ORIGINAL ARTICLE

In-process detection of grinding burn using machine learning

Emil Sauter[1,2] · Erkut Sarikaya[2,3] · Marius Winter[2] · Konrad Wegener[1]

Abstract
The improvement of industrial grinding processes is driven by the objective to reduce process time and costs while maintaining required workpiece quality characteristics. One of several limiting factors is grinding burn. Usually applied techniques for workpiece burn are conducted often only for selected parts and can be time consuming. This study presents a new approach for grinding burn detection realized for each ground part under near-production conditions. Based on the in-process measurement of acoustic emission, spindle electric current, and power signals, time-frequency transforms are conducted to derive almost 900 statistical features as an input for machine learning algorithms. Using genetic programming, an optimized combination between feature selector and classifier is determined to detect grinding burn. The application of the approach results in a high classification accuracy of about 99% for the binary problem and more than 98% for the multi-class detection case, respectively.

Keywords Grinding burn · Machine learning · Process monitoring · Acoustic emission · Time-frequency transform

1 Introduction

In industrial manufacturing, grinding often is a final production step that must fulfill high demands on precision and surface integrity. On the other hand, the economic productivity of the grinding process has to be ensured by maintaining maximal tool life and shorten the process time. Furthermore, the consideration of the manufacturing process environmental impact gains more importance, leading into aims to reduce the usage of resources. In general, these three aspects contradict, as increasing the production speed results in a higher wear of the grinding tool, a higher energy input per time, and the amount of scrap parts increases.

As almost all of the introduced energy is converted into heat, the high temperatures can cause undesirable effects onthe surface integrity. These microstructural changes comprise phase transformation, residual stresses, and reduced fatigue strength. According to Wegener [1], these workpiece damages are referred to as grinding burn, regardless of their extent, visibility, and depth. Therefore, it is necessary to identify possible deteriorations of the surface integrity to guarantee the quality of the machined components.

There are several methods for the detection of grinding burn on the machined workpiece, categorized by He et al. [2] in beforehand prediction and post-mortem detection methods. The latter include time-intensive procedures like metallographic or residual stress testing. Following Wegener [1], in industrial applications, predominantly nital etching and Barkhausen noise analysis (BNA) are applied. In contrast, beforehand methods include the modeling of forces, temperatures, and power. Additionally, process monitoring approaches are applied to realize an in situ detection of thermal damages.

In research areas, acoustic emission (AE) signals are quite common for condition monitoring of grinding processes [3–5]. Multiple studies investigate effects like chatter vibrations [6], spindle damage [7], wheel wear monitoring [8], and the detection of grinding burn.

In the following, an approach for the in situ detection of grinding burn using machine learning is proposed. Therefore, an experimental procedure to generate controlled grinding burn is developed and a comprehensive dataset is ground using high material removal rates. After applying various time-frequency transforms and extracting statistical

✉ Emil Sauter
  emisauter@ethz.ch

1 Institute of Machine Tools and Manufacturing (IWF),
   Department of Mechanical Engineering, ETH Zürich,
   Leonhardstrasse 21, 8092 Zürich, Switzerland

2 Corporate Research, Robert Bosch GmbH,
   Robert-Bosch-Campus 1, 71272 Renningen, Germany

3 Institute of Production Management, Technology and
   Machine Tools (PTW), TU Darmstadt, Otto-Berndt-Strasse 2,
   64287 Darmstadt, Germany

features, the final feature matrix comprises 391 experiments and 878 features. Based on this feature matrix, the combination of feature selection and machine learning methods is optimized to ensure maximal classification accuracy for both binary and multi-class prediction. In addition, the selected and most relevant features are studied to evaluate the benefits of the different sensor types and signal processing methods.

2 State of the art

In recent years many researchers studied machine learning approaches for grinding burn detection. Even though they all pursued the same objective, each of them contributed with different procedures and techniques. The distinctive characteristics can be found in the sensor system, signal processing methods, feature engineering and feature selection/extraction techniques, and the applied classifier, as visualized in Table 1.

In the very early stages, Eda et al. [9] already studied the behavior between AE signals and corresponding grinding burn with different degrees. A general rise in AE amplitude was found with the progress of grinding burn. Furthermore, the fast Fourier transform (FFT) was used to conclude that the characteristic frequency band in the AE spectrum lies between 100 and 300 kHz. Saxler [10] and Webster et al. [23] confirmed the findings and highlighted the importance of the AE root mean square feature (RMS). Further statistical features were studied by Wang et al. [11], who concluded that the band power, kurtosis, skewness, and autoregressive coefficients of the AE signal show good correlations to the occurrence of grinding burn. Kwak and Ha [12] and de Aguiar et al. [13] implemented a multi-sensor method by including the tool power signal into the data acquisition. Neto et al. [14] also applied a multi-sensor method with AE and vibration signals. All of them used an artificial neural network (ANN) for the two-class classification problem and performed with a classification accuracy from 95 to 98.3%. The presented works so far either use time-domain or frequency-domain methods for grinding burn detection. However, these approaches are limited since they don't include methods that consider both domains at the same time.

