is overfitting and is failing to generalize, so that it is unable to correctly recognize a class on data it has never seen.

The analysis of the performance of the learning algorithms on this dataset relies on the use of an accuracy score and a multi-class confusion matrix, as the data in X are well-balanced (Grandini, Bagli, & Visani, 2020). Any correctly classified individual is defined either as a True Positive (TP) or a True Negative (TN) and any incorrectly classified individual is defined either as a False Positive (FP) or a False Negative (FN). The accuracy is then defined by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

A multi-class confusion matrix M can also perform evaluation of any predictive model based on its classification score. The element M_{ij} of the matrix is the number of individuals that are predicted in class j while its real class is i. True Positive individuals then appear in the diagonal of the matrix.

3.1. Neural network approach

For the first training, it has been decided to use a simple neural network (NN) approach because it can produce efficient models on nonlinear data with a large number of inputs. Moreover, trained NN models do not require much time to produce predictions once in use.

Let NN be the learning Neural Network function such that:

$$NN : X, L \rightarrow f_{NN} \tag{3}$$

with X be the matrix of FFT signals, L be the label vector associated with X and f_{NN} be the prognostic model that is learned by NN. Function f_{NN} is therefore the approximation of f_{RUL} learned by the proposed NN approach. To train f_{NN}, 2 hidden layers have been used. The first one is defined with 16 neurons and the second with 32. A Rectifier Linear Unit (ReLU) is chosen as an activation function. The output has 5 neurons, one for each class of $\mathcal{C}_{RUL}$ with a softmax activation function. Categorical cross entropy loss has been chosen as it is one of the most used for multi-class classification problems and for its capacity to converge faster than other loss functions. Finally, ADAM optimizer has been chosen for the stochastic gradient descent because it is computationally efficient and well suited for large data sets with a lot of parameters (Kingma & Ba, 2014).

The accuracy score is firstly used to measure the potential overfitting of the prognostic model f_{NN} during the learning phase. When the training starts, 10% of the training set is automatically extracted as a new dataset called a validation set. This dataset is used to observe the performance evolution of the prognostic model during the learning phase. Indeed, at the end of each learning iteration (epoch), the accuracy of the prognostic model is evaluated both on the training set and the validation set, which produces two curves showing the evolution of its evaluation based on the selected metrics (accuracy score in this case). By comparing both curves, it is possible to understand the behavior of the predictive model performance during its training. If the curves converge, the model is generalizing. In case of divergence, the training can be failing to get a proper classification, or it can be overfitting. In the latter case, methods preventing overfitting can be applied, like early stopping or regularization (Ying, 2019).

In this experiment, the prognostic model has been trained by

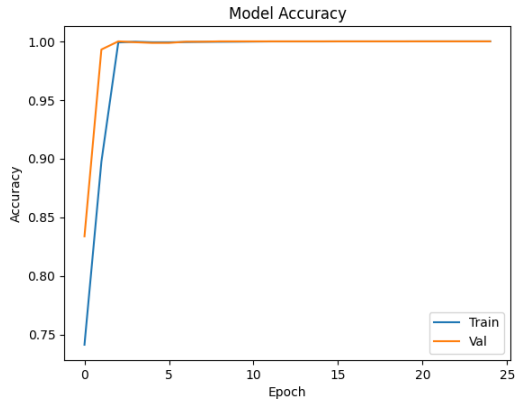

Figure 2. Accuracy - Neural Network Approach

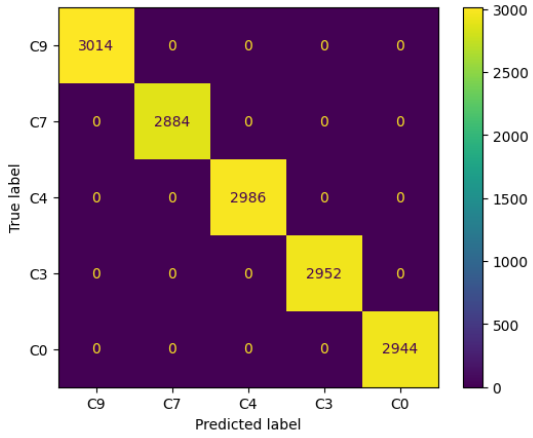

Figure 3. Confusion matrix - Neural Network Approach

setting 25 learning iterations, but it has actually converged after the second learning iteration as shown in Figure 2: this figure represents the f_{NN} prognostic model trained on dataset X after each epoch, and results are finally 100% accurate.

