Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 00 (2018) 000–000

www.elsevier.com/locate/procedia

12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 18-20 July 2018, Gulf of Naples, Italy

Anomaly Detection in Discrete Manufacturing Using Self-Learning Approaches

Benjamin Lindemann*[a], Fabian Fesenmayr [a], Nasser Jazdi [a], Michael Weyrich [a]

[a]*Institute of Industrial Automation and Software Engineering, University of Stuttgart, Pfaffenwaldring 47, 70550 Stuttgart, Germany*

* Corresponding author. Tel.: +49 711 685 67321; fax: +49 711 685 67302. *E-mail address:* Benjamin.lindemann@ias.uni-stuttgart.de

Abstract

Process anomalies and unexpected failures of manufacturing systems are problems that cause a decreased quality of process and product. A better understanding of the system's behavior with the aid of data is the key to improve reliability and process stability. Current data analytics approaches show decent results concerning the optimization of single processes but lack in extensibility to plants with high-dimensional data spaces. This paper presents and compares two data-driven self-learning approaches that are used to detect anomalies within large amounts of machine and process data. Models of the machine behavior are generated to capture complex interdependencies and to extract features that represent anomalies. The approaches are tested and evaluated on the basis of real industrial data from a metal forming process.

Keywords: Unsupervised Learning; LSTM Autoencoder; Anomaly Detection; Predictive Maintenance;

1. Introduction

An ongoing increase in the storage density of storage devices interrelated with a continuing price reduction enable manufacturers to store a huge amount of process and machine data. However, in most of the cases the measured data is either partially evaluated or not analyzed at all. A manual analysis of the data is overwhelmingly time-consuming and almost impossible in case of large data sets ('Big Data'). Machine learning algorithms have proven to be one of the most promising approaches to extract unknown relations and knowledge within large data sets. These methods are used in a wide range of industries to generate additional benefits from the recorded data.

With regard to industrial automation and manufacturing systems, a similar trend can be observed. Automation systems are equipped with an increasing number of sensors to monitor the whole process. The archived but unused data can be utilized as training data. For instance, historical process data of production plants can be investigated to predict future maintenance intervals and to maximize the availability of the plant ('Predictive Maintenance'). Another application would be the determination of the optimal system parameters and the according quality prediction of workpieces to be produced ('Predictive Quality'). As a consequence, the amount of degraded products can be minimized leading to an increased quality of the manufacturing process.

In the scope of the present paper, two algorithmic approaches for an automated and data-driven anomaly detection have been applied to show the described applicability of machine learning algorithms to industrial machine and process data. Based on the algorithms, two concepts are developed to cope with high-dimensional and mostly unlabeled time series data from an industrial wheel press that is used in a real manufacturing environment. Both real anomalies and simulated ones based on expert knowledge are available for a test phase of the learned models. Hence, implemented prototypes are used to evaluate which of the methods leads to the best detection performance. This paper is organized as follows. Chapter 2 gives a brief overview of the state of the art. The two data mining algorithms, namely the k-means clustering and the Long-Short Term Memory (LSTM), are described in chapter 3 [1]. These algorithms have been extended and adapted to cope with unlabeled, high-dimensional time series data. Consequently, two approaches for anomaly detection are proposed. The investigated data set, the data integration and preprocessing operations as well as the implementation and the empirical results of the two approaches are presented in chapter 4. They are evaluated based on both real as well as simulated anomalies. The paper is concluded in chapter 5.