On the assumption that purely time-domain or frequency-domain methods suffer under lack of robustness, Liu et al. [15] used the Wavelet Packet Transform (WPT) to create AE signal features. Distinguishing from other works, they provoked grinding burn by laser irradiation and demonstrated that the spectral density distribution of the AE magnitudes was shifted from lower to higher frequencies when heat was induced [15]. Chen et al. [16] used a similar experimental setup but applied the short-time Fourier transform

(STFT); nonetheless, they made similar observations. Griffin and Chen [6] studied grinding burn and grinding chatter with AE, power, vibration, and force signals whereby grinding burn was generated by adjusting process parameters. The observations of Griffin and Chen [6] that the characteristic frequency bands might reach over 500kHz coincides to some extent with the results of Liu et al. [15] but it should be considered that both use same typical aerospace workpiece material consisting of iron-nickel alloy. Their characteristic frequency bands may differ from those emitted by other materials such as steels utilized in the automotive industry, for example. They used the independent component analysis (ICA) for automatic feature selection and combined it with a tree-based genetic programming algorithm [6]. In a later study, Griffin [17] even obtained better results by using the classification and regression tree algorithm which promises 99 to 100% classification accuracy. Lajmert [18] introduced the Hilbert Huang transform (HHT) as a signal processing method to quantify grinding burn for machined hardened steel workpieces. In further work, Lajmert et al. [19] used AE, vibration, and force sensors focusing on automatic feature selection methods. They created many statistical features after applying HHT, reduced the number of features with principal component analysis (PCA), and established a minimum set of relevant features with decision tree algorithms Yang et al. [19]. Since the empirical mode decomposition (EMD) suffers under the mode-mixing phenomenon, Yang et al. [20] published the ensemble empirical mode decomposition (EEMD) method promising more robustness and better results. Similar to other works, a frequency shift to higher frequency bands was observed in the range of 150 to 450kHz [20]. Furthermore, Yang et al. [20] proposed, unlike Liu et al. [15], that the frequency range of 150 to 450kHz correlates with grinding burn, even though they also used laser irradiation to generate grinding burn. The main difference is that Yang et al. [20] used the middle carbon steel ASIS 1045 in his experiments, instead of the iron-nickel-based alloy. Hence, the prominent frequency bands for grinding burn might depend on the workpiece material; Guo et al. [21] applied the EEMD as well as the WPT for signal processing and used a two-stage feature selection algorithm to determine the most meaningful features. A stacked sparse autoencoder network (SSAE) was deployed to classify between burn and no burn and performed with 97.5% classification accuracy Guo et al. [21]. Another method for frequency mode decomposition is provided by the variational mode decomposition (VMD) and, in terms of grinding burn detection, first introduced by Gao et al. [22]. This method was motivated from the assumption that grinding burn is associated with the frequency distribution of the AE signal. Four degrees of grinding burn (no burn, slight burn, medium burn, and severe burn) were generated on the metallic workpiece

Table 1 Comparative overview of the state of the art

	Sensors	Sampling rate	Signal processing	Feature engineering	Feature selection/ feature extraction	Classifier	Performance
Eda et al. [9]	AE	n.i.*	FFT	PSD band power	No feature selection	–	"Good"
Saxler [10]	AE	≥ 4MHz	FFT	Statistical features	–	–	–
Wang et al. [11]	AE	2.56MHz	FFT	Band power, kurtosis, skewness, autoregressive coefficients	"Trial and error"	Neural network (radial basis function)	"Good"
Kwak and Ha [12]	AE	n.i.	FFT	Peak RMS, peak FFT	n.i.	Neural network	"About 95%"
	Power	10kHz	–	static & dynamic power			
de Aguiar et al. [13]	AE	n.i.	FFT	Statistical features	n.i.	Neural network	97%
	Power	n.i.	–	–			
Neto et al. [14]	AE	2MHz	FFT	RMS	–	Neural network	98.3%
	Vibr	2MHz	FFT	RMS	–		
Liu et al. [15]	AE	2MHz	WPT	Normalized energy	–	–	–
Chen et al. [16]	AE	5MHz	STFT	–	–	Neural network	–
Griffin and Chen [6]	AE	2MHz					
Griffin [17]	Power	n.i.	STFT	Statistical features	ICA	Genetic programming	80 to 95%
	Vibr	n.i.	WPT	Normalized energy		CART algorithm	99 to 100%
	Forces	n.i.					
Lajmert [18]	AE	≥ 0.6MHz	EMD, HHT	Statistical features	PCA	–	–
	Vibr						
Lajmert et al. [19]	AE	≥ 1MHz	HHT	Statistical features	PCA	Decision tree	100%
	Vibr	≥ 20kHz			Decision tree		
	Force	≥ 4kHz					
Yang et al. [20]	AE	2MHz	EMD,	Marginal spectrum RMS	–	–	–
	Vibr	2MHz	HHT	Spectral centroid			
Guo et al.[21]	AE	1MHz	WPT	Statistical features in time- and frequency-domain	Relief and autoencoder	SSAE, SAE, ANN, SVM	97.5% with SSAE
	Vibr	3.2kHz	EEMD				
	Force	3.2kHz					
Gao et al. [22]	AE	2MHz	VMD	RMS, center frequency	–	–	–

* n.i. no information

surface by laser irradiation with four different temperature ranges [22].

In the presented works about grinding burn detection, many different methods have been proposed. Purely time-domain and frequency-domain methods were used in the very early stages to determine relevant features and characteristic frequency bands for grinding burn detection. Especially in the last decade, time-frequency-domain methods became more popular since they contain more information and enable non-stationary signal analysis. However, after signal transformation, automatic feature selection algorithms are just rarely used for optimal feature subset selection. Furthermore, it was noticed that different workpiece materials exhibit considerably different properties for their frequency content during machining. The presented works show very good classification accuracies; however, most of them do not differentiate between different degrees of grinding burn and use a small database with few to no process variation. In addition, all of the described works either use a grinding wheel with alumina oxide grains or induced workpiece burn through laser irradiation.

