Article

A Combined Anomaly and Trend Detection System for Industrial Robot Gear Condition Monitoring

Corbinian Nentwich *[iD] and Gunther Reinhart

Institute for Machine Tools and Industrial Management, Technical University Munich, 85747 Garching, Germany; emeritus.reinhart@tum.de
* Correspondence: corbinian.nentwich@iwb.tum.de

Abstract: Conditions monitoring of industrial robot gears has the potential to increase the productivity of highly automated production systems. The huge amount of health indicators needed to monitor multiple gears of multiple robots requires an automated system for anomaly and trend detection. In this publication, such a system is presented and suitable anomaly detection and trend detection methods for the system are selected based on synthetic and real world industrial application data. A statistical test, namely the Cox-Stuart test, appears to be the most suitable approach for trend detection and the local outlier factor algorithm or the long short-term neural network performs best for anomaly detection in the application of industrial robot gear condition monitoring in the presented experiments.

Keywords: condition monitoring; industrial robots; anomaly detection; trend detection

1. Introduction

Currently, industrial robots are the workhorses of highly automated production systems [1]. A challenge to the productivity of such systems remain faults of industrial robot gears as they can cause extended downtimes. Condition monitoring (CM) of the gears can be a measure for countering this issue. CM describes a maintenance strategy in which sensor data is used to determine the health state of a robot gear. For this, sensor data is transformed into health indicators that correlate with the gear's health state. Critical monitored values within the time series of the health indicators form the decision criterion for a maintenance action [2]. Usually, there are many industrial robots operating in a production system and the health state of each of the axes must be monitored. Hence, manual monitoring is not feasible and an automated system is required. Such a system must be able to detect anomalies and trends in the health indicator data reliably. Anomalies in the data can be related to faults that occur abruptly (e.g., breaking of a gear tooth) and trends can be an indicator for increasing wear [3]. The occurrence of such events should be presented to the maintenance crew while showing only few false alarms. To the best of our knowledge, such a combined system does not yet exist for industrial robot gear condition monitoring. Hence, the contribution of our publication is threefold. Firstly, a combined anomaly and trend detection system (CATS) for industrial robot gear CM and secondly a method for selecting suitable anomaly detection (AD) and trend detection (TD) models for this defined application are presented. Thirdly, the suitability of different AD and TD models for the defined use case is evaluated by applying the method. Thus, the remainder of this publication is structured as follows: in Sections 1.1 and 1.2 an overview of industrial robot CM systems, AD and TD models is given and the addressed research gap is refined. In Section 2, CATS and the AD and TD model evaluation method is described. In Section 3, the method is applied to state-of-the-art AD and TD models and suitable models for CATS are selected. In Section 4, the limitations of the presented approach are discussed. In doing so, the outlook discussed in Section 5 is derived, which also includes a summary of our

contribution. Through the remainder of this publication the term application refers to the condition monitoring of industrial robot gears.

1.1. State of the Art

In this section, first supervised and unsupervised approaches for robot condition monitoring are presented. As this research area does not present the fields of anomaly detection and trend detection models completely, a broader overview of these research fields is given subsequently. Finally, the state of the art is summarised and the research gap is presented that we are addressing.

1.1.1. Industrial Robot Condition Monitoring

Different approaches for the CM of industrial robots exist in the literature. These can be classified by the type of model used, i.e., supervised or unsupervised machine learning models or the raw data used, which are mainly acceleration sensor data or robot controller data.

In the field of supervised models and robot controller data, several models such as XGBoost and different neural networks based on both joint specific data such as speed and torque and operational specific data (e.g., number of emergency stops) were compared from a fleet of 6000 robots. A maximum AUC value (area under the curve) of 0.87 could be achieved for a neural network model for fault detection in axis 2 [4]. A similar model comparison for logistic regression, support vector machines, random forests and ensemble stacking was performed in [5]. Here, angle, angle speed, acceleration and torque data were used from 26 robots to classify gear faults. The best AUC value of 0.77 was reached by the random forest classifier. Fault detection for loose gear belts was performed with a decision tree, a gradient booster and a random forest and statistical features derived from current data. Here, the random forest performed best with F1-scores around 0.9 [6].

In the section of unsupervised models and robot controller data, a kernel density estimator was used to detect faults based on motor angle, angle velocity and torque in combination with the Kullbach-Leibler divergence. Data from accelerated wear tests show a clear increase in the health indicator [7]. In another publication, the transferability of models was investigated for a combination of principle component analysis and Q-residuals. Anomalies were assumed if the distance measure was above a set threshold. The study shows that the use of the differences between measured and set quantities such as torques as raw data perform best in terms of transferability. In this context, transferability describes the training of the model based on the data of only one robot and then also using this model for other robots [8]. A model based on the deviations of a dynamic equation of a robot relative to actual measurements of the robot is combined with Hotelling's T^2 test statistic to determine robot faults [9]. A sliding-window convolutional variational autoencoder was used to detect anomalies in pick-and-place operations of a robot simulated by little strikes on the robot. The method outperforms benchmark models with an F1-score of 0.89 [10]. A long short-term memory neural network was successfully used to detect anomalies within the grinding process of an industrial robot based on speed, position and torque data. Anomalies were generated by applying a force to the robot hand during the process [11].

Turning to supervised learning approaches based on acceleration sensor data, multiple methods are worthy of note. A sparse autoencoder was trained with data from an attitude sensor (collecting acceleration and velocity signals at 100 Hz) attached to the tool centre point of the robot. The sensor collected data from normal behaviour and different fault conditions such as pitting and broken teeth of a gear. The classification results showed accuracy values of 90 percent [12]. Wavelet-based features in combination with a neural network were used to classify backlash faults for a six axis industrial robot [13]. Multiple supervised models such as a support vector machine, neural networks, gaussian processes and random forests were combined with different dimensionality reduction methods based

on data from acceleration sensors attached to the gear caps for gear fault classification. The SVM and GP showed the best performance with accuracy values over 91 percent [14].

In the area of unsupervised models and acceleration sensor data, a gaussian mixture model was used based on health indicators derived from time and the time-frequency domain to differentiate measurements from a degreased robot from normal measurements of the robot. Classification performances over 94 percent for recall and precision values were achieved [15]. Time domain and frequency domain features derived from a residual signal were used in combination with thresholding for gear fault detection for different test trajectories [16]. A one-class generative adversarial autoencoder was used for the detection of artificially introduced faults in a robot gear in [17]. Classification accuracies of 97 percent were achieved for the identification of different faults.

1.1.2. Anomaly Detection Models

The state of the art provides various anomaly detection models for point, collective and contextual anomalies of uni- and multivariate time series and spatial data. One possibility for clustering such models is presented in [18]. Here, anomaly or novelty detection methods are structured in probabilistic, distance-based, reconstruction-based, domain-based and information theoretic approaches. For a detailed review of anomaly detection methods, refer to [18] or more recently to [19]. Below, only those approaches that are considered in the method evaluation of our publication are presented. Different approaches from the above mentioned classification scheme are compared. From the field of probabilistic models, a kernel density estimator (KDE) based on the values of the time series [20] is used. This model fits a non-parametric probability density function on the data. By calculating the probability that a sample (one step of a time series) belongs to this density and by comparing this value with a threshold, anomalies can be determined. Furthermore, a gaussian process (GP) for one-class classification is used, which works based on a similar principle [21]. From the field of distance based approaches, the local outlier factor (LOF) [22], the isolation forest (IF) [23] and the DBSCAN algorithm [24] are used. LOF is based on determining the density of data points and detects anomalies as data points with few close neighbors. IF is based on multiple tree classifiers for one-class classifcation. DBSCAN is a clustering algorithm that determines anomalies based on their distance to reachable points from cluster core points. Multiple representatives from the reconstruction-based model class are used. An autoregressive (AR) [25] and autoregressive moving average model (ARMA) [26] are applied and compared with a convolutional and a long short-term neural network [27,28]. All four models are used as regression models between the past time steps of the signals and a time step of the signal in the future. The deviations between these predictions and the actual progress of the signal are then compared with a threshold. If the deviation exceeds the threshold, an anomaly can be assumed. Furthermore, the one class support vector machines (OCSVM) [29] as a domain-based model is included for the comparison. This model builds a domain of inliers based on support vectors and the border data points of this domain. Data points outside this border line are classified as anomalies. As a simplistic baseline model, an approach is considered where a data point is compared to a multiple of the standard deviation of the reference data (abbreviated STD). If this distance exceeds a defined threshold, an anomaly is assumed.