2. State of the Art

There have been various data-driven approaches to detect anomalies occurring in manufacturing and automation systems. In [2] a feature learning approach is used to detect anomalies of a gas turbine. To extract features from high-dimensional data, a stacked denoising autoencoder is used. On this basis, 27 sensor values could be reduced to significant features. The extreme learning machine is then applied on the extracted features to build a detection model. The approach showed solid results but demands for labeled data. An approach based on unsupervised learning is presented by [3]. The aim is to extract a state space model for hybrid timed automation systems. Different combinations of discrete signals are used to define states. The developed algorithm creates a state tree from measured data samples and merges it to generate a compact representation of the system states. Anomalies are detected based on major deviations from the learned model. The states itself contain continuous signal parts but this process data is not used to generate the discrete state model. In our case, this type of discrete signals is not available. In [4] a condition monitoring of a metal-working manufacturing system is realized. The developed algorithm is based on recurrent networks and simultaneously incorporates healthy machine data and the data to be analyzed. This special architecture helped to capture time dependencies within the time series data. In [5] an LSTM based network is combined with a support vector machine. The approach is used for anomaly prediction. The LSTM is applied to preprocess the data sequences and the SVM to identify anomalies. LSTM cells showed decent results to cope with time series data but are not able to solve dimension reduction problems. Another LSTM approach is presented in [6]. The network contains recurrent structures and is capable of learning long term dependencies with sparse representation. The Gaussian distribution is applied as anomaly detection metric. The approach showed decent results on four existing public data sets from different applications. Large amounts of industrial process data and a reduction of dimension was not part of the investigation. Hence, the development and evaluation of algorithmic approaches considering these untackled aspects will be the main focus of the following paper.

3. Data Mining Approaches

The present paper focuses on two approaches that are based on unsupervised learning techniques, namely the k-means clustering algorithm and the Long-Short Term Memory.

3.1. k-Means based Approach Using Sliding Windows

The k-means algorithm is an unsupervised learning method that generates data clusters with similar properties. Similarity is thereby characterized using distance metrics. The number of possible clusters must be defined before training. The input data is in most cases high-dimensional data where each data point is characterized by a multitude of features. This creates a high-dimensional input data space in which each data point can be represented by a vector. The algorithm aims to reduce the distance metric between the cluster centers and the acquired machine and process data. Thus, the final clustering result is characterized by a maximized homogeneity within the clusters as well as a maximized heterogeneity between the clusters. Hence, the overall optimization problem can be described as follows:

$$c^{(i)} := \arg\min_{j} || x^{(i)} - \mu_j ||^2 \tag{1}$$

The index i corresponds to the number of recorded data samples $x^{(i)}$ and index j to the number of clusters μ_j [7]. The number of potential clusters is not known but needed as an input parameter to the algorithm to conduct the training. In our case, the elbow method is utilized for the determination of an optimal number of clusters. The present approach extends the k-means algorithm due to the fact that industrial machine and process data is mostly time series data. A time-sensitive variant is created by artificially increasing the feature space on the basis of a sliding window function. As a result, observation periods of varying length can be defined by adapting the number of training features.

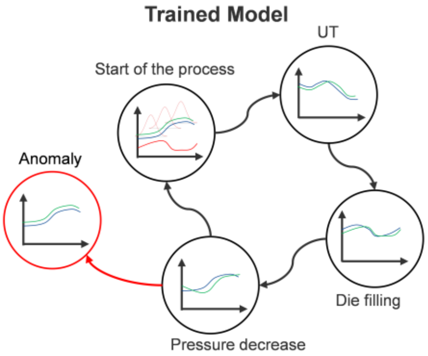

Fig. 1. Detection of anomalies through creation of new clusters.

However, increasing the size of the batch cannot be done arbitrarily due to the fact that it leads to an increased training time [8]. The k-means algorithm is assumed to be trained on a large, representative data set that reflects the healthy behavior of the underlying process. According to the idea of [3], a state space model is applied to describe the system. Based on this, the learned time-sensitive cluster structure is used as model for the system behavior. This requirement allows to pursue the following anomaly detection approach: the distance between a test data set consisting of currently acquired data and the cluster centers of the trained model is calculated with regard to the current point in time. Anomalies in the test data set are detected by a defined time-dependent limit violation to the cluster centers as well as the emergence of new, previously non-existent clusters. This circumstance is visualized in figure 1 where each cluster represents a specific machine state during one cycle of the manufacturing process. Due to the slinking emergence of an anomaly, the distance between the trained model and new data increases over time. The time-dependent limit violation of the average distance to cluster centers is used as anomaly detection metric.