Distinguishing from other works, this paper uses a metal-bonded grinding wheel with cBN grains and produces four different burn states by an emulated wear progress. Furthermore, the proposed approach applies modern signal transformation methods on multiple sensor signals. Next, the generated statistical features are reduced by the use of feature selection methods that automatically identify the most suitable features. Finally, the multi-sensor approach and the abundant amount of experimental data under varying process conditions enable multi-class prediction, which is capable to classify between different degrees of grinding burn.

3 Theoretical methods

The theoretical foundations of this study are introduced in the following. Thereon, the problem-specific application and implementation are presented in Section 4.5.

3.1 Signal processing

The most popular transformation techniques are introduced briefly in the following which are taken into account as a preprocessing step for ML. Table 2 shows a comparative overview of the Fourier, Wavelet, and Hilbert transform covering the major differing aspects with significant importance for this work.

The FFT transforms any time-domain signal $x(t)$ into the frequency-domain by convolution with periodic sinusoidal functions. Due to this assumption of periodicity, the FFT looses the time content and therefore is not suited for non-stationary signals. This limitation gave rise to the STFT

which implies, according to Isermann [25], a window function which captures the frequencies over the time by shifting a time constant τ. However, due to the time frequency uncertainty principle, the STFT is not able to provide an exact time-frequency representation of the signal [14].

To overcome this drawback, the wavelet transform was developed, which uses a group of mother wavelets $\Psi(t)$ showing sharp and oscillating behavior. The choice of wavelets is typically done according to the similarity to the original signal [25]. Once a mother wavelet is chosen, it can be time-scaled (adjusting the frequency resolution) by the factor a and time-shifted (adjusting the time resolution) by τ. Dilatation and translation of a mother wavelet $\Psi(t)$ leads to Eq. 1. [25].

$$\Psi'(t, a, \tau) = \frac{1}{\sqrt{a}} \, \Psi\left(\frac{t - \tau}{a}\right) \tag{1}$$

Hence, the wavelet transform provides an independent modulation of time and frequency resolution [15]. The continuous wavelet transform (CWT) of the signal $x(t)$ is then described as follows [25]:

$$W_{CWT}(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \, \Psi\left(\frac{t - \tau}{a}\right) \, d\tau \tag{2}$$

The wavelet theory can practically be used in combination with high-pass and low-pass filters to decompose a signal into different frequency bands. The wavelet packet transform (WPT) decomposes the signal into scaling (approximations) and wavelet (details) coefficients by convolution of the signal. Thus, the high-pass and low-pass filter impulse responses lead to a symmetric tree-structured filter bank [14].

The Hilbert transform (HT), introduced by Hahn [26], handles nonstationary and nonlinear narrow-band signals and is defined as

$$\mathcal{H}(x(t)) = H(t) = \frac{1}{\pi} \, P \, \sum_{-\infty}^{\infty} \frac{x(t')}{t - t'} \, dt' \tag{3}$$

where P describes the Cauchy principle value [27]. The HT describes a 90° phase shift of $x(t)$, forming a complex conjugate pair of $x(t)$ and $H(t)$ which can then describe an analytical signal $Z(t)$ such that

$$Z(t) = x(t) + i \, H(t) = A(t) \, e^{i\Phi(t)} \tag{4}$$

where $A(t)$ describes the instantaneous amplitude and $\Phi(t)$ the instantaneous phase [27, 28]. The instantaneous frequency can then be derived as follows.

$$\Omega(t) = \frac{d\Phi(t)}{dt} \tag{5}$$

Table 2 Comparative overview of Fourier, Wavelet and Hilbert transform, according to [24]

	Fourier		Wavelet		Hilbert	
Algorithms	FFT	STFT	DWT	WPT	(E)EMD	VMD
Nonstationary	No	Yes	Yes	Yes	Yes	Yes
Nonlinear	No	No	No	No	Yes	Yes
Frequency	Convolution		Convolution		Differentiation	
Time-Frequency Resolution	Global uncertainty		Regional uncertainty		Local certainty	

For the effectiveness of the HT, however, some limitations on the analyzed data $x(t)$ are required, e.g. it needs to be a narrow-band signal [27]. This is because the instantaneous frequency function is a scalar frequency value over time and is thus not able to describe more than one oscillatory mode at any given time [27]. For this reason, Huang et al. [27] introduced the empirical mode decomposition (EMD), which decomposes a wide-band signal into its narrow-band frequency components called intrinsic mode functions (IMF's). After that, the HT is applied to the decomposed components and then reconstructed. The combination is known as the Hilbert Huang transform (HHT). As an improvement to the EMD Wu and Huang [29] published the ensemble empirical mode decomposition (EEMD) as a truly noise-assisted data analysis method. This method builds an ensemble of white noise-added signal and takes the average as the final result to overcome the mode mixing problem which turned out as a major drawback of the EMD. As a very recent approach to decompose a signal into frequency bands, the variational mode decomposition (VMD) was proposed by Zosso and Dragomiretskiy [30]. The VMD describes an entirely non-recursive decomposition technique which extracts the modes concurrently. In contrast to the EMD or EEMD, it is mathematically well founded as a constraint variational problem, which assumes each mode to be mostly compact around its center frequency [30].

3.2 Feature engineering and data preprocessing

High-frequent sensor data typically have a low density of information, no matter if it is present in time- or frequency-domain. One way to compress the information density is provided by feature engineering, whereby statistical features are computed to represent the data as a collection of scalar values without loosing much information. After applying the signal transformation methods previously introduced, features based on time, frequency, and time-frequency information are calculated. According to Grimmett [31], the features usually include the central moments from statistics: mean, standard deviation (*std*), skewness (*skew*), and kurtosis (*kurt*). Especially for AE signals, additional features as the root mean square

(*rms*), the median (*med*), the peak values (*p2rm* and *p2p*) and the spectral entropy (*spen*) could be considered [10, 22, 23, 32]. An overview of the features with their corresponding equations used for fault detection and diagnosis was listed by de Aguiar et al. [13], Sharma and Parey [33], Teti et al. [34], and Guo et al. [21].