1.1.3. Trend Detection Models

In the context of this publication a trend is defined as the gradual change in future events from past data in a time series [30]. Trend detection can be differentiated from remaining useful life (RUL) estimation by several aspects. In contrast to RUL estimation, trend detection methods do not extrapolate existing time series into the future. Furthermore, no thresholds for the extrapolated time series are defined which describe the end of lifetime of an asset. Trend detection methods have different purposes. It is possible to differentiate between models for change point detection, trend description and identification of trend

presence in a time series. For the considered application, a model is required that answers the question of whether a trend is present. This is why the remainder of this subsection focuses on the field of trend presence identification. Here, various statistical tests exist. The Mann-Kendall test (MK) is a sign test based on pairs of all samples of a time series and their predecessors [31] to detect trends. The Cox-Stuart (CS) test uses a reduced amount of data pairs for a sign test [32] to achieve the same objective. The Wilcoxon-Mann-Whitney trend test builds a test statistic based on the signs of the slopes between samples and the rank sums of the samples with an increasing and decreasing slope [33] for this purpose. The Durbin-Watson test checks for auto-correlation in the residuals of a regression fit. If the residuals do not show autocorrelation, a trend can be assumed [34]. Furthermore, slope based approaches in combination with thresholds exist. The most simple approach from this field is to fit a linear or quadratic function to the time series data, calculate the slope of this function and compare it with a threshold. This model will be named linear regression model, short LR, for the rest of the publication. If the slope exceeds the threshold value, a trend can be assumed. A more complex approach for trend detection is based on the clustering of a time series. In a first step, a clustering algorithm (e.g., Fuzzy-K-Means) is used to detect clusters within the time series. Then, the slope between the cluster centres is determined. Finally, the slope values of the cluster centres are compared with a threshold to decide, whether a trend exists [35]. The last approach for trend detection presented in this section is based on the comparison of the time series' moving average with its overall mean (moving average model, short MA). In a first step, these two quantities are calculated. Afterwards, the time series' standard deviation multiplied by a factor is added to the overall mean to determine a threshold. Then, it is determined, whether the moving average of the signal rises above this threshold for a defined time window. If this is the case, it can be assumed that a trend is present in the signal. The principle behind this method is also illustrated in Figure 1.
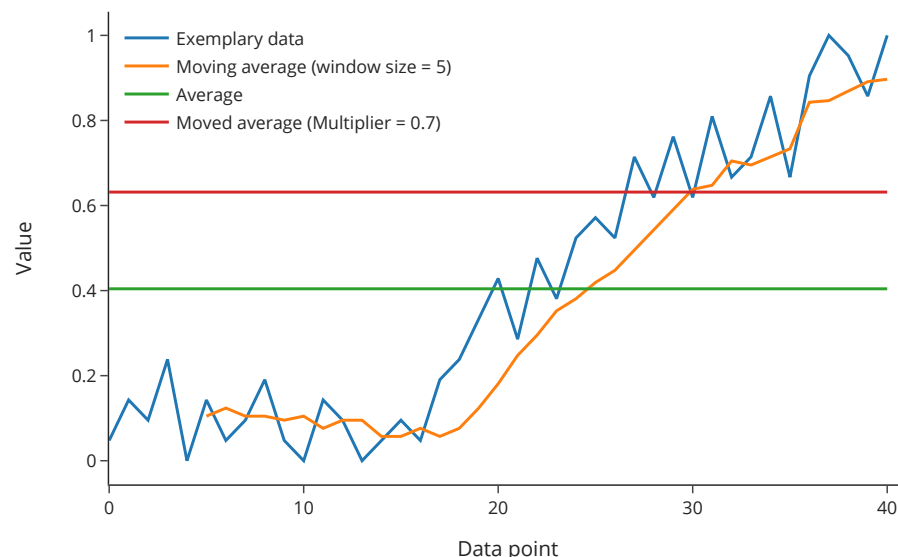

Figure 1. Example for the trend detection method using moving averages.

1.2. Considered Research Gap

In the field of industrial robot gear condition monitoring no combined AD and TD model has been presented up to now to the best of our knowledge. Therefore, the research objective of this publication is to present such a system. For the detailed design of this system, a suitable AD and TD model must be chosen. As no comparison of AD and TD models for univariate time series of HIs derived from acceleration sensors has been

performed up to date, a method to select suitable AD and TD models for the application of industrial robot gear condition monitoring is formulated. Afterwards, it is applied to choose models for the presented combined system. In the context of the framework presented in [36], we address the question of algorithm selection for the inference task. By doing so, we support the transfer of state of the art AI models into practice and reduce the effort of model selection for practitioners. The identification of suitable data acquisition systems or the selection of features is not considered in this publication. This is e.g., considered in [3]. Therefore, the presented work builds up on assumptions derived from this publication. These assumptions are summarized in Section 2.1.1. Furthermore, we limit our research frame to the field of six-axis articulated robots as we can not provide comprehensive experiments for other asset classes and hence validate our approach for such assets.

2. Materials and Methods

In this section, firstly CATS is described. Subsequently, the method for selecting suitable AD and TD models for CATS is described.

2.1. Combined Anomaly and Trend Detection Model

The objective of CATS is the reliable detection of trends and anomalies in industrial robot gear health indicator data. In the following, the assumptions that the system is based on, are defined. Then, the system itself is presented.

2.1.1. System Assumptions

The presented model builds upon certain assumptions. Data ingested in the system must be collected from a setup with a constant robot trajectory and load. The system analyses only univariate time series data of one health indicator per axis derived from acceleration sensor data. A suitable HI is described for example in [3]. The HI exhibits stationary behaviour when the robot axis is in a healthy state. The considered time series can be subject to trends $x_{trend}(t)$, seasonality $x_{seasonality}(t)$, noise $x_{noise}(t)$ and anomalies $x_{anomaly}(t)$. Noise can be caused by changing environmental conditions or sensor effects. Trends can occur due to wear. Trends due to sensor drifts are prevented by the sensor setup or suitable data preprocessing (e.g., high pass filtering of the raw data). Seasonality can occur due to changing temperatures of the gears. These temperature changes lead to variations in the HI (for example, see [37]). These temperature changes result from varying utilisation in the production system. They could be caused for instance by a three shift working model with reduced utilisation during night shift. Summarising, this time series can be expressed as in Equation (1).

$$x(t) = x_{trend}(t) + x_{seasonality}(t) + x_{noise}(t) + x_{anomaly}(t) \tag{1}$$

2.1.2. System Design

The objective of the presented system is to evaluate whether $x_{anomaly}(t) \neq 0$ or $x_{trend}(t) \neq 0$. For this, an anomaly detection model and a trend detection model are deployed in parallel. The detection of an anomaly in a defined number of sequential measurements leads to the recommendation of immediate maintenance actions. The detection of trends in the data of a defined number of a sequential measurements leads to the proposal of maintenance actions in the near future. The working principle of the system is summarized in Figure 2. The design of the system addresses different aspects of the industrial robot gear condition monitoring use case. Faults, whose manifestation but not the underlying fault mechanism progress (e.g., tracking of the growth of a crack in a gear tooth) can be tracked with HIs, will cause point or collective anomalies. The AD model will be used for the detection of such faults. Other faults, whose progress can be tracked (e.g., increasing wear), will cause trends in the HI. These trends will be detected by the trend detection model.
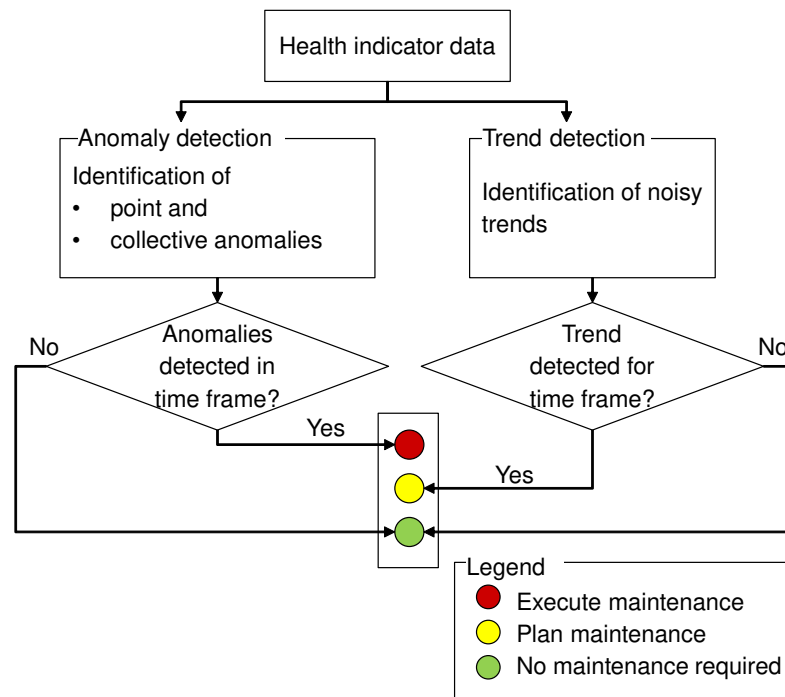
Figure 2. Overview of the condition monitoring system.

2.2. Method for Anomaly and Trend Detection Model Selection

　　In this section, the overall model evaluation method is proposed. Then, more detailed information is given about the generation of synthetic data and the model evaluation criteria.

