Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 00 (2023) 000–000

www.elsevier.com/locate/procedia

56th CIRP Conference on Manufacturing Systems, CIRP CMS '23, South Africa

Sample size prediction for anomaly detection in locks

Tim Andersson[a]*, Mats Ahlskog[b], Tomas Olsson[c], Markus Bohlin[d]

[abd]*Mälardalen University Eskilstuna 63105, Sweden*
[c] *RISE Research Institutes of Sweden Västerås 72212, Sweden*

* Corresponding author. Tel.: +46 707-787-2382, *E-mail address:* Tim.Andersson@mdu.se

Abstract

Artificial intelligence in manufacturing systems is currently most used for quality control and predictive maintenance. In the lock industry, quality control of final assembled cylinder lock is still done by hand, wearing out the operators' wrists and introducing subjectivity which negatively affects reliability. Studies have shown that quality control can be automated using machine-learning to analyse torque measurements from the locks. The resulting performance of the approach depends on the dimensionality and size of the training dataset but unfortunately, the process of gathering data can be expensive so the amount collected data should therefore be minimized with respect to an acceptable performance measure. The dimensionality can be reduced with a method called Principal Component Analysis and the training dataset size can be estimated by repeated testing of the algorithms with smaller datasets of different sizes, which then can be used to extrapolate the expected performance for larger datasets. The purpose of this study is to evaluate the state-of-the-art methods to predict and minimize the needed sample size for commonly used machine-learning algorithms to reach an acceptable anomaly detection accuracy using torque measurements from locks. The results show that the learning curve with the best fit to the training data does not always give the best predictions. Instead, performance depends on the amount of data used to create the curve and the particular machine-learning algorithm used. Overall, the exponential and power-law functions gave the most reliable predictions and the use of principal component analysis greatly reduced the learning effort for the machine-learning algorithms. With torque measurements from 50-150 locks, we predicted a detection accuracy of over 95% while the current method of using the human tactile sense gives only 16% accuracy.

Keywords: Anomaly detection; Sample size prediction; Learning curves; Machine learning; Quality control

1. Introduction

Currently, industry 4.0 is emerging where most of the systems will be automated, interconnected and self-aware such that decision-making and optimization can be done in a decentralized fashion [1]. This has led to an increased interest in Machine Learning (ML) in fields like Intelligent Fault Diagnosis [2,3] where ML is used to predict maintenance needs and classify faults based on sensor data. In the context of ML, the amount of data, i.e., sample size, used for training and validation of the algorithms is one of the most important factors to consider to get a representative sample of the population but, also to give the ML algorithm enough observations to detect the underlying patterns in the data [4]. This becomes even more important for high-dimensional datasets since the search space increases exponentially with the number of dimensions i.e., the curse of dimensionality [5,6]. From a statistical point of view, larger sample sizes are preferred [4] but, the process of gathering data is often a time-consuming task that can affect the productivity, and inherently the economy, of an organization and also increase the training time for ML algorithms. The sample size should therefore be kept to a

minimum with respect to the desired predictive power. There are methods that can be used to create synthetic data based on the already gathered data e.g., SMOTE [7] and GANS [8]. Synthetic data can be useful to balance a dataset to avoid the ML algorithm getting biased towards a specific class. It might be tempting to use a large fraction of synthetic data to save time and resources but, in some cases, synthetic data can make the precision and accuracy of the ML algorithm worse which is why it is recommended to use real data as much as possible [9].

How large a sample should be is unfortunately not a simple question to answer. It is possible to use statistical power analysis methods to determine the minimum sample size for hypothesis testing [10]. However, these methods are based on certain assumptions about the data which might not be correct, and they do not take into consideration the ML algorithms' ability to learn i.e., Learning Curve (LC) [11]. The LC shows the performance of an ML algorithm's ability to predict a certain outcome on a given domain as a function of varying learning efforts where the most common measure is prediction accuracy as a function of varying sample size for the training dataset. All ML algorithms have their own unique ability to learn with a characteristic LC that is different for every application domain.

In the lock industry, quality control is done by hand using the human tactile sense, but a recent study has shown that an automated solution using ML and a torque sensor can be more dependable [12]. In that study, nine operators independently evaluate the same collection of faulty locks (low level of lubrication) and label each lock based on the quality. Only 16% of all the labels were correct and the labels from different operators were often contradictive which indicates that the operators cannot detect this type of fault reliably.