In this work, the abovementioned features are used to build the feature matrix $\mathbf{F} = (f_{i,j}) \in \mathbb{R}^{m \times n}$ with m experimental observations and n features. The absolute values $|f_{i,j}|$ strongly depend on the selected unit of the features. These different scales and also the varying sign effect that for example using a distance-based classifier the higher feature values would outweigh the lower ones. As a consequence, standardization or z-score normalization is conducted to preprocess the data [35], whereby the feature matrix $\mathbf{F}$ is scaled column by column. Following Guyon et al. [36], this step typically results in a better performance of the machine learning method applied afterwards.

3.3 Feature selection

Not all constructed features are necessarily useful for the classification task, since there might be also redundant or even irrelevant features. Bolón-Canedo et al. [37] describe the process of feature selection as a problem of finding the most informative and compact subset of features. The main difference to feature extraction methods like the principal component analysis (PCA) is that no new features are calculated and only a subset of the full feature matrix containing the most relevant dimensions is selected [37]. According to Bolón-Canedo et al. [37] and Pilnenskiy and Smetannikov [38], feature selection methods are divided into filter, wrapper, and embedded methods.

Filter methods mainly base on univariate or multivariate statistical tests. In this work, the Fisher's exact test (`SelectPercentile`) and the family wise error (`SelectFwe`) implemented in the Python package `scikit-learn` [39] are applied. Wrapper methods consider a classification algorithm as a black-box to evaluate the performance of the selected feature subset iteratively. One of the most common wrapper methods

is recursive feature elimination (RFE, implemented in scikit-learn [39]), whereby the number of selected features is reduced recursively using the feature importance of the internal tree-based method.

Following Bolón-Canedo et al. [37], filter methods usually are computationally less expensive and have a good generalization ability. In contrast, wrapper methods give better classification results but also tend to overfit and have high computational costs [37]. Due to their strong dependency of the applied classifier and their high computational costs [37], embedded methods are not used in this work.

3.4 Classification

Supervised learning is a subtask of machine learning, where an algorithm learns a function that maps a given dataset to a target, the so-called label. In the training step, the applied method learns the key characteristics of the dataset to determine the target value. Afterwards, the algorithm could predict the label of a new observation due to its ability of generalization [40]. Within supervised learning, one typically distinguishes between regression and classification tasks. In regression tasks, the label is a continuous value, whereas in classification, an object from a finite categorical set is predicted [40].

There is a broad range of machine learning algorithms; however, this study focuses on classification methods suitable for the detection of grinding burn. In state of the art, mainly neural network approaches are applied. To improve the classification performance, this study assesses a distance- and a kernel-based classifier as well as different ensemble methods based on decision trees.

3.4.1 Distance-based methods

The most common distance-based classification method is the k-nearest neighbor (kNN) algorithm, implemented also in scikit-learn [39]. Using this method, all training tuples represent a point in n-dimensional space. The label of a new datapoint is predicted from the class label of the k-nearest neighbors. To determine the distances, any metric measure can be applied; predominantly, the euclidean distance $\|.\|_2$ is selected. According to [35], the optimal number of neighbors k has to be determined experimentally. This choice might have crucial impact on the overall classification performance (see Section 3.4.4). Furthermore, the kNN algorithm is often assigned to lazy algorithms as no real learning is involved [40].

3.4.2 Kernel-based methods

The kernel-based machine learning methods transform the dataset in a higher dimensional space using kernel functions. Due to this so-called *kernel-trick*, one expects

that the data becomes linearly separable and classification is possible. The support vector machine (SVM) is the kernel method used most frequently, whereby a separating hyperplane that maximizes the margin between two classes is calculated [40]. For SVM, most commonly linear or radial basis function (RBF) kernels are applied, whereby the choice of the hyperparameters C and γ is critical to the classification performance. In this study, again the Python implementation SVC from the Package scikit-learn [39] is used.

3.4.3 Tree-based ensemble methods

In general, only one learning algorithm is applied to solve a given problem. Using ensemble methods, a set of machine learning algorithms mainly of an identical type is trained for the same problem. According to Zhou [41], this combination makes the generalization ability often much stronger than applying a single method.

In many applications, a decision tree, a non-parametric supervised machine learning method, is selected as base learner. In classification tasks, a decision tree learns to predict the label by adapting binary splitting rules inferred from the given dataset. At each node, the best split to minimize the so-called impurity function (*gini* or *entropy*) is searched [40]. Finally, the split with the highest information content is selected and the procedure is repeated iteratively. At each leaf node of the decision tree, a prediction for the unknown class label is given [40]. However, decision trees tend to overfit and selecting the split with minimal impurity in each step does not necessarily result in the global optimum [40]. To overcome these disadvantages, ensemble methods combine multiple decision trees by *bagging*, *boosting*, or *stacking*.

For *bagging*, the original dataset is randomly resampled with replacement to obtain a new dataset (called bag) for each base learner. Each model is trained independently in parallel and the predictions are averaged over all base learners [41, 42]. The most common representatives for *bagging* are random forests, where multiple decision trees are combined to decrease the variance and improve the overall classification performance. In extremely randomized trees, an improvement of random forests, the computation of the splits is further randomized, which results in a slightly lower computational time [43]. Both RandomForestClassifier and ExtraTreesClassifier are provided by the Python package scikit-learn [39].

Boosting mainly differs in the learning procedure, as it takes place in a sequential way. Each observation is assigned a weight, whereby after the base learner is trained, the weights of the misclassified data are increased. Thus, the further training process focuses on them. The final

prediction is computed by a weighted average over all previous learners [41].