2.2.1. Overall Method and Selected Models

　　To select suitable AD and TD models for the presented system a three step approach was followed to ensure that the most suitable models are chosen. Firstly, potential models were identified in the literature. Secondly, these models were applied on synthetic data meeting defined characteristics of the considered application and evaluated in respect of different quality criteria to reduce the solution space. Thirdly, the best performing models were evaluated using real world data taken from accelerated wear tests of industrial robots. The overall selection process is summarised in Figure 3. In the following, these steps are explained in detail.
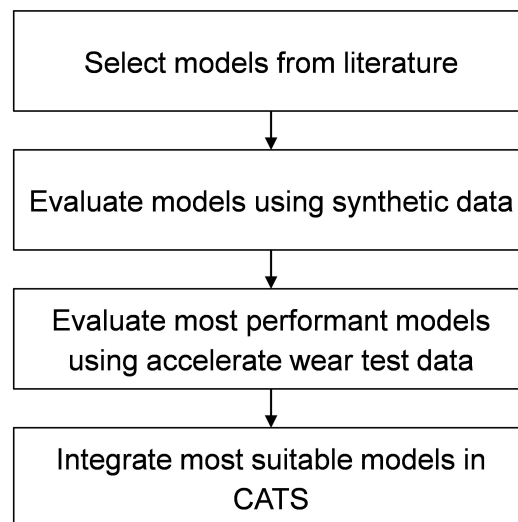

Figure 3. Overview of the model evaluation method.

As described in Section 1.1, a large number of AD and TD models exist. Hence, a holistic comparison of existing approaches is not feasible. Therefore, models from the classes as described in [18] were chosen for the AD model comparison. In detail, the models listed in Table 1 were used. The models are explained in detail above in Section 1.1.2. For TD model comparison, the MK test, the CS test as well as the LR and MA based approaches described in Section 1.1.3 were chosen. The implementation of the models is described in an open source repository [38].

Table 1. Models considered.

Anomaly Detection Model	Anomaly Detection Model Type	Reference
CNN	Reconstruction based	[28]
LSTM	Reconstruction based	[27]
AR	Reconstruction based	[25]
ARMA	Reconstruction based	[26]
KDE	Probabilistic	[20]
GP	Probabilistic	[21]
OCSVM	Domain based	[29]
IF	Distance based	[23]
DBSCAN	Distance based	[24]
LOF	Distance based	[22]
STD	Distance based	[-]
Trend detection model	**Trend detection model type**	**Reference**
MK	Statistical test	[31]
CS	Statistical test	[32]
LR	Slope based	[-]
MA	Slope based	[-]

2.2.2. Synthetic Data Generation

For the model comparison based on synthetic data, a data generator was implemented to create time series as described in Equation (1). Different trend, noise, seasonality and anomaly functions were considered. In detail, linear and quadratic trend functions were implemented. White noise and uniform noise with different variances or ranges were used as noise functions. Sine functions and a hand crafted function as described in Equation (2) were applied for seasonality. Here, t is the current time step, which would relate to the length of one hour of the time series and a is the magnifier factor, which is further described in Table 2. An example of this function is depicted in Figure 4 on the upper right side.

$$f(x) = \begin{cases} (t\%24) \times a/4 & \text{if } (t\%24) \leq 4 \\ 1, & \text{if } 4 < (t\%24) \leq 20 \\ \frac{24-(t\%24)}{4} \times a & \text{otherwise} \end{cases} \quad (2)$$

For the anomaly function, a uniform distribution was used to define the anomaly positions. Different lengths for collective and different amplitudes for both collective and point anomalies were applied. To derive reasonable parameter ranges, certain realistic assumptions were made. A time series consists of 8736 samples representing 24 measurements per day for one year. The range of the trend functions' slopes should allow a doubling of the HI value in no less than one week and no more than half a year. Noise and seasonality should as a minimum result in a deviation of the time series by the factor 0.3 and as a maximum by the factor 9 from the mean of the signal. These assumptions were based on collected HI data from industrial robots in a car manufacturing plant. Due to confidentiality reasons, this data can not be published. The different functions, their parameters, the range of the parameters used and underlying assumptions for the parameter range choice are specified in Table 2. In the first three months of the time series no anomaly or trend occurs. In the last nine months anomalies may occur. Figure 4 shows a typical synthetic time series.
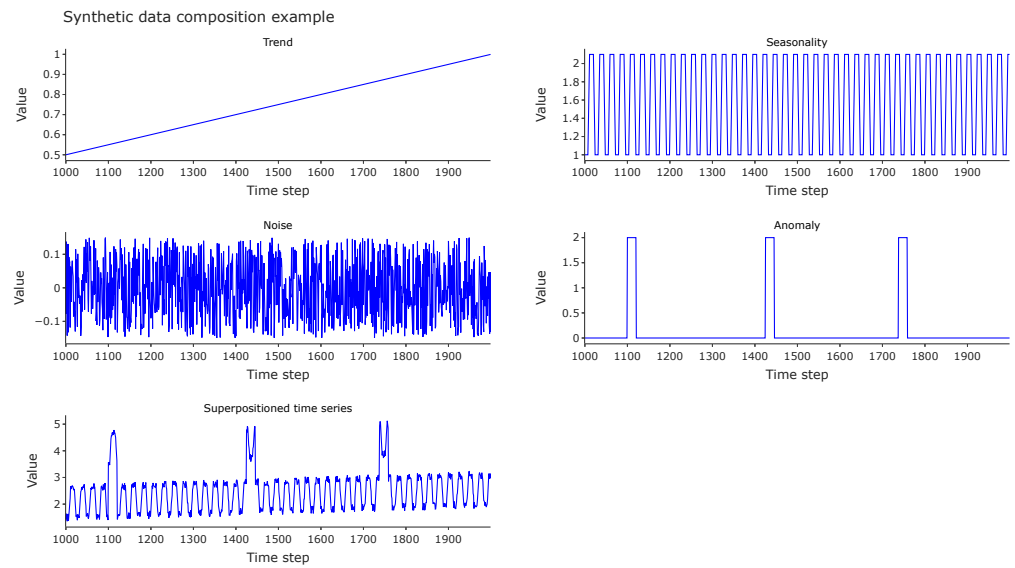
Figure 4. Example of a synthetic time series.

Table 2. Overview of used parameter ranges for the synthetic time series.

Signal	Parameter Type	Parameter Values	
		Synthetic Data Set 1	**Synthetic Data Set 2**
$x_{trend}(t)$	Trend type	Linear	
		Quadratic	
	Trend slope	Linear: 0.012	Linear: 4.58×10^{-4}
		Quadratic: 7.09×10^{-5}	Quadratic: 1.05×10^{-7}
$x_{seasonality}(t)$	Seasonality type	Sine	
		Production cycle (Formula 2)	
	Amplitudes a	Sine: 0.15	Sine: 3
		Production cycle: 1.1	Production cycle: 2
$x_{noise}(t)$	Noise type	Uniform noise	
		White noise	
	Noise parameters	Uniform noise range: 0.15	Uniform noise range: 1
		White noise mean: 0 White noise standard deviation: 0.03	White noise mean: 0 White noise standard deviation: 0.8
$x_{anomaly}(t)$	Anomaly types	Point anomaly	
		Collective anomaly	
	Anomaly parameters	Amplitude: 2	Amplitude: 1.1
		Collective anomaly lengths: 20 measurements	Collective anomaly lengths: 5 measurements

Based on this parameter range, over 26 million unique time series could be modeled. To reduce the computational effort, two reduced data sets were created. The first data set (synthetic data set 1) was used for an initial screening of the models' performance.

It consisted of time series with low noise, trends with a high slope, and large anomaly magnitude values and lengths. Furthermore, a second data set (synthetic data set 2) with more difficult conditions for the detection of trends and anomalies was generated. Here, time series with high noise, low trend slopes, and low anomaly magnitudes and lengths were calculated. In each time series 40 anomalies were present. Each created time series was analysed by each model to detect trends and anomalies. In total, 16 unique time series were analysed per data set.

2.2.3. Model Evaluation

To measure the models' performance, the ROC curves (receiver operating characteristic curves) for different parameter choices of the models were determined. This means that different model parameters were varied and the True Positive Rate (TPR) and False Positive Rates (FPR) of the models for the synthetic data were determined. More precisely, the models were presented with slices of the time series and had to determine, whether trends or anomalies were present in the time series. For the trend detection task, these slices were increased in size per time series with a window size of 1008 samples and an initial size of 2016 samples. This is equivalent to 24 measurements per day for a length of 12 weeks for the initial window. For the anomaly detection, the first 168 values were used to train the models. This is equivalent to 24 measurements per day for one week. The models were then tested on time series with a length of 6720 samples. The parameters that were varied for the different models are summarized in Table A1. The most robust models with high TPR and low FPR and high average AUC values (area under the curve) were then applied to data sets from accelerated robot gear wear tests. A data set, which is based on an accelerated wear test with an ABB IRB 6600-255/2.55, was used to test the trend detection models (Accelerated wear test 1). The experiment caused different faults in the robot gear of the second axis. In total, 2425 measurements over a time span of roughly one year were used from the experiment; these were acquired with an acceleration sensor at the robot gear cap. From this data the HI described in [3] was derived. For more information regarding the experiment, see [39,40]. The same data set and another data set, which was acquired during another accelerated wear test with an ABB IRB 7600-340/2.8, to test the anomaly detection models (Accelerated wear test 2). Here, 920 measurements were acquired over three months at the second axis gear cap with an acceleration sensor, and the same HI was calculated and various gear faults were subsequently detected in the second axis gear. As no obvious trend could be seen in this data set, it was just used for the AD model evaluation.

