Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022)
Edited by Maria Chiara Leva, Edoardo Patelli, Luca Podofillini, and Simon Wilson
©2022 ESREL2022 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-5183-4_S29-03-225-cd

An Ensemble Learning Methodology for Predicting Medical Micro-robot Degradation Classes

Paul Cardenas-Lizana[1], Liseth Pasaguayo[2,3], Sergio Lescano[3] and Zeina Al Masry[2]

[1] *Universidad de Ingenieria y Tecnologia, Jr. Medrano Silva 165, Barranco , Peru.*
E-mail: pcardenas@utec.edu.pe
[2] *FEMTO-ST institute, Univ. Bourgogne Franche-Comté, CNRS,ENSMM, 24 rue Alain Savary, Besançon cedex, 25000, France. E-mail: lisethpasaguayoec@ieee.org and zeina.al.masry@ens2m.fr*
[3] *AMAROB Technologies, 25000 Besançon, France. E-mail: sergio.lescano@amarob.com*

The new generation of medical devices for surgical operation involves to develop equipment on the scale from micrometers to millimeters in order to perform more precise microsurgical procedures. Such kinds of devices must fulfill several tests according to medical standards to be commercialized. Hence, it is necessary to model the micro-robot degradation in order to ensure the optimal performance limits during surgical acts. The work aims to predict a micro-robot degradation class using a machine-learning-based methodology and consists of classifying the degradation into three classes: healthy, degraded, and out of service. Firstly, the degraded data are collected by using a four-bar complaint mechanism. This mechanism allows obtaining relevant attributes for the micro-robot degradation behavior. Secondly, a data preprocessing analysis and feature engineering are conducted to generate representative attributes that provide a better learning representation for the machine learning (ML) algorithm. Then, non-linear supervised learning algorithms are trained to construct the prediction. Random forest outperforms other algorithms in terms of predicting the remaining useful life (RUL) while gradient boosting generates the optimal decision boundary for classification using the RUL and features generated by autoencoders in presence of noise. Finally, a pipeline for the classification of the micro-robot degradation state is provided. This methodology ensures a procedure that evaluates whether or not the ML model can represent the underlying system in presence of noise.

Keywords: surgical micro-robots, degradation classes, Remaining useful life, ensemble learning, noisy features

1. Introduction

The new generation of medical devices for microsurgery procedures involves developing equipment on the scale from micrometers to millimeters in order to perform more precise surgical procedures with a minimal impact on the patient. Those procedures include small incisions and ablations of malignant tissues inside the human body. The enterprise Amarob has a miniaturized laser scalpel embedded in a micro-robot in charge of performing intracorporeal laser surgeries. The micro-robot must meet several regulations to ensure its safety and its consistent manufacturing. Thus, it must undergo several tests before its approval by the Food and Drug Administration (FDA) in the United States and by the Medical Devices Regulation (MDR) in Europe for commercialization. The regulations are primarily based on a risk analysis according to the standard ISO 14971 for medical devices (van Vroonhoven (2020)). Thus, the micro-robot must undergo different tests that ensure optimal performance during the surgical procedure. Considering standard ISO 14971, our work is focused on analyzing random faults as degradation that could impact the micro-robot safety. This is possible when the RUL is predicted with minimal error by implementing a prognostics health management (PHM) study.

Artificial intelligence techniques are currently applied for PHM, but there is still a long way for a precise RUL prediction. It is very important to achieve very small error and high precision, especially in critical biomedical systems, since this would significantly improve the reliability and operational safety of the micro-robot and prevent fatal breakdowns that may impact human health. Machine learning (ML) tools are used for developing high-precision predictive algorithms based on data. ML tools are capable of handling high-dimensional and multivariate data and learning the patterns. Their performance depends on the appropriate choice of the adapted technique and the quality of the data.