3.4.4 Hyperparameter search

Each machine learning method has a various number of hyperparameters, e.g. the number of neighbors for the k-nearest neighbors algorithm. These model parameters are not known a priori, but their choice affects the performance significantly. There are four common approaches to optimize the hyperparameters and select the optimal model: Grid search, random search, Bayesian optimization, and evolutionary algorithm (EA).

According to Le et al. [44], the advantages of EA over the other approaches arise in the lower computational power needed, the flexibility in building multi-level tasks, and the large search space of methods and hyperparameters. The tree-based pipeline optimization tool (TPOT), developed by Le et al. [44], uses genetic programming to determine an optimal pipeline and is also implemented in Python. In this work, the term pipeline comprises the combination of one feature selection method and one classification algorithm. In the optimization process, a wide variety of different pipelines and their hyperparameters are evaluated with the objective to maximize the final classification performance.

3.4.5 Metric and scoring

The quality of the prediction given by the machine learning algorithms is quantified by the use of metrics. In general, the applied metric or metrics should be selected specific to the problem. In grinding burn classification, so far, mainly the *accuracy* is evaluated, which describes the number of right classified samples over the total number of samples presented. To ensure comparability to other works, in this study also the *accuracy* is used. Additionally, the *balanced accuracy* is calculated, which describes the average of the recall in each class and is suited for imbalanced datasets. Both of the metrics can be evaluated during training and testing of the pipeline. In this work, cross-validation (CV) is used during the training process to avoid overfitting; hence, the CV-score contains an average performance of all cross-validation folds.

4 Material and method

4.1 Experimental setup

The grinding experiments are conducted on an external cylindrical grinding machine. For all experiments, the same multi-layered metallic bonded cBN grinding wheel with a metallic wheel hub is used featuring a grit size of 150

to 180μm and a concentration of 150% (manufacturer's designation B181-C150-M-3V1-V15). The maximal tool diameter d_s measures 373.9mm and the grinding wheel width b_s of the straight profile is 7mm, featuring an inclination of 15° to compensate the tilt of the tool spindle holder. In this study, C45E (ISO EN 10269) is used as workpiece material. This quenched and tempered steel has a carbon content of 0.45% and a low content of phosphor and sulfur. It is hardened inductively to about 740 HV1 up to a hardening depth of approximately 2mm. The initial diameter of the prepared workpieces d_w measures 28mm.

The workpiece is radially and axially clamped between a jaw chuck and a grinding mandrel. The grinding wheel speed v_s varies from 80 to 100m/s and the workpiece rotates synchronously with $n_w = 250$ min^{-1}. A single experiment comprises one plunge grinding process with a constant radial infeed velocity v_{fr}, wherein radial direction 1 mm of the workpiece material is removed. In all experiments, a mineral oil–based cutting fluid of type *Wisura AKS 12* is deployed with a constant volume flow of 75 l/min. The applied process parameters are summarized in Table 3.

For the development of a robust classification algorithm, the process parameters of the recurring plunge grinding experiments have to be varied. The specific material removal rate

$$Q'_w = \pi \cdot d_w \cdot v_{fr} \tag{6}$$

describes the amount of material, which is removed per unit width of the grinding wheel within a certain time interval [45]. Another measure to compare different sets of process parameters is the quotient of the equivalent chip size h_{cu} and the geometrical contact length l_g, which is defined according to Klocke [45] as follows:

$$\frac{h_{cu}}{l_g} = \frac{\pi \cdot d_w \cdot v_{fr}/v_s}{\sqrt{f_r \cdot d_{eq}}} \tag{7}$$

Thereby, f_r describes the depth of cut during one workpiece revolution, v_w the circumferential velocity of the workpiece, and d_{eq} the equivalent diameter.

Table 3 Process parameters for the external cylindrical plunge grinding experiments

Parameter	Symbol	Unit	Value
Radial infeed velocity	v_{fr}	mm/min	20–50
Grinding wheel speed	v_s	m/s	80–100
Workpiece speed	n_w	min^{-1}	250
Radial depth of cut	a_e	mm	1.0
Workpiece diameter	d_w	mm	28.0
Wheel diameter	d_s	mm	373.09
Grinding wheel width	b_s	mm	7.0

Fig. 1 Numbered workpiece after performing the 14 plunge grinding experiments according to Table 4

As visualized in Fig. 1, each workpiece comprises 14 grinding positions. The corresponding process parameters are listed in Table 4. For the first four sets of process parameters, the radial infeed velocity is increased step-wise. This effects that the specific material removal rate Q_w' and the dimensionless value h_{cu}/l_g both increase. For the last two positions 1+2 the grinding wheel speed v_s is decreased, but keeping the quotient h_{cu}/l_g on the same level, the radial infeed velocity is also decreased.

4.2 Signal acquisition

To obtain a signal-based grinding burn detection method, the grinding machine is equipped with several system monitoring components. As visualized in Fig. 2, two different AE sensors with different positions are used for signal acquisition. In addition, the current and the power signal are recorded.