More information regarding this experiment is given in [3]. Figure 5 presents the various faults of both accelerated wear tests. For analyzing these data sets, the models' parameters were chosen that yielded the best compromise in TPR and FPR during the experiments with the synthetic data. In a real world setup, other parameter sets could be more reasonable in respect of the trade-off between false alarms and undetected faults. A method of how to choose the best parameters given the maintenance circumstances of an individual robot is discussed in Section 4. Based on the results of the accelerated wear test experiments, a suggestion of which models to use for trend and anomaly detection in the CM system is made. The detailed model evaluation method based on synthetic data is depicted in Figure 6.

Faults of the IR 7600 experiment

Faults of the IR 6640 experiment

Figure 5. Overview of the faults of the accelerated wear tests following [3,39].
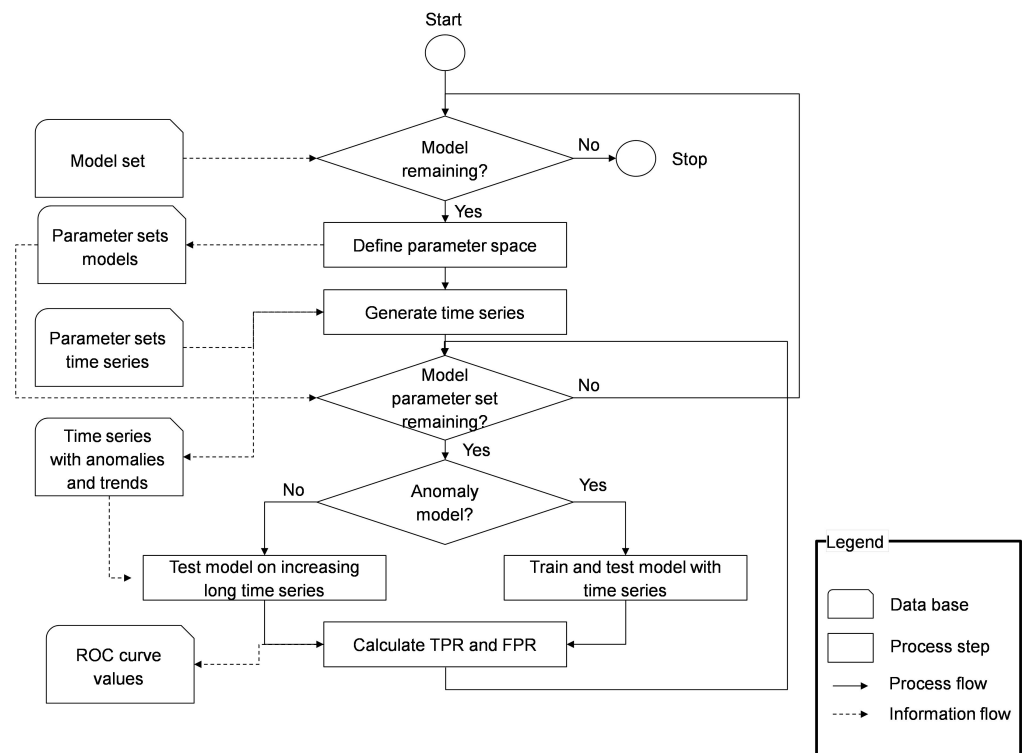

Figure 6. Overview of the model evaluation method.

3. Results

In the following, the presented method from the last section is applied to the AD and TD models listed in Table 1. First, the results for the TD models are shown, then the results of the AD models.

3.1. Trend Detection Model Comparison

Here, first the evaluation of the TD models based on synthetic data are presented. Subsequently, the results based on the accelerated wear test are analysed.

3.1.1. Evaluation Based on Synthetic Data

Figure 7 shows the ROC curve derived from the synthetic data set 1 and the model parameters described in Table A1. Ideally, the plots would show a dot in the upper left corner for a model. Such a dot would refer to a perfect classifier. This means that the model has a TPR of 1 and FPR of 0. Such a model would detect all trends and trigger no false alarms. The LR model and the MA model achieve these perfect classification results. The variation of parameters of the CS model does not influence the model performance and the MK model shows high TPR values only at the expense of an increased false positive rate. The results of synthetic data set 2 with the same model parameters are shown in Figure 8. Here, the CS model shows the best performance as a parameter combination exists where no false alarms are triggered and all trends are detected. It is followed by the MK model, which also yields a performance where all trends are detected and the FPR is small. The LR and the MA models achieve high TPR values only at the expense of increased FPR. The AUC values of the models for both data sets are presented in Table A2. Based on these results, it was decided to apply the CS and the MK model to the accelerated wear tests as they performed best on the more difficult data (synthetic data set 2) and based on their average AUC values.
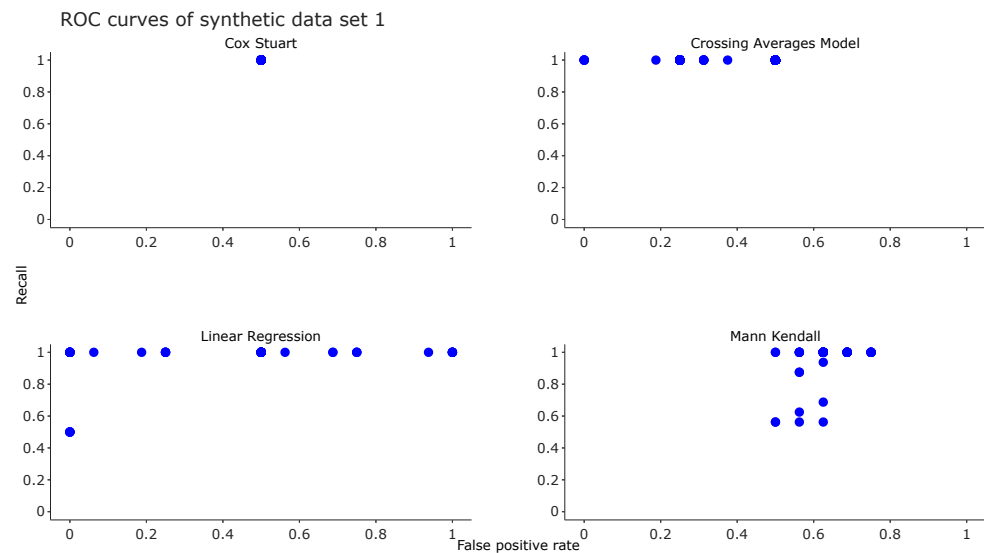

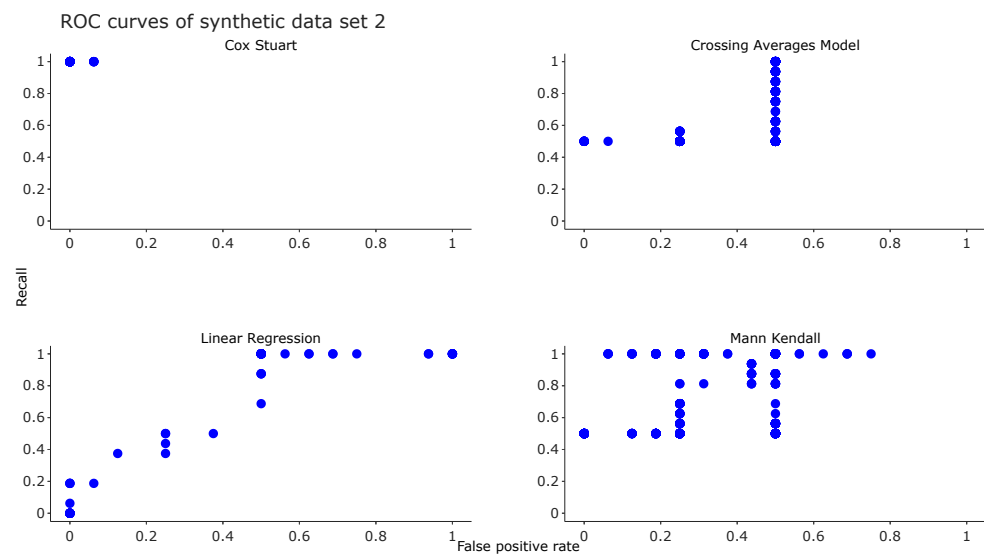**Figure 7.** Trend detection model comparison based on the synthetic data set 1.

Figure 8. Trend detection model comparison based on the synthetic data set 2.