Secondly, to see how efficient the trained models are after the learning phase and to verify the quality of the predictions for each class, multi-class confusion matrices on test sets are computed as well as the overall accuracy. No overfitting has appeared. The confusion matrix in Figure 3 shows the distribution of the individuals from the testing sets with respect to their predicted class and their true class: no individual has a wrong prediction.

This shows that a simple NN model trained with few parameters and few iterations can produce a prognostic model that is already able to get high accuracy scores and determine a class C_t of a spindle bearing based on the data acquired by the proposed experimental setup.

3.2. More interpretable prognostic model: Decision-Tree approach

With f_{NN}, high accuracy scores are obtained. However, a problem remains: f_{NN} is a black-box model. Such a predictive model cannot be interpreted, which is a big issue in a lot of work fields like manufacturing. Indeed, an operator must understand why an advisor model takes a maintenance decision. As the NN approach is very performant on the dataset (X, L), it motivates the use of another learning approach that is known to be less performant but that increases the level of interpretability: a decision-tree approach (DT).

$$DT : X, L \rightarrow f_{DT} \tag{4}$$

One of the strength of a DT approach is its capacity to provide a prognostic model f_{DT} (i.e. approximation of f_{RUL}) that is an interpretable decision tree, but its drawback is that this type of predictive model is known for having lower performance than NN and the learning process can be slower (Kim, 2008). In fact, a DT method is efficient only if the intrinsic relations between input data and their respective output decisions are simple and general. The way the NN approach converges on the dataset (X, L) (see Figure 2) leads to the conclusion that (X, L) could be simple enough for a DT-approach to perform well.

For this approach, it has been decided to train without tuning hyperparameters. There is no max depth for this tree, and the quality of a node of the decision tree is measured through the Gini index. The Gini index has been chosen for its computational speed compared to entropy. Its purpose is to indicate the probability that a given node of the tree makes a wrong decision to classify an individual. Let n be the current node of the decision tree. Let $X_n \subset X$ be the set of individuals that are passing through this node. Let $X_{t,n}$ be the set $X_{t,n} = \{x \in X_n, x \in C_t\}$ the impurity of the node n is given by the following Gini index G_n:

$$G_n = \sum_{t=1}^{|\mathcal{C}_{RUL}|} p_{t,n} \times (1 - p_{t,n}) = 1 - \sum_{t=1}^{|\mathcal{C}_{RUL}|} p_{t,n}^2 \tag{5}$$

with $p_{t,n}$ being the estimation of the probability that any individual $\mathbf{x}$ passing through the node n has a class C_t, that is:

$$p_{t,n} = \frac{|X_{t,n}|}{|X_n|}.$$

For each node, the Gini index gives a score between 0 and 0.5 with 0.5 being the highest level of impurity. This means that an individual has a random probability of being wrongly classified in the node. A node reaching a Gini index of 0 means that every individual $\mathbf{x}$ in it belongs to one class only. Thus, this node becomes a leaf node displaying a unique class C_t.