3.2. LSTM based Approach Using Autoencoder Structure

In case of time series data, multiple time steps have to be correlated. Hence, a special architecture of recurrent neural networks (RNN) is applied, namely the LSTM developed by [1]. The LSTM architecture has already been successfully applied to many different problems where time-dependencies are highly relevant (for instance: translation [9], speech synthesis [10], audio analysis [11], etc.). Figure 2 illustrates the used LSTM block. The main elements of a LSTM block and concurrent the major advantages compared to a recurrent neuron are the three gates and the memory cell. The three gates control the interaction with other blocks of the network. In our approach, we embedded these gates in the network structure in a particular way. The network input layer is artificially extended to incorporate multiple time sequences at a certain point in time. Thus, the compressed information of hidden layers show an increased time-sensitivity and the gates do not transform single but multiple time sequences. Thus, the forget gate $\vec{f}$ can be described as follows:

$$f_j(t) = \sigma(\sum_{k=0}^{m-1} w_{kj} x_k(t) + \sum_{i=0}^{n-1} w_{ij} h_i(t-1) + b_j) \qquad (2)$$

where σ is the sigmoid function, m and n the number of weighted connections w to the input gate and the recurrent gate. It can be controlled which information is added to the memory and in which degree it is added. The memory itself can be mathematically described as multidimensional state vector $\vec{s}$ at a certain point in time capturing all relevant information. This vector changes depending on new information that is available through the gates. In our case, the state change can be described as follows:

$$\vec{s}(t) = \vec{f}(t) * \vec{s}(t-1) + \vec{a}(t) \qquad (3)$$

where $\vec{a}(t) = \vec{i}(t) * inv(\vec{f}(t))$ (4)

and $\vec{i}(t) = \tanh(W\{\vec{x}(t); \vec{h}(t-1); \vec{b}\})$ (5)

represent the influence of the add gate $\vec{a}$ and $\vec{i}$ is calculated as in (2) but with tanh. The forget gate controls whether the value of the memory cell will be discarded. In our cell architecture, the add gate is realized as inverse of the forget gate.

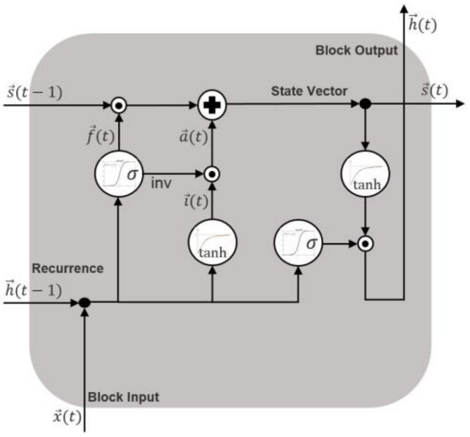

Fig. 2. Architecture of the applied LSTM cells - including inversion between forget and add gate.

Consequently, the oblivion always goes along with the addition of new information. This is useful in case of analyzing new time sequences. The output gate controls the output value of the LSTM block depending on the new values of memory cell vector [12]. LSTM blocks can be used within various RNN architectures. In our approach, the LSTM cells are integrated in an autoencoder structure to be able to process unlabeled data. The input and output layer of the autoencoder have the same dimension whereas the hidden layer consists of a smaller number of neurons. Thus, we reduce dimensions with our approach but don't explicitly use the hidden layer neurons as input for consecutive steps. The vector $\vec{y}$ represents the compressed information in the lower-dimensional data space. Our aim is to learn a process model that approximately maps the output vector $\vec{z}$ on the input vector $\vec{x}$. This reconstruction is conducted under minimized deviation, so that $\vec{y}$ captures a maximized entropy [13]. The autoencoder can be described as follows: Firstly, the deterministic function

$$\vec{y} = f_\theta(\vec{x}) = s(W\vec{x} + \vec{b}) \qquad (6)$$

maps the input vector $\vec{x}$ onto the hidden layer $\vec{y}$. Parameter W is the weight matrix of the connection weights between the neurons, $\vec{b}$ the bias vector with the connection weights of the bias neuron and θ the abbreviation for all parameters to be learned. For the projection we apply a sigmoid activation function as follows:

$$s(x) = f_{act}(x) = \frac{1}{1+e^{-x}} \qquad (7)$$

Secondly, the output vector is reconstructed using the compressed information of the hidden layer. The weight matrix corresponds to the transposed weight matrix W' of the input mapping. The training algorithm adapts the corresponding parameters of the model to minimize the average reconstruction error. Thus, our overall optimization problem is the following:

$$\theta^*, \theta'^* = \arg\min_{\theta,\theta'} \frac{1}{n}\sum_{i=1}^{n} L(x^{(i)}, g_{\theta'}(f_\theta(x^{(i)}))) \quad (8)$$