In Pasaguayo et al. (2021), the authors proposed some basic degradation models for lifetime estimation of the flexure hinges of a micro-robot dedicated to intracorporeal surgeries. Their work has opened a window for future studies to involve more advanced techniques in order to precisely estimate the RUL. Hence, this study aims to predict a micro-robot degradation classes using a machine-learning-based methodology. It consists of classifying the degradation state into three classes: healthy, degraded, and out of service. The thresholds of the classes are used to construct the decision boundary generated from the RUL. The methodology includes data preprocessing, model selection, training, and validation to obtain the best ML model. Supervised learning algorithms are used to construct the final model and to provide the classification of three health states of the micro-robot. The methodology ensures a procedure that evaluates whether or not the ML model can represent the underlying system. This is very important since the environment and settings of critical biomedical micro-robots may change over time, leading to the degradation of the model performance. The remainder of the paper is as follows: Section 2 describes the micro-robot, Section 3 presents the proposed methodology, Section 4 discussed the results, and finally, Section 5 gives the conclusion.

2. System description

The micro-robot used for this study was created and tested in laboratory conditions in FEMTO-ST Institute and is shown in Fig 1. This is encapsulated in a cube of 10x10x10 mm^3, it has a parallel kinematic structure composed of links made of carbon fiber, several flexure hinges made of polyimide, and in some parts are placed piezoelectric cantilevers actuators to generate displacements (Lescano (2015)). These flexure hinges are subjected to cyclical loads during their operation, thus these can cause complex deformations or failures that can affect the reliability of the micro-robot.

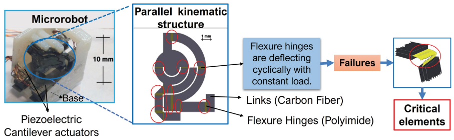
Microrobot
10 mm
Base
Piezoelectric
Cantilever actuators
Parallel kinematic structure
1 mm
Flexure hinges are deflecting cyclically with constant load.
Failures
Links (Carbon Fiber)
Flexure Hinges (Polyimide)
Critical elements

Fig. 1.: Micro-robot dedicated to intracorporeal surgeries, (a) micro-robot assembly with actuators, (b) parallel Kinematic structure made with SCM technique.

This study is focused on analyzing the behavior of these flexure hinges through data degradation extraction that provides information on the performance of the micro-robot. These flexure hinges are considered the critical elements in the micro-robot, thus in this work, they will be analyzed through machine learning techniques mainly to determine their RUL and define a boundary decision. A four-bar compliant mechanism was proposed (Pasaguayo et al. (2021)) to analyze the dynamic behavior of one flexure hinge and to collect data through simulation. This is because the parallel structure of the micro-robot integrates multiple flexure hinges and this makes complex the system to be analyzed.

3. Proposed methodology

The background of ensemble learning is first presented. Then, the RUL and degradation class prediction framework is provided.

3.1. *Machine learning models*

In this study, an ensemble learning pipeline is proposed to predict medical micro-robot degradation classes. Ensemble learning (Sollich and Krogh (1995)) combines a collection of learning algorithms to construct a predicting model by combining the strengths of a simpler base models. There are multiple forms to construct ensemble learning algorithms, and the most popular are random forests (Breiman (2001)) and boosting (Chen et al. (2015)). They are used as “out-the-box” learning algorithms that provide good predictive performance, and decision trees are commonly used as building blocks to construct and represent the model.

Random forests (RF) build a large and finite collection of uncorrelated trees and then averages their output as the prediction. Trees are ideal base learner candidates because if they are grown sufficiently deep, they have relatively low bias. RF uses bagging to average many noisy but unbiased tree models, and hence indirectly reduce the variance. For instance, each tree is grown in a bootstrapped dataset and before each split it selects a small set of the input attributes at random as candidates for splitting. Boosting is one of the most powerful learning procedures that combine the outputs of many "weak" classifiers to produce a powerful "committee". Weak learners have an error rate slightly better than a random one. At each boosting step, the data is modified by applying weights to each of the training instances where their weights are individually modified. Finally, all learners' outputs are combined through an average to produce the final prediction.

