Predictive Maintenance: Using Recurrent Neural Networks for Wear Prognosis in Current Signatures of Production Plants

Thomas Küfner and Frank Döpper
University of Bayreuth, Chair of Manufacturing and Remanufacturing Technology, Bayreuth, Germany
Email: thomas.kuefner@uni-bayreuth.de, frank.doepper@uni-bayreuth.de

Daniel Müller and André Gerhard Trenz
Fraunhofer IPA, Project Group Process Innovation, Bayreuth, Germany
Email: daniel.mueller@ipa.fraunhofer.de, andre.trenz@ipa.fraunhofer.de

***Abstract*—On the way to Industry 4.0, the digitization, networking and automation of industrial plants are the core challenges for manufacturing companies. These developments are accompanied by a rapid increase in the amount of available data, which often remains unused, resulting in waste in the sense of lean production. One reason for this is the great effort involved in the subsequent analysis of large amounts of data. In this context, predictive maintenance is a promising means of benefiting from data in terms of early wear prediction. In order to implement predictive maintenance, the approach presented here uses machine learning methods to generate a model for wear and plant status detection and, based on this, an algorithm for wear prediction. Only current signatures of production facilities are used for this. These signatures are available in every electrical system, have a high information content and can be measured with minimal effort and expense. Following the CRISP-DM methodology, a short-time Fourier transform is applied to the continuously acquired current signatures in order to extract features. In the modeling phase, recurrent neural networks are trained with these features. To create the right conditions, the current signatures are generated with a test setup for wear simulation, which is also used for the evaluation and verification of the developed models and algorithms. Especially in the area of critical wear, the trained recurrent neural network models provide correct classifications with an accuracy of over 95 percent. The developed algorithm for predictive maintenance therefore delivers reliable wear forecasts so that maintenance can be planned at an early stage. Finally, the models and algorithms are implemented and tested in a developed embedded system to perform wear detection and prediction at the machines edge in almost real-time.**

***Index Terms*—Current signature analysis, lean data, predictive maintenance, recurrent neuronal networks, short-time Fourier transform**

I. Introduction

There is a general awareness that the efficient processing of the exponentially growing volume of data in Industry 4.0 represents is a decisive business success factor and secures competitive advantages for manufacturing companies [1], [2]. In practice, however, the analysis and profitable use of the accrued data are one of the main challenges for companies to be mastered while implementing an Industry 4.0 strategy [3]. One of the most important technologies for analyzing large amounts of data is machine learning [2]. Due to its diverse applications, machine learning, which enables the automated generation of knowledge from data [4], [5], developed from a research object to one of the most important universal technologies, among other in Industry 4.0 [2], [6]. According to a recent study, 73 percent of German companies are already focusing on machine learning, which will increasingly establish itself as a standard [7]. There are many possible applications in production. From a process and product perspective, machine learning is used, for instance, to optimize processes or product design and to predict product quality. For machines and plants, it brings significant advantages for predictive maintenance [8]. Predictive maintenance is a core component and key innovation contributing to the concept of Industry 4.0 [9], [10]. One of the technical challenges that companies must overcome for a successful implementation of predictive maintenance is the selection and evaluation of data [11]. The use of electrical current signatures as a database offers great potential for – in the sense of predictive maintenance – determining the wear of production plant [12]. In order to handle the constantly accumulating volume of data in production, solutions are needed that instantly derive the knowledge from the data (e.g. current signatures) with the help of machine learning in decentralized computing systems and thus reduce the forwarded data volume that can subsequently be directly used. We refer to this as the Lean Data approach [13].

This paper introduces a practical solution for the Lean Data approach by presenting an algorithm for decentral predictive maintenance with a developed edge device

Manuscript received January 25, 2021; revised April 15, 2021.
corresponding author: thomas.kuefner@uni-bayreuth.de

doi: 10.18178/ijmerr.10.11.583-591

using machine learning models based on current signatures. The current signatures are pre-processed using a short-time Fourier transform, so these can be used for more in-depth analyses and enable wear predictions. Furthermore, a model for wear detection is developed with machine learning methods and applied in an algorithm in order to verify the suitability of predictive maintenance by using current signature analysis on a brake system in a reproducible way. Recurrent neural networks are used for the modeling, as this machine learning approach enables the processing of sequential data [14]. The developed models are evaluated with the current signatures of a test setup, designed for the simulation of wear on brake systems. The algorithm is implemented and tested on the developed edge device for the purposes of practical evaluation.

II. Related Work

The suitability of current signatures for wear detection and prediction has been scientifically proved. For example, in their development of a predictive maintenance approach for turbogenerators, Pellicel et al. highlight the fact that current signatures can contain information about electrical or mechanical problems. Monitoring the frequencies and their changes in the current signatures makes it possible to draw conclusions about the evolution of the operating conditions of a machine. [15] Position figures and tables at the tops and bottoms of columns.