The function $g(x)$ maps from hidden to output layer, L corresponds to the loss function and θ^* as well as θ'^* determine the optimal network parameters of decoder and encoder. There exists a wide range of LSTM cell architectures. Our approach distinguishes itself through the fact that the gate vectors at a certain point in time do not consider single but multiple data sets. This is realized through the extension of the autoencoder structure that incorporates and reduces several time sequences at a time. Thus, the autoencoder structure is particularly adapted to enhance the time-sensitivity of the LSTM state vector. In order to evaluate the approach and to conduct an objective comparison, a quantitative anomaly detection metric has to be defined. The root of the quadratic error deviation between the input values and the reconstructed values is considered as suitable. Hence, the violation of a dynamic, time-dependent limit of this reconstruction error is used as anomaly detection metric.

4. Empirical Investigations in Discrete Manufacturing

4.1. Data Set and Data Integration

The investigated data set is acquired from a hydraulic press in real manufacturing environment of a German massive forming company. The press produces wheel rims for the automotive industry. The data set has a size of 2 TB and was measured over a time span of 4 months. The data space included up to 86 parameters of machine and process data captured in parts from the PLC and in parts directly from sensors. The sampling rate varied in the range of 50 μs to 10 ms. The press contains eight pumps that generate the total pressure of the press. Figure 3 shows exemplary the curve of one feature (here the pressure) of a pump. If one pump behaves abnormal, other pumps can compensate the anomaly so that the total output pressure remains at the desired level and the anomaly will not be visible. As a consequence, the process output is in tolerance but the heavily loaded pumps wear faster. This can result in an abrupt and unforeseeable failure of the whole system. The suspicion is the existence of correlations between parameters indicating the misconduct. Thus, an accurate prediction of the behavior of each subsystem on the basis of measured data could enhance the reliability. Therefore, all subsystem are equipped with a wide range of sensors. The aim is to predict anomalies based on a learned model that includes all subsystems. But before the dataset can be used for training and evaluation, the data has to be preprocessed and transformed. Thus, all features are normalized to achieve a faster and better learning result of the approaches. Due to the size of the dataset it is not possible to load the entire dataset in the working memory and train the algorithms accordingly. Instead, the algorithms are trained incrementally by separating and sequentially loading several data chunks. Both aforementioned approaches are developed to support online processing and are implemented accordingly so that they can be used for incremental training. The procedure of loading data chunks from the database is also conducted sequentially and during preprocessing operations.

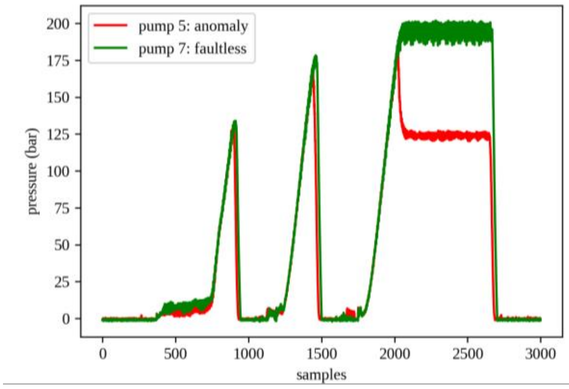

Fig. 3. Pressure curves of one press cycle from a healthy (green) and faulty pump (red).

4.2. Empirical Results of the First Approach

The first step of the analysis is the training of the k-means clustering based approach. An appropriate number of clusters is determined using the elbow-method. The training of the algorithm is conducted with GPU support on two Tesla K80 to enhance the calculation time. The optimal cluster centers are calculated during training. In our case, 6 major clusters are generated where each cluster represents a time-chronological machine state, namely press is inactive, pump pressure noise, swinging drawbar angles are minimal, pump pressure increase, maximal pressure and swinging drawbar angles close. To detected anomalies, a suitable anomaly metric has to be defined. Thus, the time-dependent mean Euclidian distance between the samples and the trained cluster centers is used. The basic idea of this metric is the fact that a major part of anomaly-free training samples is similar to one of the cluster centers. Consequently, the mean distance to the related cluster center is low in case of no anomalies. If the mean distance is higher than a time-dependent dynamic threshold, the data sample is considered as abnormal.