The purpose of this study is to evaluate the state-of-the-art methods to predict and minimize the needed sample size for commonly used machine-learning algorithms to reach an acceptable anomaly detection accuracy using torque measurements from locks. The type of lock used in this study can be seen in Fig. 1. The main research questions are:

- How will the sample size and dimensionality affect the classification accuracy?
- How accurately can the accuracy be predicted for a specific sample size using learning curves?

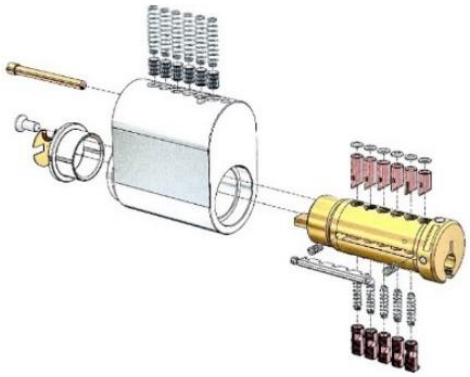

Fig. 1. Exploded model of the lock used in this study [11].

2. Literature review

In intelligent fault diagnostic there are two types of ML approaches that are commonly used called supervised- and unsupervised- learning [2,3]. In the former method, a function $f: X \rightarrow Y$ is learned such that the input features $x \in X$ can be mapped by f to the output labels $y \in Y$. In contrast, the latter method does not use corresponding labels but instead tries to make sense of the data based only on the input features [13].

In practice, it is often easier and cheaper to gather data related to a normal condition of a system, which is where an unsupervised learner can be used to detect anomalies [14]. However, there are also supervised learners that can be used, called one-class learners, which detect anomalies based on the assumption that all the provided input data are from the same class [15,16]. The drawback of one-class learners and unsupervised learners is that they generally are more sensitive to the curse of dimensionality [16], which means that they require a larger training dataset compared to the regular classifiers. Within intelligent fault diagnostics, a set of ML algorithms was selected to be used in this study namely One-Class Support Vector Machine (OCSVM), K-Nearest Neighbour search (KNN search) and Isolation Forest (IF) [2,17,18]. These ML algorithms' goal is to differentiate the approximated fraction of unknown outliers in the training data which in this study is set to 5% of the normal class. OCSVM's objective is to define a hyperplane that creates a decision boundary between the outliers and the rest of the training data. IF creates an ensemble of decision trees to isolate the outliers and KNN search classifies observations as outliers if the distance to their K_{th} nearest neighbour is greater than a specific threshold (95th percentile's distance based on the training data in this study).

When working with high-dimensional real datasets most of the discriminant information is frequently found in lower dimensions, these are called the intrinsic dimensions [19]. In the case where the dimensionality is higher than the number of observations in the dataset, the curse of dimensionality can affect the performance of an ML algorithm negatively. The most common approach to solve this is to find an approximation of the intrinsic dimensions by doing a linear and orthogonal projection of the data down to a lower dimension such that the variance in the data is maximized, this method is called Principal Component Analysis (PCA) [17] and is the method used in this study.

When using LCs for sample size prediction the mathematical function that defines the shape of the curve needs to be decided initially. In [20], a literature review has been made regarding the shape of learning curves where they concluded that power law and exponential functions have the most theoretical and empirical evidence support but, they also found studies where logarithmic functions were used or combinations. In this study, logarithmic, exponential and power law functions were evaluated.

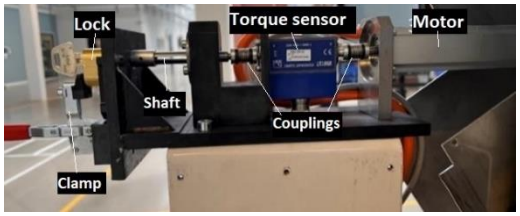

Fig. 2 The experimental setup used to measure the needed torque to turn the lock.