The empirical evidence of a correlation between motor current and (tool) wear is provided by Li et al. Using the spindle motor current and the feed drive current as a baseline, wear detection via current signatures is proven and the advantages of this method are elaborated. In addition to its simple application and effectiveness, the authors also include the low cost and the smooth machining process. [16]

Mandal summarizes the current state of research in his examination of the applicability of tool condition monitoring methods in conventional milling, specifically micromilling: there is scientific consensus that on the one hand the use of current signatures is characterized by the simplicity of implementation and by low costs. No external sensors are required, since the current of the drive can be used. On the other hand, Mandal emphasizes that current signatures are unaffected by mechanical noise as another advantage of this method. [17] Stavropoulos et al. also emphasize the low-cost aspect when using current signatures in their comparison of common condition monitoring methods [18].

Unlike Mandal, Klaic et al. consider tool wear during drilling. One advantage Klaic et al. point out when using current signatures for condition monitoring is that the machining process is not disturbed due to indirect measurement. On the contrary, Klaic et al. note that indirect methods – compared to direct methods – are generally less accurate. For example, the information in current signatures can be dependent on or affected by motor dynamics and motor temperature. [19]

Another positive aspect of current signatures as used in condition monitoring is provided by Praveenkumar et al. In a comparison with vibration signals, acoustic signals and current signatures for early fault diagnosis in gearboxes, the best results are achieved – regardless of the model choice – when current signatures are used as baseline data. [20]

An overview of further scientific publications dealing with condition monitoring of tools based on current signatures during drilling is given by Jantunen [21].

In summary, it can be said that:

- The suitability of current signatures for wear detection and prognosis is empirically proven,
- A more cost-effective use of current signatures for wear detection and prognosis can be stated compared to other indirect methods and
- Current signatures are well suited as baseline data, since modern production plants are almost exclusively electrically operated [22] (high data availability).

Despite the scientific proof of the suitability of current signatures for wear monitoring and the advantages mentioned in comparison to other indirect methods, it is not yet widely used in practice. Although first publications are already several years old and do not show any concrete obstacles or disadvantages, it is difficult to find a clear reason why this technique has not been pursued. While there are already initial applications, the method of wear detection and prediction in current signatures is still in the development stage [23]. This fact attests to the novelty and relevance of the research subject of this work.

III. Methodology

A. Description of the Experimental Setup

Fig. 1 shows the experimental test setup, designed and built for the reproducible simulation of wear on a brake system. The brake disc is driven by an asynchronous geared motor. The motor has a rated power of 180 Watt and a maximum speed of 1360 revolutions per minute. A worm gear with a gear ratio of 15:1 reduces the maximum speed to 90 revolutions per minute and delivers a maximum torque of 16 newton meter. The asynchronous motor is supplied with a voltage of 230 Volt via a frequency converter. The frequency converter allows variable control of the speed of the electric motor.

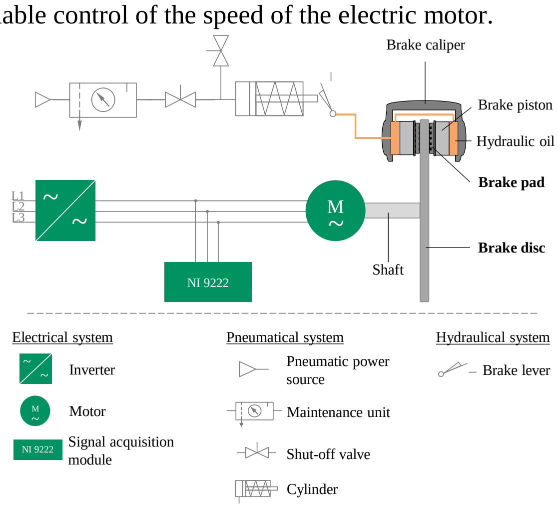

Figure 1. Schematic structure of the experimental test setup

In order to make the test results of the wear demonstrator reproducible and comparable, a pneumatic cylinder and a maintenance unit are installed. In order to be able to perform measurements with a defined braking force, the air pressure can be precisely specified with the help of the maintenance unit. The pneumatic cylinder actuates the brake lever with a known force resulting from the set pressure, which triggers the braking process. In this way, the force is transmitted hydraulically via the brake lever to the two brake pistons installed in the brake caliper. The braking process is initiated by pressing the brake pads against the brake disc. The test setup is suitable for simulating and analyzing wear on the brake disc on the one hand, and for examining brake pad wear on the other. This work is focused on brake pad wear.

A total of six brake pads with different *wear reserves* are considered. In addition to new brake pads (t_0 = 3.7 mm), brake pads with a thickness of t_1 = 3.5 mm, t_2 = 2.8 mm, t_3 = 2.6 mm, t_4 = 2.4 mm and t_5 = 2.2 mm are used (dimensions including backing plate). The wear of the brake pads analyzed in this work is measured by the two key figures *wear reserve* and *degree of wear*. The *wear reserve* indicates the brake pad thickness in millimeters that is still available until the brake pads need to be replaced.

$$\textit{wear reserve} = \textit{current brake pad thickness} - t_5 \quad (1)$$

With t_5 as the intervention limit. The *degree of wear* refers to the brake pad thickness still available and indicates, in an interval from 0 to 100 percent, how much the brake pads are worn. This available brake pad thickness is calculated from the difference between the original thickness of the brake pads (3.7 mm) and the brake pad thickness of the defined intervention limit (2.2 mm) amounting to 1.5 mm of usable brake pad.

$$\textit{degree of wear} = \frac{(1.5 - \textit{wear reserve})}{1.5} \cdot 100\,\% \qquad (2)$$

A 0 percent *degree of wear* means that the brake pad is new and not yet subject to wear. 100 percent *degree of wear* means that the intervention limit has been reached and the brake pads must be replaced.