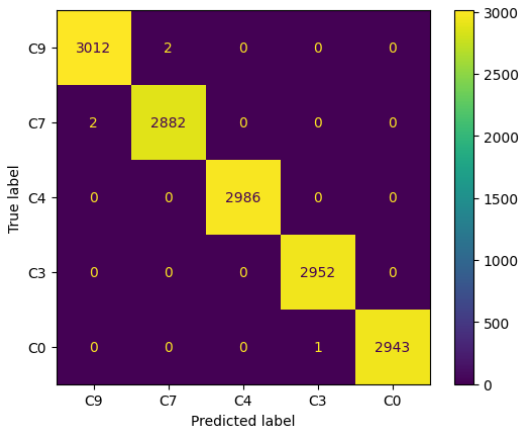

Figure 4. Confusion matrix - DT Approach

Figure 4 shows the confusion matrix of the learned prognostic model f_{DT} on the dataset (X, L). From this matrix, it follows that f_{DT} performs with 99.96% of overall accuracy. The confusion matrix shows few miss-classed individuals for C_9, C_7 and C_0. As expected, the dataset (X, L) is easy to model and that is why the trained model performs with a high accuracy score. However, the interest here remains in the understanding of why and how the predicted class C_t has been chosen by f_{DT}. The strength of DT remains in its capability of providing an interpretable flow chart as a decision tree.

The obtained prognostic model f_{DT} is entirely presented in Figure 5. It is simply composed of 13 nodes. To assign the class C_t to any newly available individual $\mathbf{x}$, it only consists in selecting the unique branch from the root node to the leaf node such that $\mathbf{x}$ holds all the conditions in the branch and assigning to $\mathbf{x}$ the class of the leaf node of the branch. The interpretability of this model relies on the fact that the set of conditions leading to the leaf node of the selected branch are physical features (frequency amplitudes from the underlying FFT signal). Indeed, nodes split according to the amplitude a_i assigned to a frequency i of the FFT spectrum in $\mathbf{x}$.

The model f_{DT} performs with a high accuracy score and is now easy to interpret as opposed to f_{NN}. Indeed, an operator can clearly understand that, if the current measured vibratory amplitude at the frequency 6392 Hz is lesser than 0.826 m/s^2 but greater than 0.164 m/s^2 for 6368 Hz then the RUL of the spindle is about 3 months (C_3).

Finally, by analyzing the tree produced by the DT approach (see Figure 5), it can be visually noticed that the classes are somehow ordered: the shorter the RUL of the class, the further to the right the class is on the tree. It results from a decision by f_{DT} that if the amplitude of the signal $\mathbf{x}$ at a given frequency is greater than a predetermined amplitude, the corresponding spindle is expected to be more worn out. This leads to the conjecture that a prognostic model decides not really by selecting a specific frequency in a FFT signal, but by selecting over the amplitude differences between individuals.

4. Discussion

This technical brief describes an experimental setup to acquire and process data for learning prognostic models for bearings in spindles. Previous experimental analyses first show that the quality of acquired data is correct in the sense that the tested learning methods reach similar results: on one hand both methods obtain a model with high accuracy score and on the other hand these models generalize well and avoid overfitting. Comparing these two approaches, DT shows that it could perform as well as NN because the acquired data hold simple intrinsic relations between the input and the output. DT has the advantage to offer interpretable results to the maintenance operators that make them understand the decision of the model through the tree. From a maintenance operator viewpoint, DT should be as simple as possible (not having too many nodes) to be user-friendly. From a data analysis viewpoint, the obtained DT f_{DT} leads to the conjecture that the decision is more a matter of selecting amplitude differences than selecting specific frequencies in the spectrum. Furthermore, this experiment shows that the tradeoff between precision and interpretability for a prognostic model can be minimal as soon as the data are well processed and homogeneous. This is why it remains important to evaluate this tradeoff as losing a few percentage of accuracy to bring interpretability gives to the operator more confidence to apply the maintenance decisions from the prognostic model.

The presented experiment is a seminal study with encouraging results that still needs to be improved. The studied industrial dataset relies on a set of spindles, with every degradation class C_t being associated with one spindle only. It must be enriched with data from more spindles associated with the same classes C_t and spindles associated with other classes C_t closer to C_0 representing the end of life.