3.1.2. Evaluation on Accelerated Wear Test Data

The data from the accelerated wear test was analysed using the two chosen models. The results are depicted in Figure 9. The blue line shows the health indicator values, the dots indicate the models' decision of whether a trend is present in the time window of the last 504 samples (which equals a time frame of 2.5 months) while the horizontal yellow line shows, when more then 50 percent of the last 504 decisions were positive.

In such a case, a maintenance action should be planned. It can be seen that both models show similar behaviour for the beginning of the data set where they both detect a trend in the data after the initialisation phase of the first 504 measurements. The outlier at measurement 1000 leads to the rejection of the hypothesis that a trend is present for the following measurements in the MK model. It can be assumed that the CS model interprets the outlier correctly so that even for the following measurements a trend is detected. Both models detect the more stationary behaviour of the time series at its end. As the CS model handles the outlier around measurement 1000 better compared to the MK model, it is suggested to use the CS model in CATS. In this experiment, the confidence level parameters from the ROC curve of synthetic data set 2 were chosen for the models that yielded the highest TPR values with the lowest FPR at the same time.
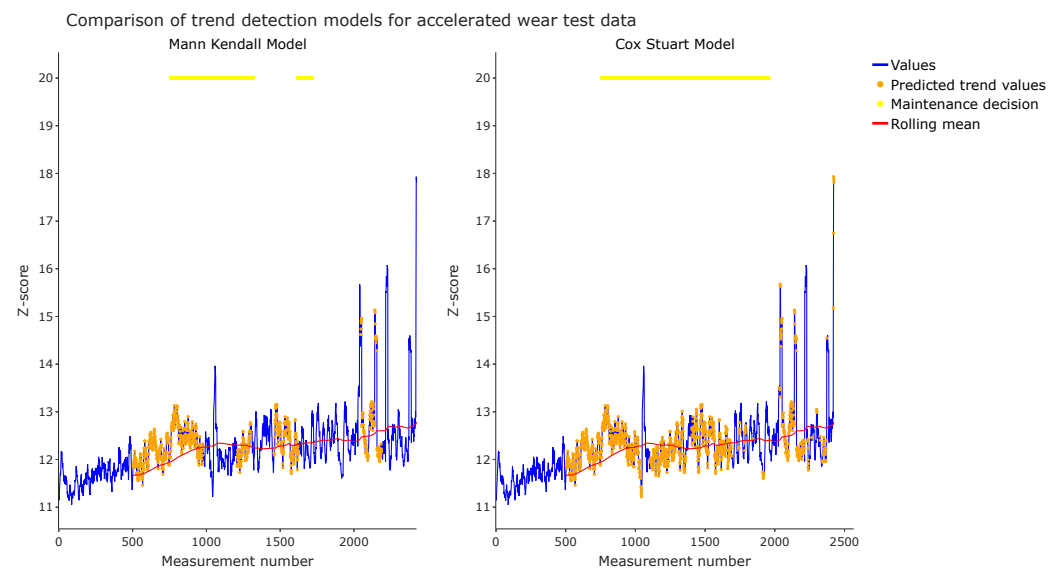

Figure 9. Results of the trend detection models based on accelerated wear test data.

3.2. Anomaly Detection Model Comparison

The presentation of the results of the AD model comparison follows the same scheme as Section 3.1.

3.2.1. Evaluation Based on Synthetic Data

The ROC curves of different models for the synthetic data set 1 are shown in Figure 10. Again, as described in Section 3.1.1 the plot would ideally show dots for the models at the upper left corner. Most of the models show good results except the OCSVM for which parameter combinations exist that yield poor classification performance. This means that all models are capable of identifying anomalies reliably and with a low false alarm rate in the case of high anomaly amplitudes and low noise level. In contrast, the models' overall performance regarding the synthetic data set 2 is rather poor. Figure 11 summarises the ROC curves for this data set. No perfect classifier was found for all models and the distance of the models' ROC curves to the upper left corner is large. Here, it can be concluded that the models struggle to detect anomalies at high noise levels and low anomaly amplitudes. This fact will also be discussed in Section 4. The AUC values for all models and both data sets are provided in Table A3. The individual ROC curves of all models for bothd data sets

are presented in Figures A1 and A2. The best overall performance show the LSTM, STD and LOF models based on their average AUC values. Hence, it was decided to use the LSTM, STD and the LOF model on the accelerated wear test data.
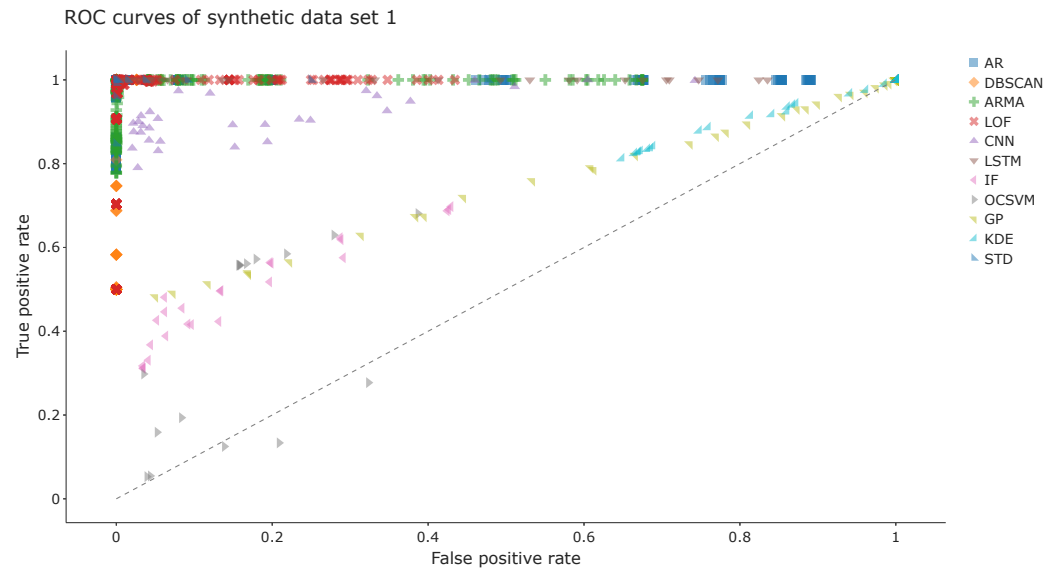

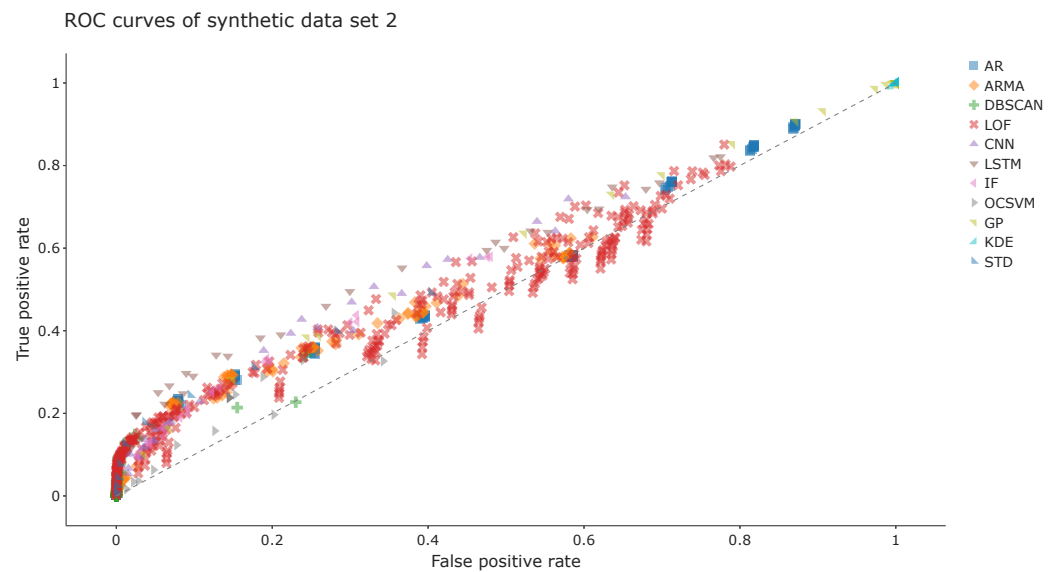**Figure 10.** Results of the anomaly detection models based on synthetic data set 1.

Figure 11. Results of the anomaly detection models based on synthetic data set 2.