To record the current signatures of the asynchronous motor, the current signals of the three output phases of the frequency inverter are each connected to a current transformer (Hall-effect current transformer HX 03-P from LEM). The NI 9222 module from National Instruments, which has a resolution of 16 bit, is used to digitize the measured signal. A sampling rate of 20 kHz is specified for the experiments. A value of 1.15 bar is selected for the air pressure and the speed is set to the maximum value of 90 revolutions per minute.

A total of three measurement series are carried out for the brake pads of each pad thickness. Between each measurement series, the brake pads of each thickness are reinstalled. This serves to ensure that the classification of the individual degrees of wear is actually based on the *degree of wear* and not, for example, on possible characteristics in the current signatures due to a possible deviation in the installation of the brake pads. For each installation, 30 braking operations with a duration of approximately 1-2 seconds were performed. Thus, each measurement series contains 30 braking processes. In total, current signatures from 90 braking operations were recorded for each brake pad thickness. Only the brake pad thickness is used to label the measurement series, regardless of the installation, in order to minimize the distortion effects described above. An overview of the experiments is shown in Table I.

TABLE I. DESIGN OF EXPERIMENTS

Series Number	Brake pad thickness in mm	Brake pad thickness number	Installation number	Number of braking operations
1	3.7	t_0	1	30
2	3.7	t_0	2	30
3	3.7	t_0	3	30
4	3.5	t_1	1	30
5	3.5	t_1	2	30
6	3.5	t_1	3	30
7	2.8	t_2	1	30
8	2.8	t_2	2	30
9	2.8	t_2	3	30
10	2.6	t_3	1	30
11	2.6	t_3	2	30
12	2.6	t_3	3	30
13	2.4	t_4	1	30
14	2.4	t_4	2	30
15	2.4	t_4	3	30
16	2.2	t_5	1	30
17	2.2	t_5	2	30
18	2.2	t_5	3	30

In order to counteract any possible influence of the engine temperature and dynamics, the measurements for brake pads of the same wear condition are not performed in direct succession.

B. Preparation of the Current Signatures

In addition to filtering and transforming the data to extract the relevant information, preparation also consists of formatting the data, which is required for a recurrent neural network. Fig. 2 shows the main steps of the data preparation.

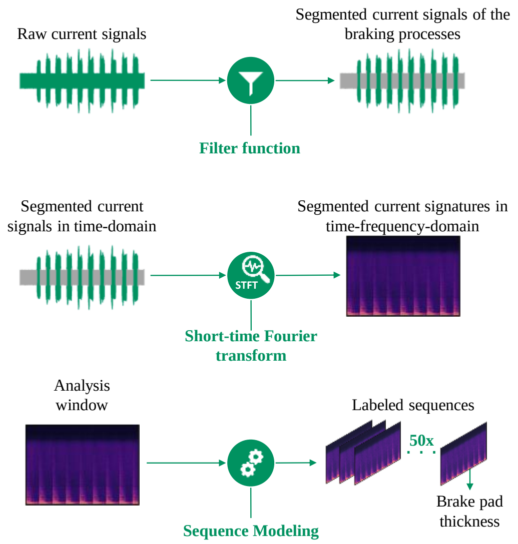

Figure 2. Main steps of current signature preparation (data pre-processing)

The approach in this section is based on the cross-industry standard process for data mining (CRISP-DM), an open standard process methodology for data mining and covers the phases of data understanding and data preparation.

First, the current signatures are filtered so that only the segments of the braking processes remain. The idle operation is completely removed. Based on a spectrum analysis, one can see that the current signatures are predominantly composed of low frequencies. In a second step, the filtered current signatures are transformed into the time-frequency domain using the short-time Fourier transform (STFT). This allows conclusions to be drawn about the changes in frequencies over time. The von Hann window is selected as the analysis window of the STFT. In addition to this input parameter of the function, the parameter for the overlap of the analysis windows is set to 50 percent of the window length. The length of an analysis window is 1024 measuring points. Each analysis window is labeled with the brake pad thickness of the transformed current signature. For the next step, a number of n analysis windows are required to view changes in spectral properties over time. A total of 50 analysis windows are combined to form a sequence. Starting with the fiftieth analysis window, the current analysis window is modeled as a sequence with the previous 49 analysis windows. Each generated sequence is labeled with the corresponding current brake pad thickness.

After the preparation of the current signatures, the data is split into training and test data in a ratio of 80 to 20. The training data is then used to train recurrent neural networks for brake pad thickness identification to determine the *degree of wear*.