Autoencoders (AEs) are widely used for anomaly detection since they are good at learning to replicate the most frequently observed characteristics in the training data. The AE is composed of two main structures: an encoder and a decoder that are multilayer neural networks (NN). The first encodes the input data into a latent representation, while the second decodes this latent representation to an approximation or reconstruction of the original data. The variational autoencoder (VAE) has the same functions as the AE, since it is made up of an encoder and a decoder. VAE is a generative model that combines bayesian inference and neuronal network efficiency to obtain a low-dimensional nonlinear latent space.

3.2. *Remaining useful life prediction*

A RUL of a device is defined as the length of time left until the device is likely to operate before it requires repair or replacement, and it reaches its end of useful life (Wang et al. (2019)). The $RUL(t)$ is given by $T_{EoL} - t$, where T_{EoL} is the failure time to a predefined degradation threshold and t is the current time. If the RUL is known, maintenance can be scheduled in advance to optimize operating efficiency and reduce unplanned downtime. For this reason, estimating RUL is a top priority in predictive maintenance for critical systems and developing data driven methodologies for accurate RUL prediction is exceptionally beneficial for PHM and predictive maintenance.

3.3. *Degradation Classes*

There have been proposed multiple ML schemes for PHM (Biggio and Kastanis (2020)), however, today, ensemble learning such as RF and boosting are by far the most popular and used non-linear ML models for tabular datasets. A study found that RFs performed best on 121 UCI datasets against 179 other classifiers (Fernández-Delgado et al. (2014)) and a study found that RF and gradient boosting (GB) are the most widely used non-linear ML model right now (Kaggle (2021)).

In this study, ensemble learning is preferred as the main regressor and classifier over other non-linear techniques since they require less computing power, relatively fewer data compared to other algorithms such as neural networks, and are best at handling noise (Gupta and Gupta (2019)). Feature engineering (FE) could improve model performance, and it is a crucial step prior to training the ML model. The FE step extracts and creates attributes from raw data, since ML models do not create new features, and they only process the given ones. Since there is a myriad of ML models and data, the best features could be only defined in the context of the specific application, the model, and the data (Hui et al. (2017); Qiu and Zhang (2020)). The FE algorithms are the key to easing the learning process and enabling the model to predict better results. The RUL calculated from the historical data is required for determining the micro-robot degradation classes. Thus, we propose a methodology for creating the best features and for predicting the degradation classes based on the degradation data and RUL in the presence of noise.

The proposed pipeline is to use FE on the raw data to produce important attributes that best describe degradation. First, statistical learning is used to estimate the partial autocorrelation (PA) on the raw data and to determine the number of contiguous prior observations needed to accommodate the time series dependence in the model. Using PA, a rolling window of 10 lags is defined and applied on the raw data where all statistics are calculated inside the window to create new features such as mean, median, standard deviation (STD), minimum/maximum, skewness/kurtosis, 3-quantiles, and a cumulative sum. An exponentially weighted window is applied to raw data to calculate the mean, and STD inside the defined window. We define a relative strength index (RSI) inside the rolling window. Finally, an expanding window is applied on the raw data to create additional features such as mean, median, standard deviation (STD), minimum/maximum, skewness/kurtosis, 3-quantiles, and a cumulative sum. A total of 26 features are created.

Ensemble learning algorithms are chosen for building the RUL regressor for their performance when compared with other ML algorithms. The FE is directly coupled to the RUL regressor. A joint probability distribution of the features are constructed based on feature permutation and feature removal using RF, extremely randomized trees (ERT), and GB. Only 50% of the most important features are considered as core for a forward and backward wrapper-based selection. The wrapper algorithms are employed to end up with the best final attributes. The searching constrained for the maximum number of the best attributes is set to be no more than half the number of attributes. The wrapper algorithms for feature selection find the most relevant explanatory variables based on reducing the error both in training and testing. During this procedure, the best RUL regressor is found to be a RF with 9 attributes. A bias corrector is included for a RF after obtaining the best model and features, since RF only reduce the variance while leaving mostly unchanged the bias in regression problems (Zhang and Lu (2012)).