3.2.2. Evaluation on Accelerated Wear Test Data

The results of applying the LSTM, STD and LOF models to the data from the accelerated wear test 1 are depicted in Figure 12. For this, all models were trained based on the first 500 measurements with model parameters of the ROC curves that yielded the best compromise between high TPR and low FPR values. It can be seen that all models correctly identify the anomalies at the end of the time series. The LOF model detects the outlier around measurement 1000 as an anomaly. Given a maintenance action decision criterion of 10 detected anomalies in the last 24 measurements, maintenance actions would have been triggered at the end of the data set for all models and a false alarm would have been triggered around measurement 1000 for the LOF model and for many more time ranges for the STD model. The AD models' behaviour on the second data set are summarized in a similar manner in Figure 13. In this scenario, the models were trained

using the first 200 measurements with the same model parameters. It can be seen that the LSTM model and the STD model detect more anomalies than the LOF model along the time series. The apparent anomaly at the end of the time series is detected by all models.The LSTM triggers two false alarms around measurement 300. The STD model triggers many false alarms. Summarising, the STD shows more false alarms compared to other models. The LOF and LSTM model detect only the apparent anomalies with a low false alarm rate. Hence, it is suggested that either the LOF model or LSTM model is used in CATS as the AD model.
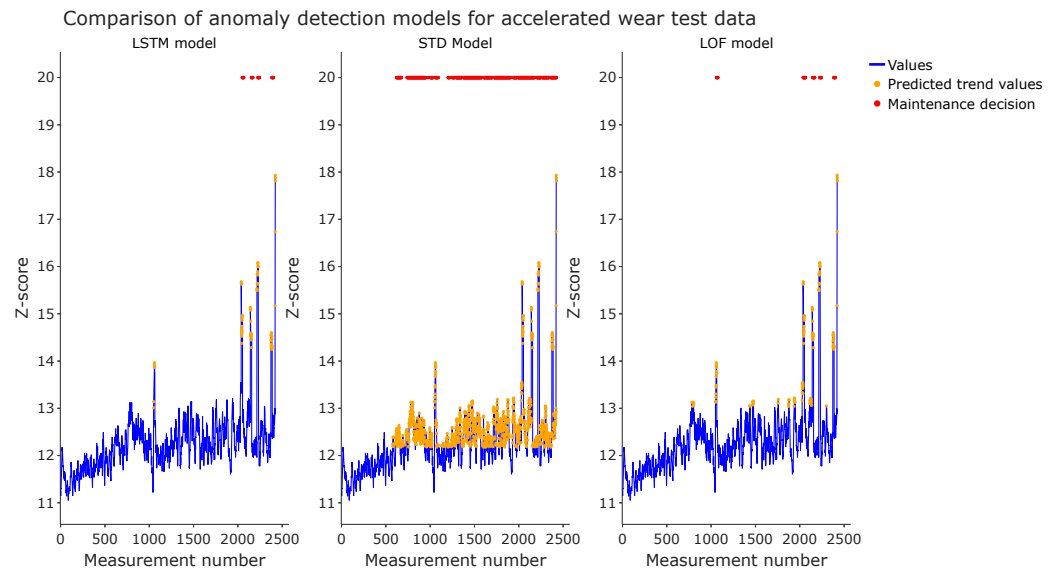
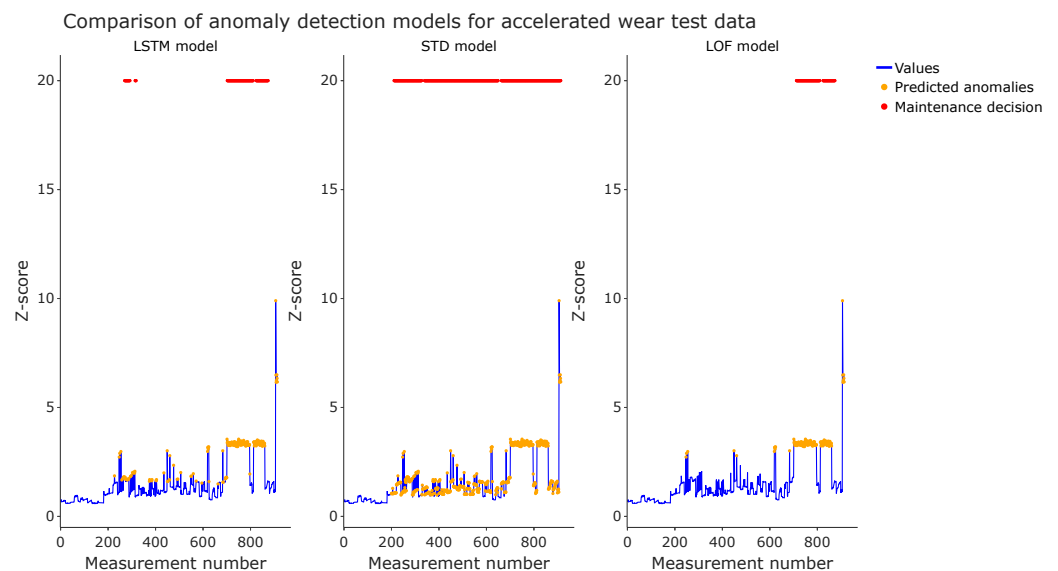**Figure 12.** Results of selected anomaly detection models for accelerated wear test 1.

Figure 13. Results of selected anomaly detection models for accelerated wear test 2.

4. Discussion

The presented results highlight some interesting aspects that will be discussed in this section. We will justify our initial choice of models and highlight some aspects of the models' performance on the synthetic data. Then, we will explain the models' parameter choice and end with organisational thoughts regarding the integration of CATS in a real world production site.

As emphasised in Section 1.1.2, a comprehensive comparison of AD and TD models is not feasible due to the high variety of existing models. Our motivation for selecting models from different categories as presented in [18] was to test how their underlying detection mechanisms cope with the different characteristics of time series. The fact that AD and TD models were found that detect the trends and anomalies in the accelerated wear test data reliably, strengthens the argument that the comparison of the selected models is sufficient for the application. From our point of view, the results of the AD model comparison based on synthetic data set 2 clearly highlights the limitations of anomaly detection models in general. High noise levels in the data make it difficult for such models to detect anomalies. Figure 14 shows a typical time series of this data set. Even as a human operator, it is difficult to identify the anomalies. However, from our experience, such extreme noise does not appear in the HI time series as shown in Figure 9 or Figure 13 for the accelerated wear tests. When deploying AD or TD models in real world applications, suitable model parameters must be chosen. For this, from our point of view, the parameters have to be configured for the individual robot considering the common trade-off between false alarms (higher FPR) and undetected faults (lower TPR). If no ideal anomaly or trend detection model can be used considering the ROC curves, this trade-off can be tackled by considering a maintenance score for an individual robot. This maintenance score can be influenced for example by the position of the robot in the production systems in respect of the distance to buffers or the effort required to exchange the robot. Other criteria could be the required calibration effort after the replacement or the response time of the maintenance team if a replacement is required. For robots with a higher maintenance score, model parameters with high TPR and higher FPR should be chosen. For robots with a lower maintenance score, model parameters with lower TPR and low FPR should be selected. This principle is also depicted in Figure 15. The reconfiguration of such models might also be required if the FPR or TPR do not meet the expected behaviour over time. Finally, the implications that the formulated assumptions in Section 2.1.1 yield must be discussed. To meet these assumptions, two aspects must be considered in a real world application. First of all, a measurement trajectory must be used for data acquisition so that the HI data is comparable and has a low noise level. Secondly, CATS must be extended by mechanisms to ensure that anomalies or trends in the HI data are only due to wear and not changing environmental conditions, new robot programs or faulty data acquisition systems.
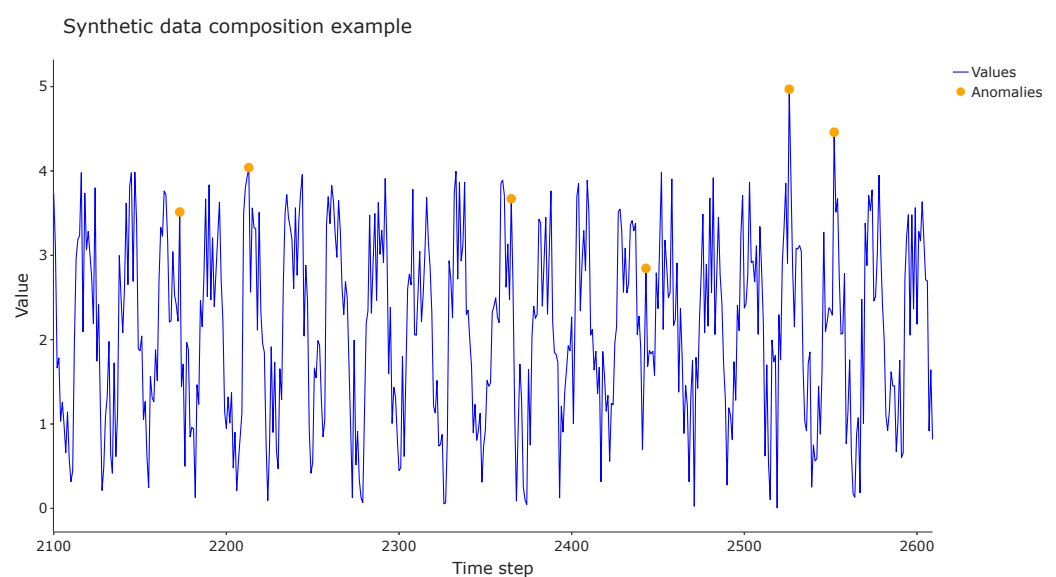

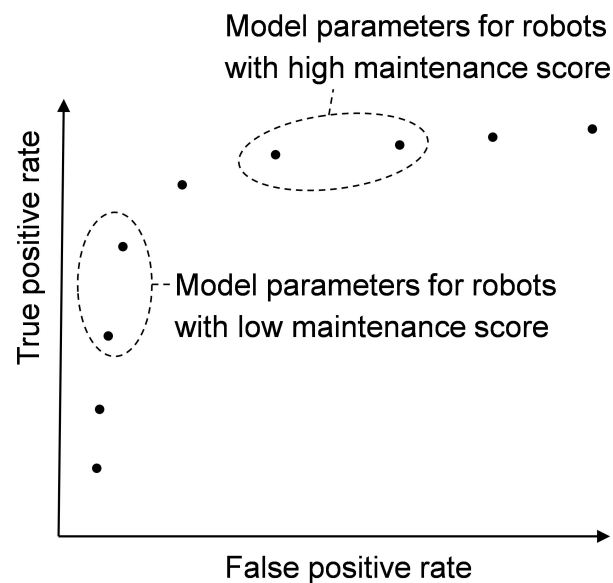**Figure 14.** Example of a noisy time series from synthetic data set 2.