C. *Description of the Analysis Model for Condition Detection*

This section covers the modeling phase of the CRISP-DM. To create the recurrent neural network, the Keras library in Python is used. Compared to the LSTM-architecture (Long Short-Term Memory), the GRU-architecture (Gated Recurrent Unit) achieved better results in a direct comparison and is therefore used in this work. Fig. 3 shows the topology of the recurrent neural network. By adding batch normalization layers, the training process is accelerated and overfitting is reduced.

```
model.add(GRU(1024, activation='relu', return_sequences=True,
input_shape= (num_steps, num_features)))
model.add(BatchNormalization())
model.add(GRU(512, activation='relu', return_sequences=True))
model.add(BatchNormalization())
model.add(GRU(512, activation='relu', return_sequences=True))
model.add(BatchNormalization())
model.add(GRU(256, activation='relu', return_sequences=True))
model.add(BatchNormalization())
model.add(GRU(256, activation='relu', return_sequences=True))
model.add(BatchNormalization())
model.add(GRU(128, activation='relu', return_sequences=True))
model.add(Flatten())
model.add(Dense(6, activation='softmax'))
```

Figure 3. Topology of recurrent neural network for wear classification

The standard activation function for GRUs is the tangent hyberbolic function. However, better results are obtained when rectified linear units (ReLU) are used, which is why – with the exception of the output layer – ReLU is assigned as the activation function to all layers of the recurrent neural network.

The number of neurons in the output layer corresponds to the number of brake pad thicknesses to be classified – i.e. six. The activation function is the softmax function, which is widely used for classification tasks. The output of the recurrent neural network is an array with six values. Due to the softmax activation function, each value corresponds to the probability of a class.

Adaptive moment estimation (Adam) is used as an optimization algorithm for training. The characteristic of this optimizer is its adaptive character, i.e. the dynamic adjustment of the learning rate in the course of training. The error function used is the categorical cross entropy, which is a common error function for classification tasks.

Due to the large amount of training data, the number of epochs is set to 4 when training the recurrent neural networks. The batch size is set to 32. During the iterative process, training processes with a batch size of 16 show that the error function does not converge and when choosing a larger batch size of 128, the model quality is not satisfactory.

D. *Evaluation of the Analysis Model for Condition Detection*

The approach in this step is based on the CRISP-DM and covers the phases of evaluation and deployment. The test data set is used to evaluate the trained model. This still unseen data is passed to the recurrent neural network and then classified with respect to the brake pad thickness. To ensure reproducibility, a total of five recurrent neural networks with constant topology are trained with identical training parameter settings. Since the results of the individual models differ from each other, the average correct classification rate of all five models is determined in Table II.

TABLE II. AVERAGE CORRECT CLASSIFICATION RATE (*CCR*) OF THE FIVE MODELS

Modell	1	2	3	4	5	Average
CCR in %	91.44	92.27	85.34	91.86	90.36	90.25

The confusion matrix of the average values of the trained recurrent neural networks is shown in Fig. 4. One can see that the probabilities of the individual classes vary. When the confusion matrices are viewed horizontally, no specific pattern can be identified that can be used to explain incorrect assignments to other classes. The highest correct classification rate of 97.89 percent is achieved with a brake pad thickness of 2.2 millimeters. The average probability of correct assignment across all classes is 90.25 percent.

It can also be seen that the average probability of correct classification increases with decreasing brake pad thickness. For example, the probability of predicting a brake pad thickness of 3.7 mm (t_0) – which corresponds to the original thickness of a new brake pad – is 85.35 percent. As the brake pad is worn down, i.e. with increasing wear, the average probability that the predicted value is correct also increases. Consequently, the prediction can be considered highly accurate in the critical range of the *degree of wear*, which is advantageous for the reliability of the subsequent algorithm towards the end of the lifetime. However, this pattern can only be observed when looking at the average values of the five models and does not necessarily occur with each of the trained recurrent neural networks.

	t_5	t_4	t_3	t_2	t_1	t_0		PREDICTED VALUES
	3912 18.98%	78 0.38%	11 0.05%	15 0.07%	1 0.01%	19 0.09%	4035 96.97% 3.03%	t_5
	3 0.02%	3040 14.75%	24 0.12%	60 0.29%	19 0.09%	4 0.01%	3151 96.69% 3.32%	t_4
	39 0.19%	220 1.07%	2809 13.63%	102 0.50%	19 0.09%		3189 89.28% 10.72%	t_3
		136 0.66%	213 1.03%	3127 15.17%	95 0.46%	21 0.10%	3592 89.78% 12.79%	t_2
	2 0.01%	208 1.01%	94 0.46%	44 0.21%	2970 14.41%	110 0.54%	3428 86.28% 13.08%	t_1
	40 0.20%	99 0.48%	61 0.30%	44 0.21%	227 1.10%	2743 13.31%	3214 85.35% 14.65%	t_0
	3996 97.89% 2.11%	3781 80.40% 19.60%	3212 87.45% 12.55%	3392 92.18% 7.82%	3331 89.16% 10.84%	2897 94.69% 5.31%	**20609** **90.25%** **9.75%**	
ACTUAL VALUES	t_5	t_4	t_3	t_2	t_1	t_0		

Figure 4. Confusion matrix of the average values of the five recurrent neural networks

E. Description of the Algorithm for Wear Prediction

The described model makes it possible to diagnose the present *degree of wear* of the brake pads for the six defined levels of wear and thus represents a condition monitoring system based on current signatures. In order to make predictions, the time-related component for determining the remaining useful life (*RUL*) is still missing. For this purpose, an algorithm is presented below in which the model is applied.