Two binary classifiers are generated based on the output of an AE and a VAE using an activation unit that detect whether there is an anomaly or not in the input vector. Both AE and VAE use the best 9 FE as their input vector and the threshold of the activating units are learned from data where two states are produced and correspond to normal (N) or anomaly (A). The EA and the VAE add a noise resistant anomaly detection method, and during training they are imposed to learn with more precision normal class. Finally, the feature engineered attributes, the RUL, and the output of the 2 binary classifier are stacked to train an ensemble learning classifier such as a RF and GB. The final classifier is selected for its capacity to predict the medical micro-robot degradation classes with high precision in the presence of noise. The methodology is illustrated in Figure 2. This architecture is composed of a feature engineering (FE) unit, a RUL predictor with a bias corrector (BC), a scaling unit (S), an AE and VAE with their respective activation units, and a final classifier.

The raw dataset includes 81 degradation trajectories of the micro-robot and each trajectory is composed approximately of 10K instances. The total instances are then classified as "degraded", "healthy", or "out of service" class based on the manufacturer's specifications (Table1) and using the RUL. This dataset is composed by 81K instances where 11% corresponds to the "degraded" class, while the "healthy" class and the "out of service" class are 57% and 32%, respectively. The dataset is then split in a stratified manner according to the classes into training and test set using the 80/20 rule for the learning process. The training set is used with cross-validation to learn the parameters and for model selection. The test set is used for measuring the generalization error of the final chosen model. The root-mean-square error (RMSE) is used to measure the performance of the RUL regressor and is defined as the differences between the ground truth and the predicted values. The confusion matrix (CM) is used to examine the performance of a multinomial classifier. The

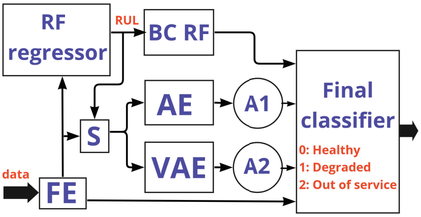

Fig. 2.: Proposed methodology for RUL estimation and the classification of the micro-robot degradation state.

elements (CM_{ij}) of a CM represent the number of observations known to be in group i and predicted to be in group j, and they are normalized with respect to the total ground truth of each class. The results are presented using k-fold cross-validation ($k = 5$) for model training and model selection. To assess the generalization error of the final model in the presence of noise, 50 noisy test set are constructed using the test set and by adding a Gaussian noise with specific noise level from 10% to 50% and their results are reported. Since CFs have a STD less than 0.01%, they are not shown in the CF tables.

4. Experiment and results analysis

In this section, the generation of the dataset is described, and the results are discussed.

4.1. *Generation of the degradation dataset*

In order to generate the degradation data of the flexure hinge, the bending motion of the hinge is simulated using SolidWorks®, and its 2D motion is recorded during an operating cycle. This trajectory is used to tracks the degradation of the flexure hinge over time. 81 samples are simulated under different conditions to characterize the degradation profile. The torsion spring constant is varied in each simulation to mimic wear down and provide different mechanical behavior of the mechanism. The length of the flexure hinge can change from one to many order-of-magnitude (10^3) due to overuse; thus modifying the torsion spring constant as this is given by $K = EI/L$, where I is the moment of inertia of the cross-section of the flexible segment, E is its modulus of elasticity, and L is its length (Ugwuoke (2008)).

Table 1.: Micro-robot Degradation Classes given by manufacturer.

Class	RUL
0: Healthy	$200 < \text{RUL}$
1: Degraded	$80 < \text{RUL} \leq 200$
2: Out of service	$\text{RUL} \leq 80$

This parameter best represent the degradation fo the hinge and is varied in the simulation to collect 81 different trajectories. Each sample is simulated for a total of 1000 seconds and every second a value for the x and y coordinates is recorded.