Figure 15. Selection of model parameters based on a maintenance score.

5. Conclusions

A combined anomaly detection and trend detection system for the condition monitoring of industrial robot gears has been presented. To select suitable models for these tasks, a method in which models are evaluated based on synthetic and accelerated wear test data was formulated. The synthetic data consists of time series with noise, cyclic behaviour, trends and anomalies based on realistic assumptions that were gathered from industry data. The accelerated wear test data was collected during two experiments with six-axis industrial robots, which provoked multiple gear faults and exhibited both trends and anomalies. By applying the presented method, it was found that the Cox-Stuart test is most suitable for trend detection and the local outlier factor algorithm or the long short-term neural network are capable of detecting the anomalies in the accelerated wear test data. For future research, we believe that the considerations in Section 4 regarding the extensions of CATS with functionalities to detect reasons for false alarms such as robot program changes or the change of the robot tool and the automatic reconfiguration of models in case of too many false alarms are the most important topics for enabling the automated condition monitoring of industrial robot gears in industry.

Appendix A. Model Parameters for ROC Curves

Table A1. Considered Models.

Model	Parameter	Range
MK	Confidence interval	0.9–0.999 in variablestep sizes
CS	Confidence interval	0.9–0.999 in variable step sizes
LR	Slope threshold	0–1 in variable step sizes
MA	Amplifier	min: 0.5, max: 0.8, step: 0.1
	Length above threshold	min: 24, max: 72, step: 2
	Moving Average Window	0.04×(dataset length)–0.18×(dataset length)
ARMA	Autoregression lags	min: 1, max: 9, step: 2
	Moving average lags	min: 0.2, max: 1, step: 0.2
	Anomaly threshold	min: 0.01, max: 0.1, step: 0.02
AR	Autoregression lags	min: 0.2, max: 1, step: 0.2
	Anomaly threshold	min: 0.01, max: 0.1, step: 0.02
CNN	Training epochs	10, 20, 50
	Anomaly threshold	0.1, 0.2, 0.3, 0.4, 0.5, 0.9, 0.95, 0.98, 0.99, 0.999
LSTM	Training epochs	10, 20, 50
	Anomaly threshold	0.1, 0.2, 0.3, 0.4, 0.5, 0.9, 0.95, 0.98, 0.99, 0.999
DBSCAN	Epsilon	min: 0.1, max: 1.3, step: 0.2
	Minimal number of samples	13, 21, 34, 55, 89, 144, 233, 377
GP	Anomaly threshold	0.7, 0.8, 0.9, 0.95
	Kernel upper bound	0.0001, 0.0005, 0.001, 0.002, 0.003, 0.005, 0.008, 0.013, 0.021, 0.034, 0.055, 0.089, 0.144, 0.233, 0.377, 0.61, 0.987
IF	Number of estimators	50, 100, 200
	Contamination	0.01, 0.02, 0.03, 0.05, 0.08, 0.13, 0.21, 0.34
LOF	Number of neighbors	5, 10, 20, 30, 50, 80
	Contamination	0.001, 0.01, 0.02, 0.03, 0.05, 0.08, 0.13, 0.21, 0.34, 0.5
OCSVM	Kernel	'rbf, sigmoid
	Nu	0.01, 0.02, 0.03, 0.05, 0.08, 0.13, 0.21, 0.34
KDE	Bandwidth	0.2, 0.3, 0.5, 0.8, 1.3, 2.1, 3.4, 5.5
	Anomaly threshold	0.75, 0.9, 0.95, 0.99

Appendix B. AUC Tables

Table A2. Overview of the AUC values of the trend detection models.

	Synthetic Data Set 1	Synthetic Data Set 2
Cox Stuart	0.750000	0.968750
Crossing Averages Model	1.000000	0.679688
Linear Regression	0.984375	0.707031
Mann Kendall	0.732422	0.861328

Table A3. Overview of the AUC values of the anomaly detection models.

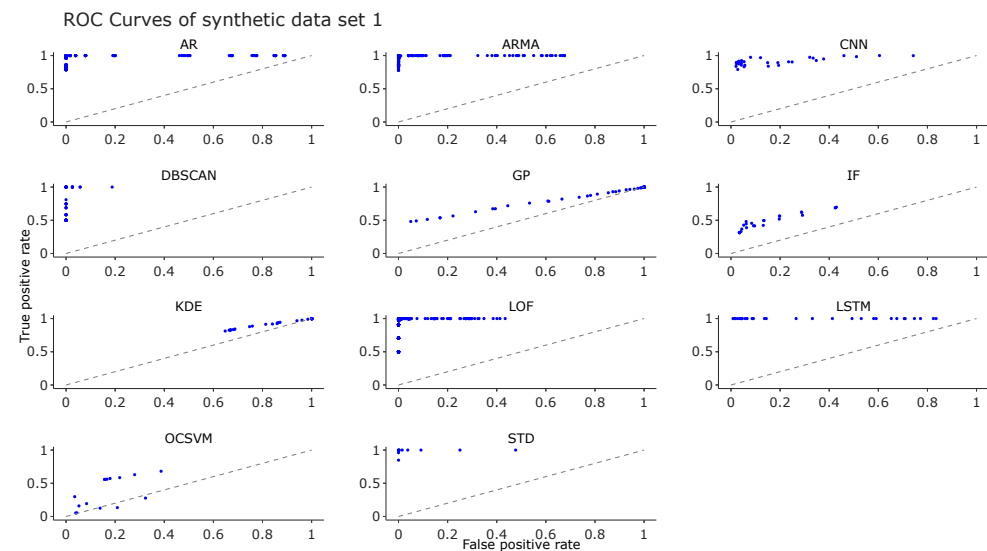|	Synthetic Data Set 1	Synthetic Data Set 2
AR	0.998641	0.550152
ARMA	0.999713	0.553447
CNN	0.955253	0.596032
DBSCAN	1.000000	0.553332
GP	0.716030	0.591000
IF	0.706679	0.580334
KDE	0.584487	0.499234
LOF	0.999880	0.550426
LSTM	0.995874	0.612189
OCSVM	0.656618	0.542888
STD	0.999951	0.576333

Appendix C. Individual ROC Curves

Figure A1. Results of the anomaly detection models based on synthetic data set 1.
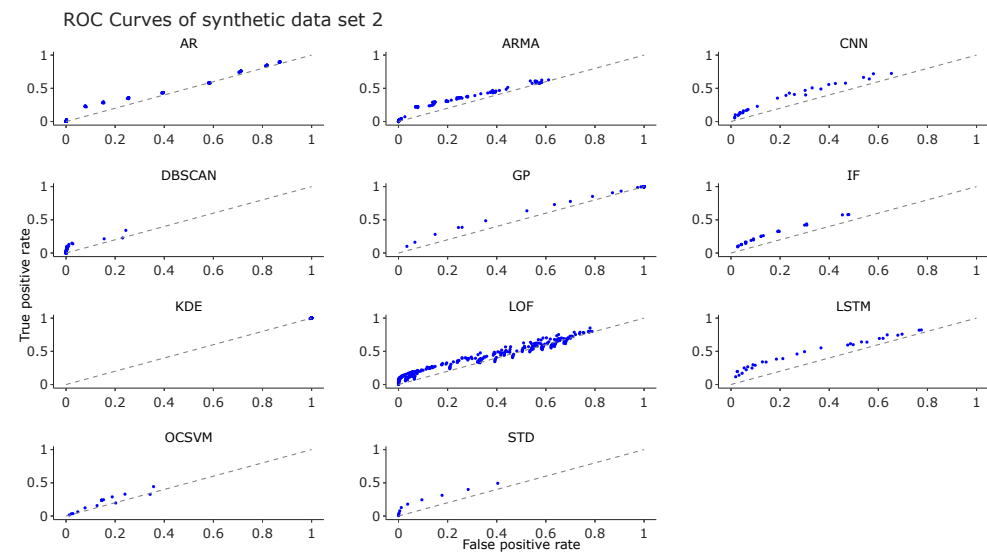

Figure A2. Results of the anomaly detection models based on synthetic data set 2.