The described data set is used to evaluate the concept. Hence, the developed approach is trained with the features of three faultless pumps (pump: 1, 3, & 8). Each pump is associated with different features, for instance pressure, swinging drawbar angle, power supply or oil temperature. On this basis, the time-dependent metric is calculated for a period of time of 30 minutes. Subsequent, the middle pump and its associated features were replaced by another pump and the mean distance is recalculated for the new features. This training process is repeated for defined pumps and the result is shown in figure 4. It shows that all faultless pumps have almost congruent diagram curves. If the faultless pump 3 gets replaced by the pump 5 with the faulty pressure curve, the mean distance increases about 2% to 4%. Due to the fluctuating course, the increase can only be identified with a dynamic threshold or if the mean distance value is calculated over a longer period of time.

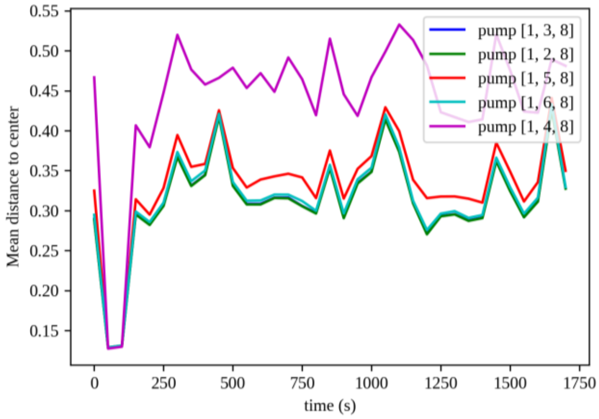

Fig. 4. k-means anomaly detection based on a data set from different hydraulic pumps.

This shows the capability of detecting anomalies with the k-means clustering based approach. If the faultless pump 3 is replaced by the defect pump 4, the mean distance is clearly increased and can be detected easily.

4.3. Empirical Results of the Second Approach

To generate a model based on the LSTM autoencoder approach, the data space of all faultless pumps is used. This training data must be transformed to time sequences due to the specific concept explained in 3.2. These sequences are cut to an identical length. During and after training, the algorithm incorporates one sequence as input and tries to reconstruct the values in this sequence as output. After the training, the learned model is applied to evaluate different test data scenarios. The test data is also transformed to time sequences according to those of the training data set. The reconstruction error is calculated over the considered test period and is thus a time-dependent metric. The reconstruction error curves over time are analyzed for each test dataset to evaluate if anomalies could be detected.

To compare the performance of both approaches, the training procedure that is applied to the LSTM autoencoder equals the procedure of the first approach. The corresponding results are shown in figure 5. The reconstruction error over one production cycle varies in a small range regarding all faultless pumps. If the faultless pump 3 gets replaced by the faulty pump 5 the reconstruction error significantly increases about 20% and can be predicted with the learned model. The comparison of this figure and figure 4 demonstrates that the fluctuation of all courses is significantly lower. This illustrates a higher detection sensitivity of the second approach. An additional improvement can be achieved by reducing the data space to active press cycles. On this basis, the anomaly prediction with the LSTM autoencoder could be improved due to the fact that a gradual threshold surveillance can be applied. The failure of the system is predicted based on the degree of deviation from the learned LSTM model. The error metric shows an exponential sensitivity in case of a linear deviation of the learned features. Hence, an early failure prediction is possible. Effects of the real anomaly could be detected after 1 day and with a confidence of 99% after 10 days. An actual failure of a subsystem occurred only after 4 months of slinking increase of the abnormal behavior.