For our study, it is important to note that although learning algorithms can be very precise, they are currently not used in critical systems due to errors that come from poor generalization and caused by overfitting or by the existence of attack systems. Thereby, this study proposes a robust and noise-insensitive system to predict the degradation classes of medical micro-robots using mainly features created from the RUL and the boundary decision defined by the RUL. It is expected that the proposed system is going to be used during surgical operations of the micro-robot, therefore it must be very robust and insensitive to noise so as not to endanger the life of the patient. The raw dataset includes 81 degradation trajectories, and the training dataset is composed by approximately by 81K instances. An example of the three classes is shown in Fig. 3. The classes are constructed by the manufacturer's technical specifications, detailed in Table 1.

4.2. *Results*

The prediction of the degraded class defined between the transition from the healthy to the out of service class is difficult and in many cases are not very precise due to the lack of data describing this state. Although the data is composed of approximately 81K observations, it is very unbalanced and does not allow the algorithm to extract and learn the pattern for this class. Importantly, creating a dataset of complete and balance degradation signals can be very costly and time-consuming, which is why developing highly accurate predictive systems is critical. The best features are obtained by the proposed technique using the complete attributes created by the FE. The wrapper method is coupled with the RF to produce 9 best features. This scheme gives excellent results by decreasing the test error (Fig.4 a) and improving the R^2 (Fig.4 b) in each iteration of the feature selection process. Fig.4 (a and b) shows error of the last 4 searching iteration using

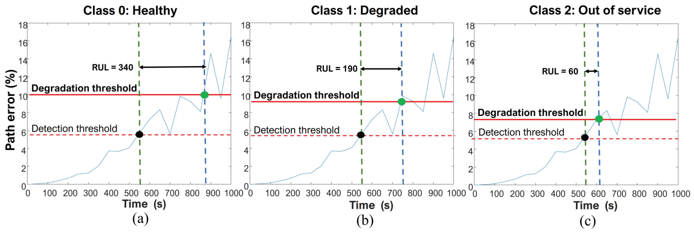

Fig. 3.: Degradation classes, (a) Healthy degradation class (b), degraded degradation class and (c) out of service degradation class.

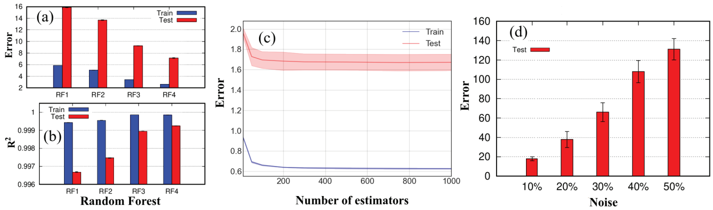

Fig. 4.: RUL prediction in the absence and presence of noise:(a) Error and (b) R^2 of the optimization process of the RF during feature selection based on instance permutation and removal. (C) The best RF regressor obtained with a test error of 1.775 cycles. (d) RUL error predictions in presence of noise.

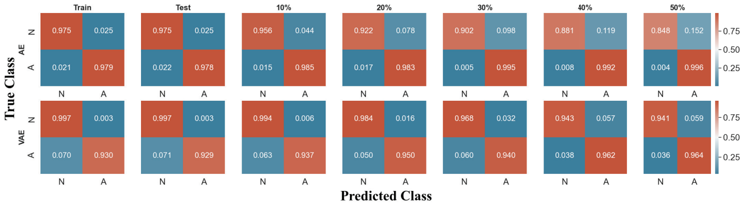

Fig. 5.: Classification of the AE and VAE outputs in two states in the absence and presence of noise that correspond to normal output (N) or anomaly output (A) defined by the activation units.