3. Method

To perform the data collection, a focus group consisting of three domain experts in testing and manufacturing of these locks was used to agree upon a set of 346 unused fully functional locks as a representative sample. This dataset was labelled OK. To generate an anomalous group of locks, the same set of 346 locks was manipulated to induce mechanical faults. Specifically, the lubrication was removed from 310 locks using alcohol. The lack of lubrication is known to give minor mechanical anomalies which are usually not detected by human operators [12] and is, therefore, a suitable test case. Additionally, 36 locks were also intentionally contaminated with sand to create major anomalies, which are not common, but it is known to have happened previously in the factory during the assembly process. The anomalous dataset was labelled NOK. The data collection was done by rotating the locks with a constant angular velocity of 30 degrees per second with an electric motor, see Fig. 2. The sensor [21] used to measure the applied torque to turn the lock has a range of 0.1-200Nm, a sensitivity of 0.02Nm and the computer used to sample the data had a sample rate of 1ms. The measurements are mapped to a specific angular position with a resolution of 0.1 degrees resulting in 3600 measurements for a 360-degree turn for each lock, resulting in two datasets with size 346x3600, i.e., 346 instances and 3600 features for each class.

3.1. Data preprocessing

When performing the rotational movement of the lock, the needed torque to keep a constant angular velocity varies during a full turn since the internal contact surfaces are not identical for each angular position. This can result in sudden changes in the angular position of the lock due to a mechanical spring-effect and play in the measuring equipment. The computer that samples the sensor data was unable to keep up with these sudden changes and therefore missed 0-5% of the measurements for some angular positions. This was solved using linear interpolation with the two neighbouring data points [22].

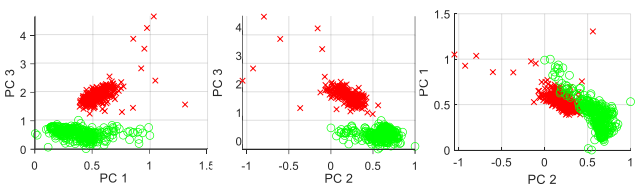

Fig. 3. This is the reduced data using 3 principal components (PC). Green is the OK class and red is the NOK class.

To get an equal impact from each feature in the data, a common range should be used for each feature in the dataset. This was achieved by normalizing the training dataset, consisting only of OK locks, such that each feature has a range of 0-1. The same derived parameters to normalize the training dataset was also used to normalize the test data, consisting of both OK and NOK locks, which is why most of the NOK locks' features are outside the range of 0-1. The normalization was done both before and after the dimensionality reduction. The dimensionality reduction was done using PCA to minimize the needed sample size. The resulting coefficient matrix used to project the data to the lower dimensions was derived from the training dataset and then applied to both the training and test data. An initial test using all the available data revealed that keeping only 3 dimensions after PCA separates the two classes relatively well (see Fig. 3).

3.2. Training and validation of ML algorithms and LCs

The training and evaluation were done by using a Monte Carlo validation approach [23] where x number of instances, not more than 80%, of the OK class, was selected for the training dataset and the remaining OK and NOK instances were used in the test dataset. To avoid getting biased test results in favour of a specific class, the test set was balanced by randomly discarding instances from the majority class (NOK).

This process was repeated one hundred times for each sample size x to evaluate the current ML algorithm's accuracy. The accuracy is expressed in terms of classification loss i.e., the percentage of wrong classifications. These results were then used to calculate one hundred bootstrapped [24] mean accuracies to better approximate the mean accuracy and its variance. The observed variance from the one hundred repeated evaluations and the bootstrapped mean variances were added together in accordance with the variance sum law [25] to calculate the z-score 95% confidence interval [26]. The training sample size x started at three and was then increased with two until 80% (275 instances) in the OK class was used for training.

From the process described above, a total of 137 data points representing the mean accuracy for different training sample sizes were used to approximate different LCs. To fit the LCs, a robust non-linear least square regression algorithm was used with a 95% confidence interval [27]. It works by adjusting the weights of the data points based on the residuals from the unweighted fitted curve and then fits a new curve with these weights. This minimizes the risk of outliers affecting the results. The curve fitting was done iteratively where the number of data points used increased with one for each iteration, until 80% of the data points were used, starting at four and the remaining data points were used as a test set. The LCs were evaluated based on their plotted Mean Absolute Error (MAE) on the training data (interpolation) and test data (extrapolation). An evaluation was also done using all 137 data points to get the best approximation of the true LC and its simultaneous observational confidence interval bounds [28].