For an accurate prognosis, the wear behavior, and the *RUL* dependent on it, between the discrete wear stages must be described and determined. Due to the homogeneity of the brake pad material and the constant parameters of the test setup, such as engine speed and constant braking force, the wear behavior of the brake pads is assumed to be linear for the purposes of this work. Based on this premise, a *wear coefficient* is determined, which results from the wear of the brake pad thickness in millimeters related to a time interval. During the measurements in braking mode with a length of 122 s, an average of 0.027 mm of the brake pad thickness is removed. The *wear coefficient* is therefore calculated as follows:

$$\textit{wear coefficient} = \frac{0.027\ \text{mm}}{122\ \text{s}} = 0.22\ \frac{\mu\text{m}}{\text{s}} \tag{3}$$

The *wear coefficient* can be used to determine the progressive removal of *wear reserve* and reduction of *RUL* between the discrete stages. In addition to the recurrent neural network for wear classification, another recurrent neural network that determines the operating status of the test setup is used in the algorithm. The average correct classification rate of the recurrent neural networks for the classification of the operating status is 98.83 percent. The complete algorithm is shown and explained in Fig. 5.

The *degree of wear* and *RUL* are updated after each sequence. If the same discrete brake pad thickness has already been detected in condition monitoring, the *degree of wear* and *RUL* are recalculated using the *wear coefficient*. By continuously updating the *RUL* and *wear reserve*, wear can be predicted in order to plan and initiate maintenance measures at an early stage.

F. Test and Evaluation with a Developed Embedded System

In [24], Küfner et al. present a practical solution for vertical data continuity by combining signal acquisition with a microcontroller and simultaneous data analysis and evaluation with a single board computer in a decentralized embedded system without the use of time-consuming external cloud solutions. The models and algorithms presented here are implemented and tested in the embedded system design, introduced in [24] to perform wear detection and prediction at the machine's edge in almost real-time.

For the evaluation new brake pads (t_0 = 3.7 mm), as well as brake pads with a thickness of t_6 = 2.7 mm and t_7 = 2.3 mm, are used. The model is also used to classify the "idling" and "off" states in addition to the different brake pads. The values speed and pressure of the test setup are as described in III.A. The current signatures are measured again with the current transformer HX 03-P from LEM.

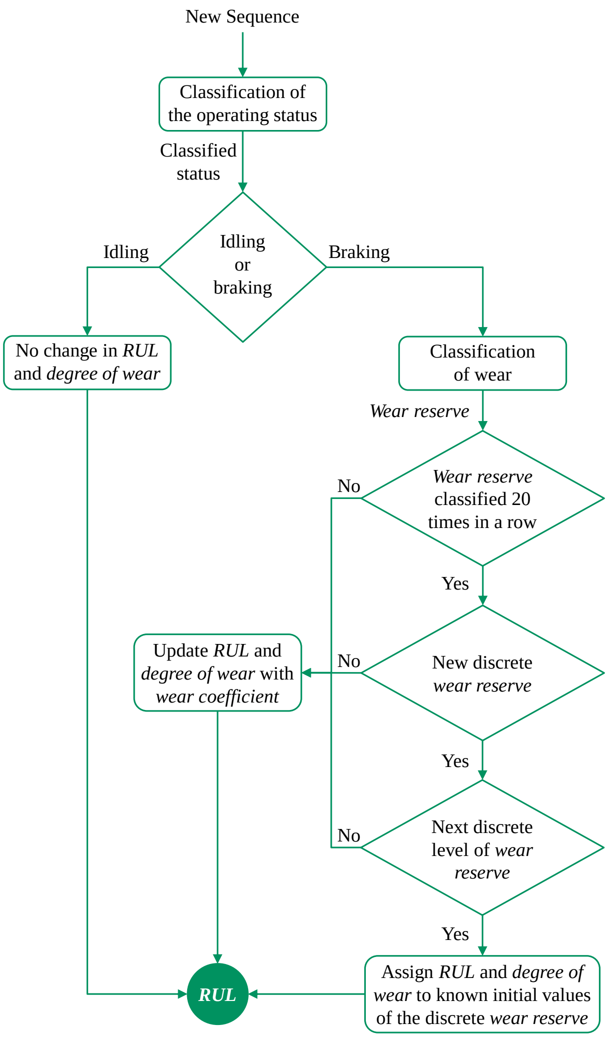

Figure 5. Flow diagram of the algorithm for wear prediction

In order to avoid potential correlations between the relevant information on the *degree of wear* in the current signatures and the influence of other factors, several series of measurements are also carried out here with newly installed brake pads. For each installation, 15 braking operations with a duration of approximately 2-3 seconds are performed. A total of four measurement series are therefore carried out for the brake pads of each pad thickness. In total, current signatures from 60 braking processes are measured for each pad thickness.