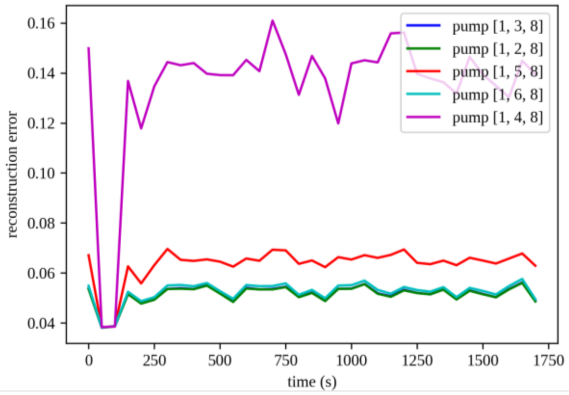

Fig. 5. LSTM autoencoder anomaly detection based on a data set from different hydraulic pumps.

To further evaluate this approach, simulated anomalies were added. These simulated anomalies represent different fault scenarios. Figure 6 shows the application of simulated anomalies to a faultless pump. They are exemplary illustrated for one feature, namely the pressure. These anomalies are created based on expert knowledge of historical failure events and a subsequent manual evaluation of the measured data. They are used to show the robustness of the learned model. The anomaly “Factor 0.7” multiplies the data with the factor 0.7. “Gaussian noise” adds an additive white Gaussian noise to the data. “Plateau Peaks” adds a peak to each pressure plateau during a press cycle. “Anomaly reconstruction 0.2” is the reconstruction of the real anomaly that is available within the data set. This corresponds to the real anomaly in an early stage. “Anomaly reconstruction 0.35” is identical to the anomaly before, but with a pressure drop of 35%.

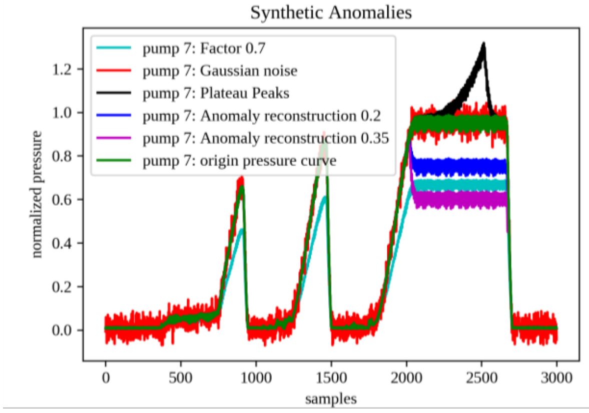

Fig. 6. Healthy course (green), real anomaly (violet) and artificial anomalies (other courses).

The same procedure as in the earlier mentioned chapters is applied to evaluate the prediction accuracy of the learned model. The simulated anomalies are incorporated into solely one subsystem, namely pump 3. Thus, nearly the entire measured data space is faultless except the features related to the abnormal subsystem. Furthermore, the intensity of the artificial anomalies increases over time. The resulting reconstruction error is shown in figure 7.

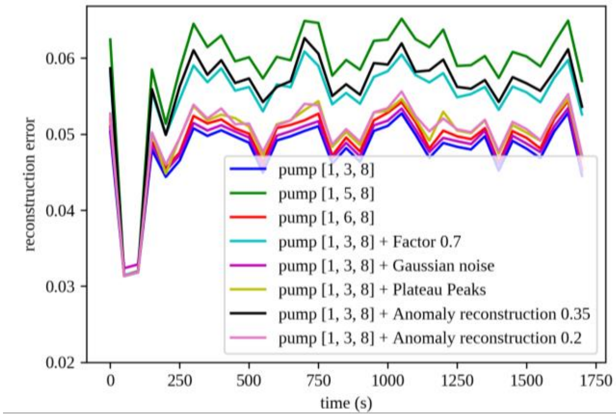

Fig. 7. LSTM autoencoder anomaly detection with simulated anomalies.

The real anomaly of pump 5 and the simulated anomalies “Factor 0.7” and “Anomaly reconstruction 0.35” can be clearly identified and thus early predicted. For instance, the future occurrence of the anomaly “Anomaly reconstruction 0.35” could be predicted with a confidence of 99% after 9 days slinking increase. The reconstruction error over time is much higher in comparison to the faultless curves. The same circumstance applies to the simulated anomalies “Anomaly reconstruction 0.2” and “Plateau Peaks” but they can only be detected if the mean reconstruction error is calculated over a longer period to smooth the curves. The anomaly “Gaussian noise” cannot be detected, because its reconstruction error is comparable to the faultless pumps 1, 3 & 8 and 1, 6 & 8. However, the occurrence of anomalies causing this type of vibrations and noise is very unlikely. The various evaluation tests show that anomalies could be detected reliably with the LSTM autoencoder, even if only a small number of features shows abnormal behavior.