Future work will aim at reproducing this experimental setup for different types of spindles available at BOSCH-Rodez manufacturing site, to select relevant sensors and deploy them on the different fleets of machine-tools having multiple spindles. As soon as these sensors will be installed, data will be regularly recorded and stored in the database, data that can be also used to enrich the previous prognostic models. This sensor instrumentation will also allow the acquisition of run-to-failure data for these spindles, such as data from the PRONOSTIA platform (Nectoux et al., 2012). These data would be used to predict their RUL through a regression method instead of a classification of their degradation state C_t. A regression method could help to refine the accuracy of the predicted RUL to hours instead of months, and thus refine the predictive maintenance policy to be more accurate. Hence, the machine-tool production uptime could be maxi-

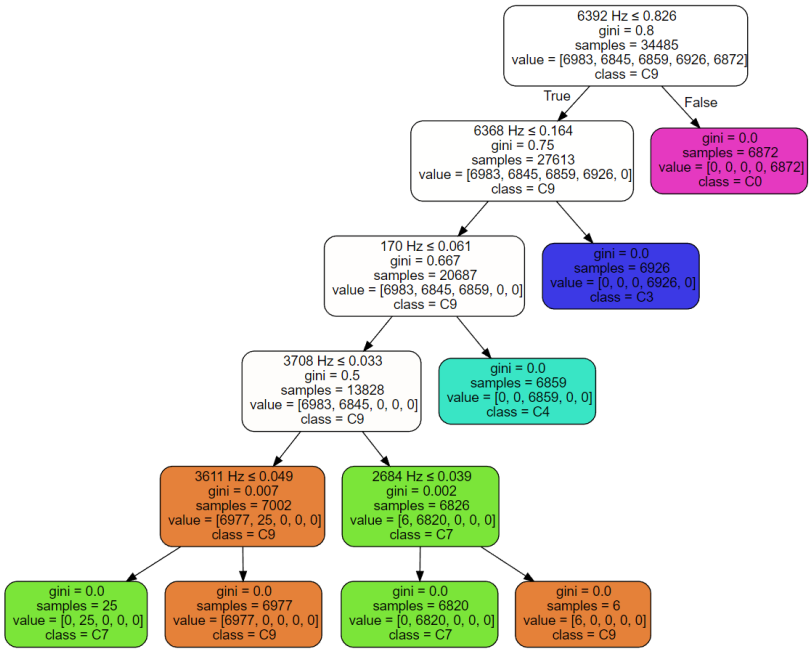

Figure 5. f_{DT}: Interpretable Prognostic Model as a Decision Tree

mized. As the data could become more complex, other interpretable methods like Generalized Additive Models (Lou, Caruana, & Gehrke, 2012) could be tested to ensure that the tradeoff between generalization and interpretability remains acceptable.

REFERENCES

Chelmiah, E. T., McLoone, V. I., & Kavanagh, D. F. (2022). Remaining useful life estimation of rotating machines through supervised learning with non-linear approaches. *Applied Sciences*, *12*(9).

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for multi-class classification: an overview.*

Kim, Y. S. (2008). Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. *Expert Systems with Applications*, *34*(2), 1227-1234.

Kingma, D., & Ba, J. (2014, 12). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining* (p. 150–158). New York, NY, USA: Association for Computing Machinery.

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. *CoRR*, *abs/1705.07874*.

Marcinkevics, R., & Vogt, J. (2020, 12). *Interpretability and explainability: A machine learning zoo mini-tour.*

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). Pronostia : An experimental platform for bearings accelerated degradation tests..

Rao, B. K. N., Pai, P. S., & Nagabhushana, T. N. (2012, may). Failure diagnosis and prognosis of rolling - element bearings using artificial neural networks: A critical overview. *Journal of Physics: Conference Series*, *364*(1), 012023. doi: 10.1088/1742-6596/364/1/012023

Riley, R. D., Debray, T. P., Moons, K. G., Schaar, M. v. d., & Hemingway, H. (2019, 01). 328Machine learning in prognosis research. In *Prognosis Research in Healthcare: Concepts, Methods, and Impact.* Oxford University Press.

Rudin, C. (2019, 05). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206-215.

Sassi, S., Badri, B., & Thomas, M. (2007). A numerical model to predict damaged bearing vibrations. *Journal of Vibration and Control*, *13*(11), 1603-1628.

Ying, X. (2019, feb). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022.