To realize the data pre-processing steps in almost real-time on the embedded system, these have to be adapted to the hardware architecture. Only the current signal of one phase is measured and analyzed by the embedded system. The phase conductor L1 was used in this setup. On the embedded system, the maximum sampling rate is limited to 12 kHz when recording one channel. An STFT is first performed with the microcontroller on the signal. The frequency resolution of the transformation and the time interval of a window are set respectively to 20 Hz and to 50 ms for easier assignment. The sampling rate is therefore reduced to 10.24 kHz and the number of points of a time window is halved from 1024 to 512. The resolution is reduced to the maximum possible value of 14 bit. The rectangular window is used as an analysis window instead of the von Hann window and there is no overlap. When modeling the STFT analysis windows of the current signatures to sequences, n-to-1 modeling is still used, but only 10 instead of 50 windows are combined to one sequence. Thus, each sequence is half a second long. To label the data, the RMS is used for a threshold analysis via the analysis window. This makes it possible to clearly distinguish the three states off, idling and braking. Since a new file is created for each measurement series and pad thickness, the respective pad thickness can be determined from the file name. As a basis for the recurrent neural network, the model described in III.C is used and adapted according to the changed parameters. These are the number of analysis windows per sequence, the features per analysis window and the number of output classes (five in total). The first three measurement series are used to train the model, the fourth measurement series is used to test and evaluate the model. Due to the reduced number of input values, more epochs are required to train the model properly. To exclude the occurrence of overfitting, 10 separate models are trained with identical parameters and training data. After each epoch, the loss of training and test data is determined and finally the respective median is formed for each epoch to mitigate outlier effects. As seen in Fig. 6, the minimum of the test loss is approximately at epoch 5, but even better results are achieved with individual models at epochs 9-11. The train loss converges to zero after about 15 epochs. The optimum number of epochs is therefore in the range of 5-10 and varies depending on the individual model. For the next steps, the number of epochs in training is set to 10.

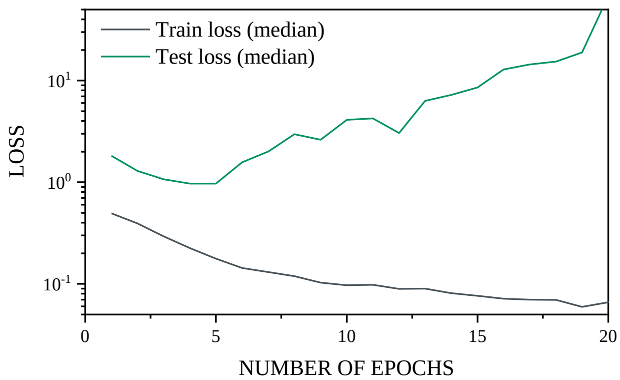

Figure 6. Loss functions of the training and test processes at the embedded system

All other training parameters are adopted identically. The model is trained on the single board computer of the embedded system, a Raspberry Pi 4 Model B (4 GB RAM). This process lasts approximately 12 hours.

After the preparation of the current signatures, the first three measurement series are split into training and test data in a ratio of 90 to 10. The training data is then used to train recurrent neural networks. The correct classification rate of the trained model is 94.61 percent for the training data and 93.87 percent for the test data. The classes “off” and “idling” have a correct classification rate of over 99 percent. The average probability of correct assignment across all classes is 94.53 percent. Fig. 7 shows the confusion matrix of the predicted values of the first three measurement series.

PREDICTED VALUES / ACTUAL VALUES	*off*	*idling*	t_7	t_6	t_0	
off	3987 21.65%		1 0.01%		1 0.01%	3989 99.95% 0.05%
idling	8 0.04%	3965 21.53%	10 0.05%	4 0.02%	2 0,01%	3989 99.40% 0.60%
t_7		10 0.05%	3195 17.35%	424 2.30%	75 0.41%	3704 86.26% 13.74%
t_6	1 0.01%	20 0.11%	103 0.56%	2995 16.26%	169 0.92%	3228 91.09% 8.91%
t_0	4 0.02%	5 0.03%	33 0.18%	138 0.75%	3266 17.73%	3446 94.78% 5.22%
	4000 99.68% 0.32%	4000 99.13% 0.87%	3342 95.60% 4.40%	3561 84.11% 15.89%	3513 92.97% 7.03%	**18416** **94.53%** **5.47%**

Figure 7. Confusion matrix of the values at the embedded system (measurement series 1-3)

The fourth measurement series is used to further evaluate the model with unseen data to ensure that the classification is actually based on the *degree of wear* and not on possible characteristics in the current signatures.

In the fourth measurement series, the motor was not turned to the "off" state. Therefore, no samples with this state exist. The average probability of correct assignment across all classes is then 98.60 percent. There are no specific patterns to explain the incorrect classifications. Consequently, a high accuracy of the classification can be stated across all classes. Fig. 8 shows the confusion matrix of the predicted values of the fourth measurement series for test and evaluation of the model.