The default parameter settings for the ML algorithms were used in this study [29], except for the IF where the number of

trees was changed from 100 to 20 to save processing time. These settings were initially tested to verify that no optimization was needed to achieve a sufficient classification loss of less than 5–10% using 80% of the OK class for training.

4. Results and discussion

The results from each ML algorithm are organized in Fig. 4. In the first row of Fig. 4, the LCs trained with all the 137 data points can be seen which shows that in most of the cases, an exponential function gave the lowest MAE. In [20], they mention that exponential LCs are common when the classes are well separated which they were in this study. However, the best-fitted curve on the training data, see row 2 in Fig. 4, does not necessarily mean that it will be the best choice to predict future values, see row 3 in Fig. 4. For small sample sizes (less than 25–50) the MAE for the predicted values is large compared to the corresponding values for the training data. A drastic decrease in MAE for the predicted values can however be observed when the LC starts to converge, i.e., where the rate of change of the classification loss for an increase of the training data, also called the learning rate, rapidly decreases e.g., at sample size 25–50 for KNN (row 3 in Fig. 4). This shows that the learning rate can be a suitable indication at which point one can use LCs to reliably predict future performance. In the cases where IF and OCSVM were evaluated, it can clearly be seen that PCA reduced the learning effort and classification loss but it had almost no effect on KNN search. All of the ML algorithms managed to reach about 2.5–5% classification loss and since the two classes are completely separated, see Fig. 3., it might be possible to reach 0% classification loss with proper tuning of the ML parameters.

5. Conclusions

In this study, we have evaluated several different mathematical functions commonly used to approximate the shape of learning curves (LC) for three different commonly used machine learning (ML) outlier detection algorithms called One-Class Support Vector Machine (OCSVM), K-Nearest Neighbour search (KNN search) and Isolation Forest (IF). The ML algorithms' task was to detect various anomalies related to the lack of lubrication and contaminations inside locks based on the needed torque to operate them. The fitted LCs were used to predict the future performance of the ML algorithms for different training sample sizes with and without a dimensionality reduction technique called principal component analysis (PCA) which resulted in a total of six different datasets to train the LCs with, i.e., two datasets from each ML algorithm. Based on the results, it can be concluded that the noise in the data points used to fit the LCs adds uncertainty to the fit, such that small sample sizes, less than 25-50 locks (13-25 data points), will not give any better prediction than approximately 5% Mean Absolute Error (MAE). But, with a sufficiently large sample size such that a distinct curvature of the LC can be captured i.e., where the learning rate starts to decrease drastically, 50–150 instances (25-75 data points) in this case, the MAE decreases to 1-2%. The best approximations of the true LCs, using a sample of 275 instances (137 data points), were an exponential and power law functions with 1-