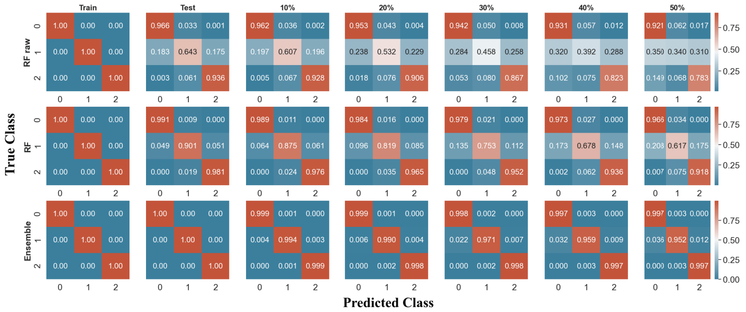

Fig. 6.: Final classifier results, (a) RF classifier trained on raw data, (b) RF classifier trained on 9 best features, (c) proposed ensemble classifying system.

RF. The best RUL regressor (RF4) is obtained by optimization using cross-validation and it finally achieves a training error of 0.7 cycles and a test error of 1.775 cycles (Fig.4 c). The RF4 has a total of 750 estimators and it is obtained including all features during each split using a mean-squared-error loss function, all trees are grown completely without pruning, and each node contains at least one sample.

The degradation profiles are obtained noise-free, however sensor acquisition is inherently prone to errors (Zhu and Wu (2004)) due to instrumentation and environmental problems during readout. To model the noise, a Gaussian distribution with mean of zero and specific standard deviation is multiplied on the data. This signal is then added to the data to mimic noise during sensing. Fig.4 (d) shows the RUL error as a function of noise power, and it is observed that the RUL prediction continuously deteriorates as the presence of noise is increased. Importantly, the RUL is primarily used to define the decision boundary. Thus, RUL alone could not be used as a predictor of micro-robot degradation classes because it is highly sensitive to the presence of noise. To overcome this problem, the ensemble system is proposed, and it includes the initial FE unit, a RF regressor with a linear bias corrector (BC) unit, the additional features created by the AE and the VAE with their respective activation (A) units. The activation thresholds are learned by imposing a bias to detect the anomaly input vector generated by the FE including the RUL. The features are standardized by removing the mean and scaling to unit variance in the scaling (S) unit prior to entering the AE and the VAE. The AE architecture is composed of an encoder and a decoder, where its structure is the mirror image of the encoder. Each structure is composed of two fully-connected layers of 10 and 8 nodes, respectively, using a code size of 4 nodes. This results in a total of 254 parameters. Similarly, the VAE architecture includes two parts, an encoder and a decoder where each is composed of two fully connected layers, and the layers have 10 and 8 nodes, respectively. The latent space is represented by two layers of 5 nodes for the encoder (the mean and the standard deviation layers) and one sampling layer of 5 nodes for the decoder, resulting in total 350 parameters. The Fig.5 shows that AE has a better precision than the VAE to detect anomaly both in training and testing in noise-free data. However, when the noise level is high the AE tend to classify mostly anomaly class deteriorating its precision on normal output (84.8%). Conversely, the VAE is resistant to noise, maintaining a good precision (94.41%) even when noise is very high (50%). The AE and VAE add a denoising function to the ensemble system.

All features are then stacked to train a RF and extreme GB (XGBoost) classifier as shown in Fig.2. The Fig.6 (a) shows a RF classifier trained directly on the raw data without FE. Although it has a good performance on training, the precision of the degraded class is poor (64.3%) and in the presence of noise it is unable to classify the degraded (34%) and out of service (78.3%) classes. The Fig.6 (a) shows a RF classifier trained on the best FE and while it has a good performance on healthy (99.1%) and degraded (98.1%) classes, it is not effective classifying the out of service (90.1%) class. This classifier completely degrades detecting the out of service class in the presence of noise going from 90.1% to 61.7%.