PREDICTED VALUES / ACTUAL VALUES	*idling*	t_7	t_6	t_0	
idling	3761 68.47%				3761 100.0% 0.0%
t_7	7 0.05%	341 17.35%	6 0.11%	1 0.02%	355 96.06% 3.94%
t_6	42 0.76%	1 0.02%	439 7.99%		482 91.08% 8.92%
t_0	8 0.15%		12 0.22%	875 15.93%	895 97.77% 2.23%
	3818 98.51% 1.49%	342 99.71% 0.29%	457 96.06% 3.94%	875 99.89% 0.11%	**5493** **98.60%** **1.40%**

Figure 8. Confusion matrix of the values at the embedded system (measurement series 4)

For use on the edge device, the model is converted to a "TensorFlow Lite" model which speeds up the inference significantly. In this way, the model can be used with the embedded system to predict the current condition every 20 ms and the system can perform wear detection as well as condition monitoring in almost real-time.

IV. Conclusion and Outlook

The subject of this work is the realization of a decentral wear prognosis in electrical current signatures of production plant at the machines edge. Therefore, a test setup for the simulation of wear on brake pads is developed and tests with different wear conditions are carried out, whereby the current signatures of the electric motor of the test setup are measured. The collected data is processed to train recurrent neural networks that can classify six degrees of wear of the brake pads. The preprocessing step (data preparation) includes filtering, short-time Fourier transformation to the time-frequency domain, and sequence modeling. Subsequently, five identical recurrent neural networks were trained, resulting in reproducibility of the results. The average correct classification rate of the wear classification is 90.25 percent.

An analysis of the results of the models shows that the average probability for correct classification rises with increasing wear. This demonstrates that the models developed in this work are suitable for identifying wear in current signatures of production plants. In particular, the models also provide accurate information in the critical wear range, which is particularly important for ensuring high plant availability due to the short remaining useful life. The developed models are thus an essential basis for predictive maintenance and reliable wear forecasts, which enable early planning and initiation of maintenance measures.

To realize a predictive maintenance approach, an algorithm was developed to enable reliable wear predictions in addition to wear identification by the trained recurrent neural networks. In addition to the recurrent neural network for wear classification, another recurrent neural network that determines the operating status of the test setup is used in the algorithm. It is shown that the developed predictive maintenance algorithm provides a reliable wear prediction based on the key parameters "*degree of wear*" and "*remaining useful life*", generated in the algorithm. These two key parameters are elementary in order to be able to plan maintenance measures in advance.

To implement the approach on a decentralized embedded system, the algorithm is adapted to the hardware architecture and tests with different wear setups are carried out. The collected data is pre-processed and reduced. A model is trained for the execution on an edge device with an average correct classification rate of 98.60 percent. This solution can reliably detect the states "off" and "idling" and three degrees of wear of the brake pads. The system therefore performs wear detection on an edge device in almost real-time.

A future goal is to further increase the correct classification rate so that the statements of the models are even more reliable. Potential further optimization by increasing the number of GRUs or hidden layers is only possible to a limited extent due to the available resources, in particular due to the hardware at the machine's edge. Another restriction that arises due to unavailable resources is the number of analysis windows that are modeled into one sequence. Apart from using better resources, alternative ways to optimize the models need to be explored. These include, for example, the amount of training data available. In addition, the models must be capable of recognizing additional conditions, such as defective brake pads, faults on the brake disc or an oily or rusted brake systems. Moreover, the developed algorithm, as well as the models, must be transferred to and tested in other manufacturing applications.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Thomas Küfner conducted the research and was responsible for conceptualization, methodology, formal analysis, validation as well as project management, writing review and editing; Daniel Müller was competent for the experiment execution, data acquisition and analysis, model optimization, visualization as well as result preparation and presentation; André Gerhard Trenz transferred the models to the embedded system and was responsible for test and evaluation as well as optimization and visualization; Frank Döpper managed the project, was responsible for the supervision, funding acquisition and editing; all authors had approved the final version.

ACKNOWLEDGEMENTS

The presented research work is partially funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (StMWi) as supporters of the project “KI-noW – artificial intelligence to sustainably optimize value creation”. The authors are responsible for the contents of this publication.

REFERENCES

[1] M. Steven, Industry 4.0. Basics – subsectors – perspectives, Stuttgart: W. Kohlhammer, 2019.

[2] C. Herta, Data science and machine learning for Industry 4.0: possible applications and skilled labor requirements, In: Industry 4.0: Fundamentals and applications: Industry meeting of Berlin science and industry. Berlin, Vienna, Zurich: Beuth, pp. 27–34, 2015.

[3] G. Reinhart, Ed, *Handbook Industry 4.0. Business Models, Processes, Technology*, Munich: Hanser, 2017.

[4] Z. Les and M. Les, *Shape Understanding System – Knowledge Implementation and Learning*, Berlin: Springer, 2013.

[5] A. N. Fish, *Knowledge Automation. How to Implement Decision Management in Business Processes*, Hoboken: Wiley, 2012.

[6] W. Huber, Industry 4.0 compact – how technologies are changing our economy and our companies, Transformation and change of the entire company, Wiesbaden: Springer Vieweg, 2018.

[7] B. Reder. *Studie Machine Learning 2020*. [Machine learning study 2020]. [Online]. Available: https://www.lufthansa-industry-solutions.com/fileadmin/user_upload/dokumente/studie-machine-learning-2020-idg-de-lhind.pdf, Munich: IDG Business Media GmbH, 2020.

[8] J. Krauß, J. Dorißen, H. Mende, M. Frye, and R. H. Schmitt, Machine learning in production – application areas and freely available datasets, In: I40M, vol. 35, no. 4, pp. 39–42, 2019.