The AE sensors capture acoustic waves which can be described as transient waves that propagate through a solid material [46]. According to Chen et al. [16], AE sensors are easy to mount, even on rotating parts, at relatively low cost. Since the grinding zone is the critical region, the sensor is typically placed as close as possible to the grinding zone [47]. For this purpose, one rotating AE sensor (Marposs, type M), in the following called *AE-S*, is mounted at the tool spindle center and rotates with the grinding wheel. The position at the spindle center provides a constant distance independently of the grinding position and close to the grinding source. Another AE sensor (Kistler, type 8152B) is placed on the tailstock to capture the waves which originate from the workpiece and is abbreviated with *AE-T*. Since the distance from this sensor to the grinding source depends on the grinding position, it has been proven first

Table 4 Variation of the process parameters for the 14 plunge grinding positions within one workpiece

Position	12–14	9–11	6–8	3–5	1+2
v_{fr} in mm/min	20	30	40	50	30
v_s in m/s	100	100	100	100	80
Q_w' in mm³/mm s	27.2	40.5	54.5	68.1	40.5
h_{cu}/l_g in mm/mm	0.195	0.239	0.276	0.309	0.299
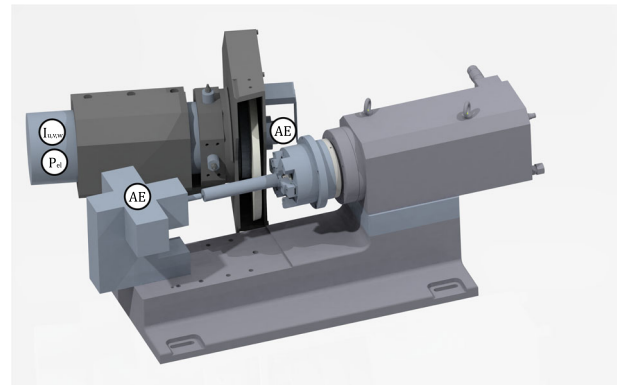
Fig. 2 Experimental setup and application of the sensors

that the varying distance has no significant impact on the signal characteristics. Both AE signals are preprocessed by the Kistler piezotron coupler of type 5125B, which provides built-in functionalities such as signal amplification and different low-pass and high-pass filters designed as plug-in modules. The amplification is set to 10dB and a band-pass filter of 100 to 500kHz is deployed. The data acquisition system is realized with a National Instruments CompactDAQ-system and organized in LabView. The digitizer works with a sampling rate of 2MHz for each AE-signal.

For the measurement of the tool spindle current, Chauvin Arnoux current clamps of type MN39 are installed on each electrical wire to acquire the 3-phase current components (I_u, I_v, I_w). The current components are then digitized with a sampling rate of 100kHz. In the next step, RMS transformation is applied in LabView to convert them into the equivalent direct current I_{rms}.

$$I_{rms} = \sqrt{\frac{I_u^2 + I_v^2 + I_w^2}{3}} \qquad (8)$$

The electrical tool spindle power P_{el} is measured with a sampling rate of 80kHz using the true power meter Artis MU-3 system.

4.3 Controlled generation of grinding burn

In this study, a touch dressing procedure is applied to the metal-bonded wheel to generate blunt grains and provoke controlled grinding burn. For the dressing process, a CVD-diamond profile roller is used rotating in opposite to the grinding wheel. In each dressing cycle, the infeed a_{ed} is about $3\mu m$ and divided into three single runs with $1\mu m$. To check, whether the grinding wheel has already contact with the dresser, the signal of the AE-sensor mounted on the grinding wheel spindle is used. All dressing parameters are summarized in Table 5.

Table 5 Dressing parameters for the touch dressing procedure of the metal bonded cBN grinding wheel using a CVD-diamond profile roller

Parameter	Symbol	Unit	Value
Tool speed	v_s	m/s	100
Dresser rotational speed	n_d	1/min	10,432
Dressing depth	a_{ed}	μm	1.0
Axial feed velocity	v_{fa}	mm/min	100
Dresser diameter	d_d	mm	80.01
Dresser tip radius	r_d	mm	0.15
Dressing overlap ratio	U_d	–	0.885

To emulate the wear of the grinding wheel, the following procedure is proposed. The unused grinding wheel is mounted on the tool spindle. In a pretreatment step prior the actual experiment, a complete workpiece is ground with a low infeed velocity ($v_{fr} = 5$mm/min) in order to achieve a quasi-stationary wear state of the tool, due to the removal of loosely bound grains with a high grain protrusion. Afterwards, the reference set with no touch dressing is ground, which includes six workpieces. Next, the first touch dressing cycle is performed and about 3μm are removed from the grains bonded on the grinding wheel. Then, three workpieces are machined to see the impacts of the first touch dressing cycle. This procedure is repeated until all stages of grinding burn including rehardening are reached. Based on experience, this state is achieved after a total amount of 15 to 18 touch dressing strokes.

4.4 Detection of grinding burn using nital etching

As introduced, in industrial application, predominantly nital etching and Barkhausen noise analysis (BNA) are applied [1]. According to Wegener and Baumgart [1] and He et al. [2], the latter is non-destructive and fast, but requires a complex calibration process due to the strong dependency on local material differences. Apart from BNA, nital etching capitalizes the discoloration of steels with thermal damage after a chemical reaction. The test procedure requires a difficult handling of the workpieces and the interpretation of the discoloration is very subjective [1, 2]. Besides these disadvantages, nital etching is the most reliable detection method in industry standardized in ISO 14104: *Surface temper etch inspection after grinding* [48].

Accordingly, nital etching is used in this study following the type 2 etching procedure from ISO14104 [48] recommended for low-alloy steels. It has been found out that the cleaning and drying procedure before each acid is important to obtain well-interpretable results. Finally, the class code from ISO 14104 [48] is used to classify the different stages of discoloration. The four classes range from no grinding burn (class A), over light and heavy tempering (classes B + D) to rehardening (class E). In this study, no additional classification has been carried out with regard to the percentage of surface area affected. A visualization is given in Fig. 3.

4.5 Implementation

The implementation of a method for the detection of grinding burn proposed in this study requires multiple substeps. In the following, the implementation is divided into two separate procedures.