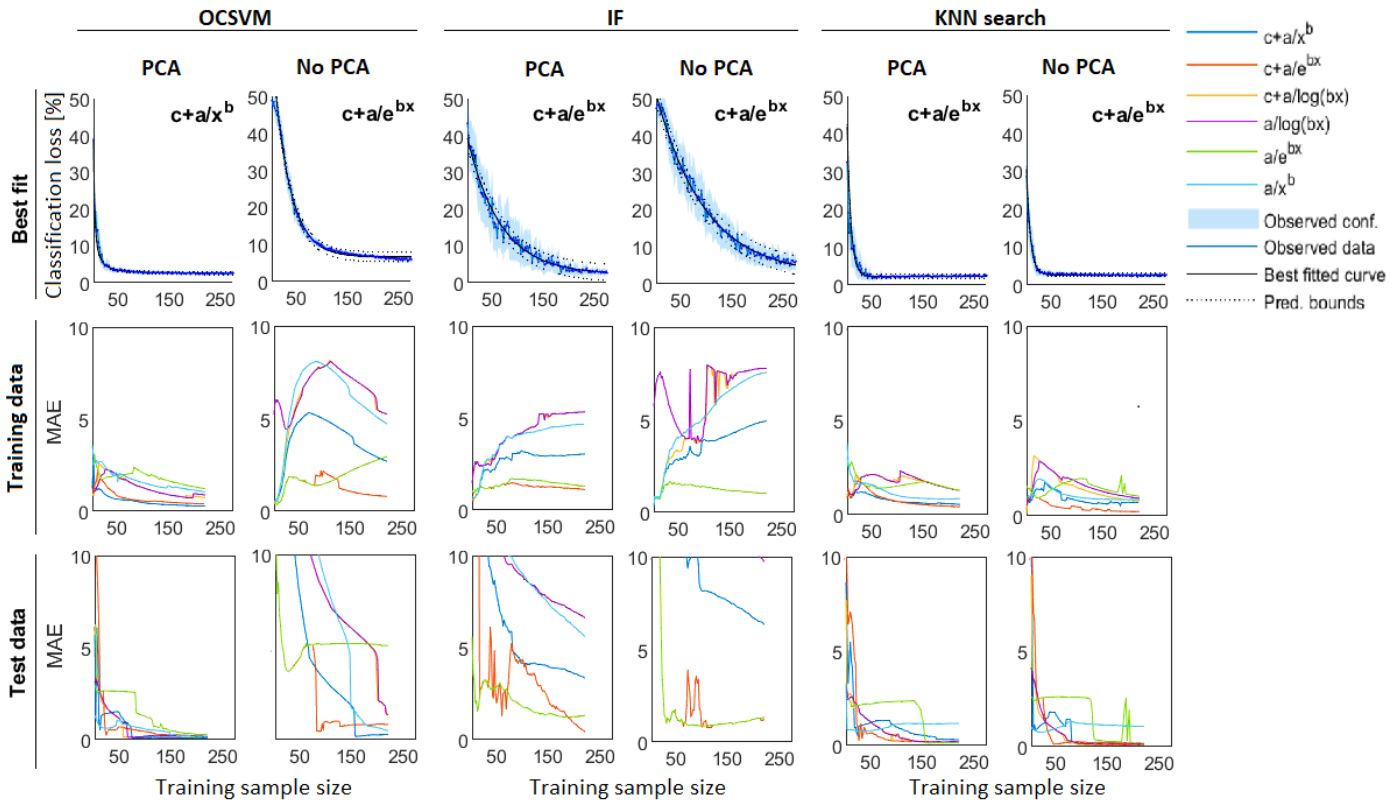

Fig. 4. First column is the results with PCA and the second column is without PCA for each corresponding ML algorithm. The first row shows the best fitted LC using all data, the second row shows the MAE for the LCs on the training data in relation to varying sample sizes of the training data and the third row is the MAE for the LCs on the test data in relation to varying sample size of the training data.

2% MAE. Using PCA, we successfully demonstrated how the needed sample size can be reduced to reach a certain anomaly detection accuracy by more than 50% and how the maximum accuracy can be improved depending on the ML algorithm. The ML algorithms reached 2.5-5% classification loss, using a sample of 50-150 locks, outperforming the current method of using the human tactile sense which typically cannot detect anomalies resulting from the lack of lubrication.

The result of this study can be used in the deployment of an automated solution for quality control in the lock industry. It also is one of few studies where torque measurements are used to detect mechanical anomalies of a rotating mechanical product and can therefore act as a framework for how it can be achieved and/or as a point of departure in similar applications.

6. Limitations and future research

The ML algorithms' parameters and the pre-processing techniques for the data were not optimized during the training to be able to reduce processing time. This can be one of the reasons why the data points have a noisy behaviour between each sample size's classification loss. Also, without optimized parameters, no conclusions can be made regarding which algorithm is most suitable for anomaly detection of locks even though KNN search seems promising. All the LCs have not yet converged within the limited sample size of 275 locks, this can affect certainty in the approximation of the true shape of the LC. For future research, optimized parameters should be considered for the ML algorithms and also compare different fitting techniques for learning curves.

Acknowledgement

This project has received funding from The Knowledge Foundation, Mälardalen University and Assa Abloy under grant agreement No 20200132 01 H.