[9] C. Brückner, R. Bopp, and F. Krauss, Quality management. The practical handbook for the automotive industry, Munich: Carl Hanser, 2019.

[10] S. Feldmann, R. Laessig, O. Herweg, H. Rauen, and P. M. Synek, Predictive maintenance: Service of the future – and where it really stands. [Online]. Available: https://www.rolandberger.com/de/Insights/Publications/Predictive-Maintenance.html, Munich: Roland Berger GmbH, 2017.

[11] F. Duscheck, R. Blameuser, and S. Gehrmann. Machine availability up, maintenance and service costs down. Opportunities and challenges of predictive maintenance in industry. [Online]. Available: https://www.bearingpoint.com/files/BearingPoint_Studie_Maintenance_.pdf, Frankfurt/Main: BearingPoint GmbH, 2017.

[12] C. Nath, *Integrated Tool Condition Monitoring Systems and Their Applications: A Comprehensive Review*, Procedia Manufacturing, vol. 48, pp. 852–863, 2020.

[13] T. Küfner, T. H. J. Uhlemann, B. Ziegler, “Lean data in manufacturing systems: using artificial intelligence for decentralized data reduction and information extraction,” in *Proc. 51st CIRP Conference on Manufacturing Systems*, Sweden, 2018.

[14] A. Géron, Practical introduction to Machine Learning with Scikit-Learn and TensorFlow, Heidelberg: O'Reilly, 2018.

[15] A. Pellicel, G. Cássio, M. Lopes, L. E. Borges da Silva, E. L. Bonaldi, L. L. de Oliveira, J. G. Borges da Silva, G. Lambert-Torres, P. Rodrigues, *Intelligent Data Mining for Turbo-Generator Predictive Maintenance: An Approach in Real-World*, 2011.

[16] X. Li, J. Wang, S. K. Tso, “Real-time tool condition monitoring using wavelet transforms and fuzzy techniques,” in *Proc. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 30, pp. 352–357, 2000.

[17] S. Mandal, “Applicability of tool condition monitoring methods used for conventional milling in micromilling: A comparative review,” *Journal of Industrial Engineering*, vol. 2014, pp. 1–8, 2014.

[18] P. Stavropoulos, A. Papacharalampopoulos, E. Vasiliadis, G. Chryssolouris, “Tool wear predictability estimation in milling based on multi-senso-rial data,” *The International Journal of Advanced Manufacturing Technology*, vol. 82, pp. 509–521, 2016.

[19] M. Klaic, T. Staroveski, T. Udiljak, “Tool wear classification using decision trees in stone drilling applications: A preliminary study,” In: *Procedia Engineering*, vol. 69, pp. 1326–1335, 2014.

[20] T. Praveenkumar, M. Saimurugan, K. I. Ramachandran, “Comparison of vibration, sound and motor current signature analysis for detection of gear box faults,” *International Journal of Prognostics and Health Management*, 2017.

[21] E. Jantunen, “A summary of methods applied to tool condition monitoring in drilling,” *International Journal of Machine Tools and Manufacture*, vol. 42, pp. 997–1010, 2002.

[22] R. Neugebauer, Handbook resource oriented production, Munich: Hanser, 2014.

[23] ISO 13379-1:2012-05, Condition monitoring and diagnostics of machines – Data interpretation and diagnostics techniques – Part 1: General guidelines, Berlin: Beuth, 2012.

[24] T. Küfner, S. Schönig, R. Jasinski, A. Ermer, “Vertical data continuity with lean edge analytics for industry 4.0 production, In: Computers in Industry,” vol. 125, February 2021.

Thomas Küfner was born in Bayreuth (Germany) in 1986. He studied Renewable Energies at the University of Stuttgart and Energy Science and Technology at the University of Bayreuth and received his master’s degree in 2015. He is head of the working group Intelligent Value Creation with the Chair Manufacturing and Remanufacturing Technology at the University of Bayreuth and has an established background in the field of Data Science and Machine Learning in production, in which he has been active for several years.

Daniel Müller was born in Regensburg, (Germany) in 1995. He studied Business Administration and Engineering at the University of Bayreuth and received his bachelor´s degree in 2018 and his master's degree in 2020. Since 2020, he is a research assistant at the Fraunhofer IPA Project Group Process Innovation in Schweinfurt. His research focusses on Artificial Intelligence in production.

André Gerhard Trenz was born in Pegnitz (Germany) in 1996. He studied Engineering Science at the University of Bayreuth and received his bachelor's degree in 2019. He is a master´s student in Automotive and Mechatronics at the University of Bayreuth and a research assistant at the Fraunhofer Project Group Process Innovation in Bayreuth. His research interests include Data Science and Machine Learning in production.

Frank Döpper was born in Aachen (Germany) in 1970. He studied Mechanical Engineering and completed his PhD at the RWTH Aachen. During his almost 20-year industrial career he worked for several international engineering companies in managerial positions. Since 2017 he is professor of Manufacturing & Remanufacturing Technology at the University of Bayreuth and is head of the Fraunhofer Project Group Process Innovation in Bayreuth.