References

1. Krockenberger, O. Industrial Robots for the Automotive Industry. In *SAE Technical Paper Series*; SAE International: Warrendale, PA, USA, 1996. [CrossRef]
2. 17359: 2018-05, Zustandsüberwachung und-Diagnostik von Maschinen—Allgemeine Anleitungen (ISO_17359: 2018). 2018. Available online: https://www.iso.org/standard/71194.html (accessed on 4 November 2021).
3. Nentwich, C.; Reinhart, G. A Method for Health Indicator Evaluation for Condition Monitoring of Industrial Robot Gears. *Robotics* **2021**, *10*, 80. [CrossRef]
4. Costa, M.A.; Wullt, B.; Norrlöf, M.; Gunnarsson, S. Failure detection in robotic arms using statistical modeling, machine learning and hybrid gradient boosting. *Measurement* **2019**, *146*, 425–436. [CrossRef]
5. Sathish, V.; Orkisz, M.; Norrlof, M.; Butail, S. Data-driven gearbox failure detection in industrial robots. *IEEE Trans. Ind. Inform.* **2019**, *16*, 193–201. [CrossRef]
6. Cerquitelli, T.; Nikolakis, N.; O'Mahony, N.; Macii, E.; Ippolito, M.; Makris, S. (Eds.) *Predictive Maintenance in Smart Factories*; Information Fusion and Data Science; Springer: Singapore, 2021. [CrossRef]
7. Bittencourt, A.C.; Saarinen, K.; Sander-Tavallaey, S. A Data-driven Method for Monitoring Systems that Operate Repetitively-Applications to Wear Monitoring in an Industrial Robot Joint1. *IFAC Proc. Vol.* **2012**, *45*, 198–203. [CrossRef]
8. Sathish, V.; Ramaswamy, S.; Butail, S. Training data selection criteria for detecting failures in industrial robots. *IFAC-PapersOnLine* **2016**, *49*, 385–390. [CrossRef]
9. Trung, C.T.; Son, H.M.; Nam, D.P.; Long, T.N.; Toi, D.T.; Viet, P.A. Fault detection and isolation for robot manipulator using statistics. In Proceedings of the 2017 International Conference on System Science and Engineering, Ho Chi Minh City, Vietnam, 21–23 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 340–343. [CrossRef]
10. Chen, T.; Liu, X.; Xia, B.; Wang, W.; Lai, Y. Unsupervised Anomaly Detection of Industrial Robots Using Sliding-Window Convolutional Variational Autoencoder. *IEEE Access* **2020**, *8*, 47072–47081. [CrossRef]
11. Wen, X.; Chen, H. Heterogeneous Connection and Process Anomaly Detection of Industrial Robot in Intelligent Factory. *Int. J. Pattern Recognit. Artif. Intell.* **2020**, *34*, 2059041. [CrossRef]
12. Hong, Y.; Sun, Z.; Zou, X.; Long, J. Multi-joint Industrial Robot Fault Identification using Deep Sparse Auto-Encoder Network with Attitude Data. In Proceedings of the 2020 Prognostics and Health Management Conference (PHM-Besançon), Besancon, France, 4–7 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 176–179. [CrossRef]
13. Jaber, A.A.; Bicker, R. Industrial Robot Backlash Fault Diagnosis Based on Discrete Wavelet Transform and Artificial Neural Network. *Am. J. Mech. Eng.* **2016**, *4*, 21–31.
14. Nentwich, C.; Junker, S.; Reinhart, G. Data-driven Models for Fault Classification and Prediction of Industrial Robots. *Procedia CIRP* **2020**, *93*, 1055–1060. [CrossRef]
15. Cheng, F.; Raghavan, A.; Jung, D.; Sasaki, Y.; Tajika, Y. High-Accuracy Unsupervised Fault Detection of Industrial Robots Using Current Signal Analysis. In Proceedings of the IEEE International Conference on Prognostics and Health Management (ICPHM), San Francisco, CA, USA, 17–20 June 2019.
16. Kim, Y.; Park, J.; Na, K.; Yuan, H.; Youn, B.D.; Kang, C.S. Phase-based time domain averaging (PTDA) for fault detection of a gearbox in an industrial robot using vibration signals. *Mech. Syst. Signal Process.* **2020**, *138*, 106544. [CrossRef]
17. Pu, Z.; Cabrera, D.; Bai, Y.; Li, C. A one-class generative adversarial detection framework for multifunctional fault diagnoses. *IEEE Trans. Ind. Electron.* **2021**. [CrossRef]
18. Pimentel, M.A.; Clifton, D.A.; Clifton, L.; Tarassenko, L. A review of novelty detection. *Signal Process* **2014**, *99*, 215–249. [CrossRef]
19. Braei, M.; Wagner, S. Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art. *arXiv* **2020**, arXiv:2004.00433.
20. Bishop, C.M. *Pattern Recognition and Machine Learning*, corrected at 8th printing 2009 ed.; Springer: New York, NY, USA, 2009.
21. Kemmler, M.; Rodner, E.; Wacker, E.S.; Denzler, J. One-class classification with Gaussian processes. *Pattern Recognit.* **2013**, *46*, 3507–3518. [CrossRef]
22. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF. *ACM SIGMOD Rec.* **2000**, *29*, 93–104. [CrossRef]
23. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 413–422. [CrossRef]
24. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, Portland, OR, USA, 2–4 August 1996; AAAI Press: Palo Alto, CA, USA 1996; pp. 226–231.
25. Kumar, V.; Banerjee, A.; Chandola, V. Anomaly Detection for Symbolic Sequences and Time Series Data. 2009. Available online: https://conservancy.umn.edu/bitstream/handle/11299/56597/Chandola_umn_0130E_10747.pdf;jsessionid=A026AFB0 208E56D91DD811BA4A134056?sequence=1 (accessed on 4 November 2021).
26. Aggarwal, C.C. *Outlier Analysis*, 2nd ed.; Springer International Publishing: Cham, Switzerland, 2017. [CrossRef]
27. Chauhan, S.; Vig, L. Anomaly detection in ECG time signals via deep long short-term memory networks. In Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, France, 19–21 October 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–7. [CrossRef]
28. Munir, M.; Siddiqui, S.A.; Dengel, A.; Ahmed, S. DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. *IEEE Access* **2019**, *7*, 1991–2005. [CrossRef]

29. Zhang, R.; Zhang, S.; Muthuraman, S.; Jiang, J. One Class Support Vector Machine for Anomaly Detection in the Communication Network Performance Data. 2007. Available online: http://www.wseas.us/e-library/conferences/2007tenerife/papers/572-618.pdf (accessed on 4 November 2021).
30. Sharma, S.; Swayne, D.A.; Obimbo, C. Trend analysis and change point techniques: A survey. *Energy Ecol. Environ.* **2016**, *1*, 123–130. [CrossRef]
31. Kendall, M.G.; Gibbons, J.D. *Rank Correlation Methods*, 5th ed.; Arnold: London, UK, 1990.
32. Cox, D.R.; Stuart, A. Some Quick Sign Tests for Trend in Location and Dispersion. *Biometrika* **1955**, *42*, 80. [CrossRef]
33. Crawford, C.G.; Slack, J.R.; Hirsch, R.M. *Nonparametric Tests for Trends in Water-Quality Data Using the Statistical Analysis System*; USGS Numbered Series; U.S. Geological Survey: Reston, VA, USA, 1983. [CrossRef]
34. DURBIN, J.; WATSON, G.S. Testing for Serial Correlation in Least Squares Regression. I. *Biometrika* **1950**, *37*, 409–428. [CrossRef]
35. Melek, W.W.; Lu, Z.; Kapps, A.; Fraser, W.D. Comparison of trend detection algorithms in the analysis of physiological time-series data. *IEEE Trans. Bio-Med Eng.* **2005**, *52*, 639–651. [CrossRef]
36. Modi, S.; Lin, Y.; Cheng, L.; Yang, G.; Liu, L.; Zhang, W.J. A Socially Inspired Framework for Human State Inference Using Expert Opinion Integration. *IEEE/ASME Trans. Mechatron.* **2011**, *16*, 874–878. [CrossRef]
37. Carvalho Bittencourt, A. *Modeling and Diagnosis of Friction and Wear in Industrial Robots*; Linköping University Electronic Press: Linköping, Sweden, 2014; Volume 1617. [CrossRef]
38. Nentwich, C. CATS—A Combined Anomaly Detection and Trend Detection Model. Available online: https://github.com/xorbey/CATS_public (accessed on 4 November 2021).
39. Danielson, H.; Schmuck, B. Robot Condition Monitoring: A First Step in Condition Monitoring for Robotic Applications. Master's Thesis, Lulea University of Technology, Lulea, Sweden, 2017.
40. Karlsson, M.; Hörnqvist, F. Robot Condition Monitoring and Production Simulation. Master's Thesis, Lulea University of Technology, Lulea, Sweden, 2018.