References

[1] Dossou P-E, Torregrossa P, Martinez T. Industry 4.0 concepts and lean manufacturing implementation for optimizing a company logistics flows. Procedia Computer Science 2022;200:358–67.
[2] Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK. Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical and System Signal Processing 2020;138.
[3] Carvalho TP, Soares FAAMN, Vita R, Francisco R da P, Basto JP, Alcalá SGS. A systematic literature review of machine learning methods applied to predictive maintenance. Computer & Industrial Engineering 2019;137.
[4] Asiamah N, Kofi Mensah H, Fosu Oteng-Abayie E. Do Larger Samples Really Lead to More Precise Estimates? A Simulation Study. American Journal of Educational Research 2017;5:9–17.
[5] Murphy KP. The curse of dimensionality. In: Dietterich T, editor. Probabilistic Machine Learning: An Introduction, MIT Press; 2022, p. 544–6.
[6] Alkhudaydi MH. Learning and Generalisation for High-dimensional Data 2021.
[7] Chawla N v., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2002;16.
[8] Vega-Márquez B, Rubio-Escudero C, Riquelme JC, Nepomuceno-Chamorro I. Creation of Synthetic Data with Conditional Generative Adversarial Networks. 14th International Workshop on Soft Computing Models in Industrial and Environmental Applications, Springer, Cham; 2020, p. 231–40.
[9] Juba B, Le HS. Precision-Recall versus Accuracy and the Role of Large Data Sets. Proceedings of the AAAI Conference on Artificial Intelligence 2019;33.
[10] Shmueli G. To Explain or to Predict? Statistical Science 2010;25.
[11] Webb GI, Sammut C, Perlich C, Horváth T, Wrobel S, Korb KB, et al. Learning Curves in Machine Learning. Encyclopedia of Machine Learning, Boston, MA: Springer US; 2011, p. 577–80.
[12] Andersson T, Bohlin M, Olsson T, Ahlskog M. Comparison of Machine Learning's- and Humans'- Ability to Consistently Classify Anomalies in Cylinder Locks. Advances in Production Management Systems. Smart Manufacturing and Logistics Systems: Turning Ideas into Action, 2022, p. 27–34.
[13] Russell S, Norvig P. Forms of Learning. Artificial Intelligence A Modern Approach, 3rd ed., New Jersey: Prentice Hall; 2010, p. 693–5.
[14] Sgueglia A, di Sorbo A, Visaggio CA, Canfora G. A systematic literature review of IoT time series anomaly detection solutions. Future Generation Computer Systems 2022;134.
[15] Perera P, Oza P, Patel VM. One-Class Classification: A Survey. ArXiv 2021;abs/2101.03064.
[16] Johannes DavidM. One-class classification Concept-learning in the absence of counter-examples. PhD dissertation. Technische Universiteit Delft, 2001.
[17] Nassif AB, Talib MA, Nasir Q, Dakalbab FM. Machine Learning for Anomaly Detection: A Systematic Review. IEEE Access 2021;9:78658–700.
[18] Carvalho TP, Soares FAAMN, Vita R, Francisco R da P, Basto JP, Alcalá SGS. A systematic literature review of machine learning methods applied to predictive maintenance. Computer & Industrial Engineering 2019;137.
[19] Murphy KP. The manifold hypothesis. Probabilistic Machine Learning An Introduction, MIT Press; 2022, p. 686.
[20] Viering T, Loog M. The Shape of Learning Curves: a Review. Journal IEEE Transactions on Pattern Analysis and Machine Intelligence 2022.
[21] Hottinger Baldwin Messtechnik. T21WN-Data Sheet. A4776-1.0 ed. Hottinger Baldwin Messtechnik; 2017.
[22] Caruso C, Quarta F. Interpolation methods comparison. Computers & Mathematics with Applications 1998;35:109–26.
[23] Xu QS, Liang YZ. Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems 2001;56:1–11.
[24] Murphy KP. Bootstrap approximation of the sampling distribution of any estimator. Probabilistic Machine Learning: An Introduction, London: MIT Press; 2022, p. 154–5.
[25] Caron P-O, Lemardelet L. The variance sum law and its implications for modelling. The Quantitative Methods for Psychology 2021;17:80–6.
[26] Moore DS. Confidence Intervals: The Basics. In: Capuano L, editor. The Basic Practice of Statistics, 4th ed., New York: Craig Bleyer; 2007, p. 343–55.
[27] Fit Linear Interpolant Models Using the fit Function. MATLAB Curve Fitting Toolbox User's Guide, ver. 3.8, MathWorks, Inc.; 2022, p. 6–7.
[28] Compute Prediction Intervals. MATLAB Curve Fitting Toolbox User's Guide, ver. 3.8, MathWorks, Inc; 2022, p. 55–6.
[29] MATLAB Statistics and Machine Learning Toolbox. ver. 12.4. MathWorks, Inc.; 2022.