The proposed ensemble system (Fig.6 (c)) uses a XGBoost with a DART booster, a soft-max loss function, and a learning rate of 0.08. The trees have a maximum depth of 5 and a regularization is conducted by dropping out the tress where all trees are equally selected and have a dropout rate of 0.2 with a 50% probability of not performing the dropout at all. During learning, new trees are equally weighted as the dropped trees. These parameters are obtained by searching through the hyper-parameter space using Bayesian optimization (Hyperopt) to find the best possible values that minimized the loss function. This system has a perfect precision in both training and testing, and even when noise is very high (50%), its precision is still 99% for healthy and degraded, and 95% for out of service class. The system is capable of handling the noisy attributes during sensors readout. The results show that the combination of the FE output, RUL regressor and the AE and VAE outputs into the XGBoost classifier, the system becomes very robust and noise insensitive and provides an excellent precision for the degraded class.

5. Conclusion

This study shows that classifying the degraded class using directly the RUL is not appropriate and in the presence of noise the error increased significantly. To overcome this, a new approach is proposed for predicting the degradation classes of a micro-robot based on ensemble learning combined with AEs. Although the proposed methodology is applied to a single degradation signal, it could be generalized to multivariable sensor signals. The proposed FE algorithm creates the vector attributes using the degradation profile for training the RF, AE, and VAE and their outputs are used to train the final XGBoost classifier. To highlight the advantage of the proposed methodology, a high level of noise is added to the degradation signal before the FE, and it is observed that the trained scheme is very resilient to the presence of noise. This study also explore the possibility of generating binary input features based on autoencoders. Future work will explore the potential of using nonlinear units for both the bias-corrected and the activation units, and include multivariable sensor signals.

Acknowledgement

This work has been supported by the EIPHI Graduate school (contract "ANR-17-EURE-0002"), the Centro Bio and the Office of Research at UTEC, and the ANRT-AMAROB CIFRE Thesis. This work used the UTEC Khipu supercluster, the authors are thankful for the support.

References

Biggio, L. and I. Kastanis (2020). Prognostics and health management of industrial assets: Current progress and road ahead. *Frontiers in Artificial Intelligence 3*.

Breiman, L. (2001). *Machine Learning 45*(1), 5–32.

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2 1*(4), 1–4.

Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim (2014, jan). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res. 15*(1), 3133–3181.

Gupta, S. and A. Gupta (2019). Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science 161*, 466–474.

Hui, K. H., C. S. Ooi, M. H. Lim, M. S. Leong, and S. M. Al-Obaidi (2017, 12). An improved wrapper-based feature selection method for machinery fault diagnosis. *PLOS ONE 12*(12), 1–10.

Kaggle (2021, October). State of machine learning and data science 2021.

Lescano, S. (2015, nov). *Design, Fabrication and Control of a Microrobot for Laser Phonomicrosurgery*. Ph. D. thesis, Université de Franche-Comté.

Pasaguayo, L., Z. A. Masry, and S. Lescano (2021, 9). Degradation modeling analysis for microrobots flexure hinges using intracorporeal surgeries. pp. 1272–1279.

Qiu, Y. and C. Zhang (2020). Wrapper feature selection algorithm for the optimization of an indicator system of patent value assessment.

Sollich, P. and A. Krogh (1995). Learning with ensembles: How overfitting can be useful. In D. Touretzky, M. Mozer, and M. Hasselmo (Eds.), *Advances in Neural Information Processing Systems*, Volume 8. MIT Press.

Ugwuoke, I. (2008, 12). A simplified dynamic model for constant-force compression spring. *Leonardo Journal of Sciences 7*.

van Vroonhoven, J. (2020). *Risk management for medical devices and the new BS EN ISO 14971*. ASM International.

Wang, Q., S. Zheng, A. Farahat, S. Serita, and C. Gupta (2019). Remaining useful life estimation using functional data analysis. In *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pp. 1–8.

Zhang, G. and Y. Lu (2012, January). Bias-corrected random forests in regression. *Journal of Applied Statistics 39*(1), 151–160.

Zhu, X. and X. Wu (2004, November). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review 22*(3), 177–210.