The first procedure comprises the steps from the grinding process to the final prediction and is visualized in Fig. 4. The data acquisition is started, when the implemented CNC-Trigger is activated few seconds before the grinding wheel comes into contact with the workpiece. All of the analog signals are first amplified and filtered before a conversion into a digital signal is carried out and the raw data is saved on local hard drive using LabView. Applying a MATLAB script, the four sensor signals are imported and cut to the relevant time period during grinding (grinding detection and time reduction). Thus, all signals have a unique length of 0.5 s independent from the actual process time. Next, the signal components outside the frequency band of 100 to 500kHz are dampened (digital filtering). Afterwards, the preprocessed signals are used to apply the time-frequency transforms introduced in Section 3.1 including PSD, STFT, WPT, EEMD, and VMD. Based on both time and frequency information, the statistical features listed in Section 3.2 are calculated. Finally, one feature vector contains all relevant information of the performed experiment and is stored in the aggregated feature matrix. In this work, all workpieces are tested on grinding burn using the nital etching procedure introduced previously so each feature
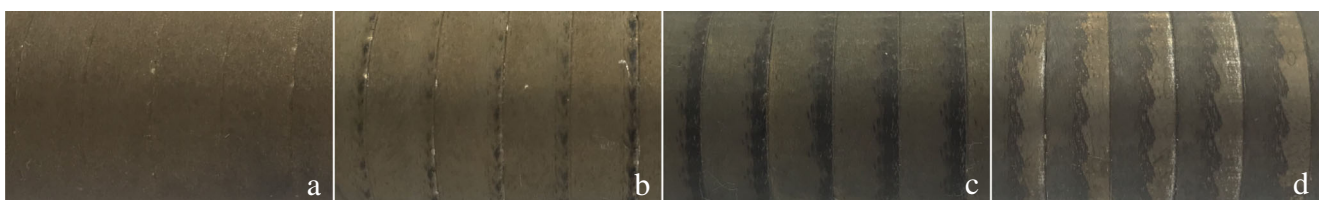
Fig. 3 The four classes of grinding burn according to ISO 14104. **a** No burn (class A), **b** light burn (class B), **c** severe burn (class D), and **d** rehardening (class E)
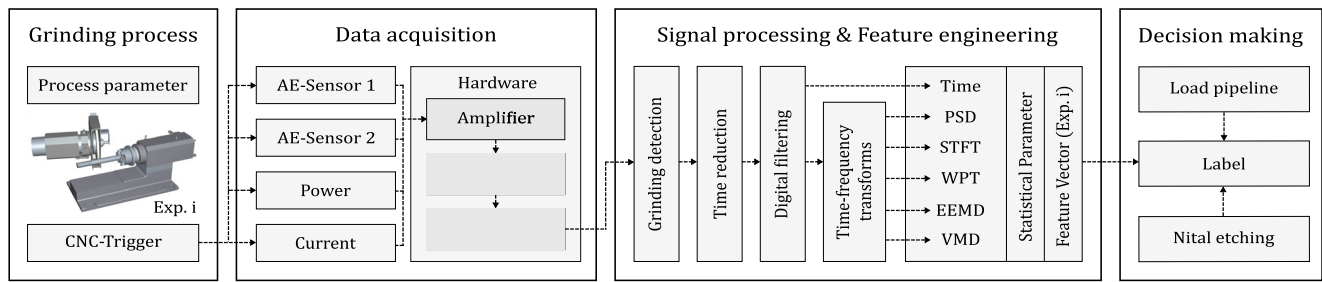
Fig. 4 Procedure to calculate statistical features and obtain a corresponding grinding burn label for one grinding experiment by capturing and processing signals from different sensors

vector gets its corresponding label. If a pretrained machine learning pipeline is available, a prediction on the grinding burn class can be provided to the user.

To obtain a trained machine learning pipeline, the procedure described above is repeated several times covering multiple wear states of the grinding wheel and as a consequence all of the four grinding burn classes (following Section 4.3). Hence, a high-dimensional feature matrix and the corresponding class labels obtained through nital etching are collected. In the second procedure visualized in Fig. 5, the machine learning part is implemented in Python based on the feature matrix.

First, a subset of the feature matrix is selected, which comprises either all features or a sensor-specific selection. For datasets holding a strong class imbalance, a resampling strategy is recommended to improve the training process. After resampling or in most cases directly after the selection of the sensor subset, the experiments of the feature matrix are divided into a training and a testing dataset (ratio 0.85/0.15). As a last preprocessing step, the feature matrix subset is standardized according to Section 3.2. Using only the training dataset, multiple feature selection and classification methods are preselected and their corresponding hyperparameter bounds are set according to Sections 3.3 and 3.4. As introduced in Section 3.4.4, genetic programming is applied to select the best combination of feature selector and classifier with the optimal set of hyperparameters. In this study, the objective of the hyperparameter optimization is to maximize the mean

classification accuracy averaged over ten cross-validation folds. Thus, an initial population with 100 pipelines is selected, trained, and evaluated using the stated metric. A new generation is built by copying (crossover-rate of 0.2) or adapting (mutation rate of 0.8) the best pipelines of the previous generation. The algorithm converges if the best CV-score of five consecutive generations is not improved or if the maximal number of 50 generations is reached. Using the best pipeline from the hyperparameter optimization, different test scores are calculated on the previously separated test dataset. In addition, a robustness analysis covering 100 randomly generated splits of the train- and test-dataset is performed. Finally, the procedure is repeated for each sensor subset.

5 Results and discussion

So far, an approach to find a well-performing combination between feature selection and classification algorithm for the in situ detection of grinding burn has been introduced. In the following, the results are presented and discussed.