5. Conclusion

The present paper investigated two approaches for anomaly detection of industrial manufacturing systems. The first approach based on a k-means clustering algorithm combined with sliding window techniques to capture time dependencies. The second approach utilizes a special architecture of LSTM cells that are embedded in an extended autoencoder structure. It is capable of reducing dimension and capturing time dependencies at once. The training data set is composed of a large amount of machine and process data from an industrial press. The data was acquired in real manufacturing environment of a German massive forming company. The empirical investigation of both approaches showed decent results concerning abrupt anomalies whereat the LSTM autoencoder approach was more accurate in case of slinking anomalies. An abrupt anomaly caused an average increase of the reconstruction error of 20%. The k-means based approach allowed a detection using a dynamic threshold of 2% to 4%. Furthermore, the learned LSTM model generated an exponentially increasing reconstruction error over time whereby the anomaly showed linear behavior. This strong sensitivity enabled the early detection of anomalies during operation, in the presented case approx.. 3.5 months before failure. Future investigations will focus on the test of different LSTM architectures to further optimize the detection accuracy.

Acknowledgements

This work was funded by the German Federal Ministry for Economic Affairs and Energy in the scope of the R&D program ‘Digital Technologies’ (PAiCE). We want to thank the Otto Fuchs KG for providing the data and supporting this work with their expert knowledge of the forming process.

References

[1] Hochreiter, S., Schmidhuber, J.: Long-Short Term Memory. In Neural Computation 9 (8): pp. 1735-1780, 1997.

[2] Yan, W., Yu, L.: On Accurate and Reliable Anomaly Detection for Gas Turbine Combustors: A Deep Learning Approach. In: Annual Conference of the Prognostics and Health Management Society. 2015, H. 6, pp. 1-8.

[3] Niggemann, O., Vodencarevic, A., Maier, A., Windmann, S., Kleine Büning, H.: A Learning Anomaly Detection Algorithm for Hybrid Manufacturing Systems. The 24th International Workshop on Principles of Diagnosis, October 2013.

[4] Kuo, C. J., Ting, K.-C., Chen, Y.-C.: State of product detection method applicable to Industry 4.0 manufacturing models with small quantities and great variety: An example with springs. In: Proceedings of the 2017 IEEE International Conference on Applied System Innovation. 2017, H. 1, pp. 1650-1653.

[5] Ergen, T., Mirza, A., Kozat, S.: Unsupervised and semi-supervised anomaly detection with LSTM neural networks. Eprint arXiv:1710.09207, Conrell University, October 2017.

[6] Malhotra, P., Vig, L., Shroff, G., Agarwall, P.: Long-Short Term Memory Networks for Anomaly Detection in Time Series. ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015.

[7] Jain., A.: Data clustering: 50 years beyond k-Means. Pattern Recognition Letters. Volume 31, Issue 8, 1 June 2010, pp. 651-666.

[8] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T. P.: On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In: ICLR 2017. 2017, pp. 1-16.

[9] Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., et al.: Addressing the Rare Word Problem in Neural Machine Translation, 2015.

[10] Fan, Y., Qian, Y., Xie, F., Soong, F. K.: TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks. In: Interspeech. 2014, pp. 1964-1968.

[11] Marchi, E., Ferroni, G., Eyben, F. et al.: Multi-resolution linear prediction based features for audio onset detection with bidirectional LSTM neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014, pp. 2164-2168.

[12] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., Schmidhuber, J.: LSTM: A Search Space Odyssey. In: IEEE Transactions on Neural Networks and Learning Systems. 2017, H. 10, pp. 2222-2232.

[13] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A.: Extracting and Composing Robust Features with Denoising Autoencoders, In: Proceedings of the 25th International Conference on Machine Learning. 2008, pp.1096-1103.