5.1 Dataset

In this study, the feature matrix comprises in total 391 experiments with 878 features, respectively. After nital etching, each experiment is matched with a corresponding grinding burn class. The distribution of the experimental
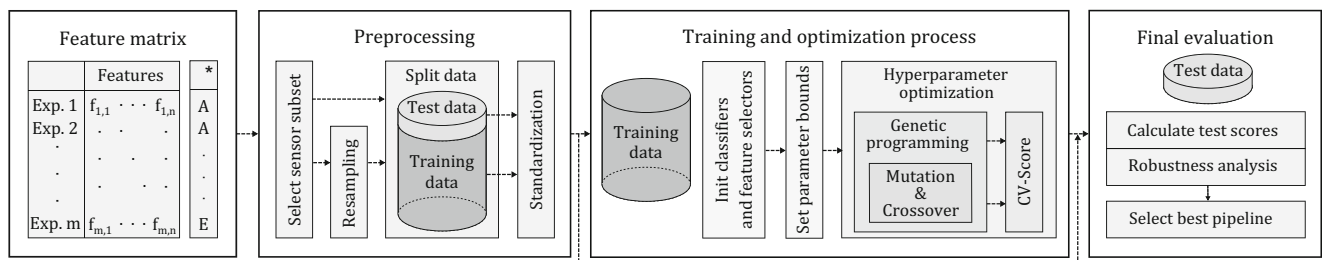
Fig. 5 Procedure to obtain an optimized machine learning pipeline to predict grinding burn based on a high-dimensional feature matrix and the corresponding labels (*)
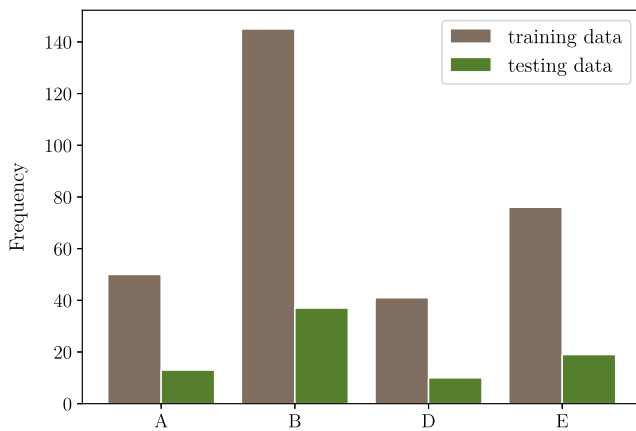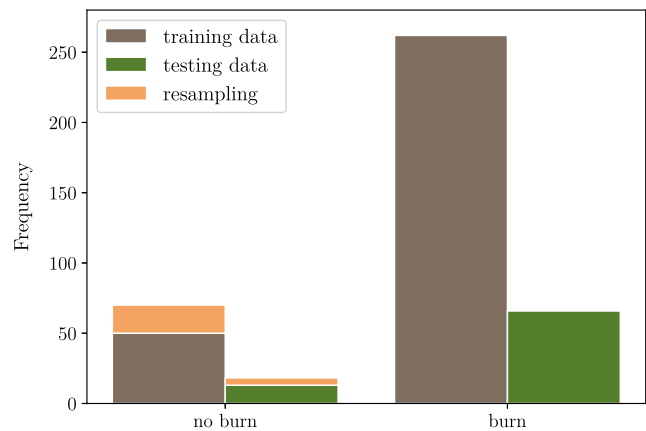
Fig. 6 Distribution of the experimental observations with respect to the four classes of grinding burn and the train/test split

Fig. 7 Distribution of the experimental observations with respect to the two classes of grinding burn and the train/test split (SMOTE resampling is used for the *no burn* class)

observations over the four classes is visualized in Fig. 6, whereby the classes with light burn (B) and rehardening (E) contain more experimental datapoints than the classes with no burn (A) and severe burn (D). As introduced in Section 4.5, the feature matrix is randomly separated into a training and a testing dataset ensuring that both contain observations of all four classes.

To estimate the impact of each sensor on the final classification result, the steps presented in Fig. 5 are performed for each sensor subset separately resulting in four feature matrices $\mathbf{F}_i$. Hence, the dataset $\mathbf{F}_{AE-T}$ contains only features extracted from the signals of the AE sensor mounted on the tailstock. The subsets $\mathbf{F}_{AE-S}$ and $\mathbf{F}_{Irms}$ comprise the features of the AE sensor in the spindle center and the features based on the current measurement, respectively. From the power signal, only one feature is extracted; thus, it is added to the $\mathbf{F}_{Irms}$ subset. Finally, the fourth dataset $\mathbf{F}_{all} = \mathbf{F}$ includes all of the 878 features.

5.2 Two-class prediction

In two-class prediction, the machine learning algorithm is trained to differentiate experiments with grinding burn from those without. Therefore, the labels are changed to *burn* and *no burn*, whereas *burn* comprises the classes B, D, E and *no burn* only class A. Consequently, the observations with grinding burn are over-represented compared to the *no burn* experiments. To improve the classification ability, a resampling method is applied to equalize the class distribution. The best final performance is achieved using synthetic minority oversampling technique (SMOTE) introduced by Chawla et al. [49]. As a consequence, the training dataset size increased to 390 and the testing dataset to 69, respectively. The distribution of the labels considering the two-class problem is visualized in Fig. 7.

Using the procedure of Fig. 5, an optimized pipeline containing a feature selection and a machine learning algorithm is obtained for each feature matrix $\mathbf{F}_i$. The final pipelines and their mean cross-validation and testing score are summarized in Table 6. In addition, Fig. 8 displays a confusion matrix for each of the four feature matrices, calculated with a fixed random state. Furthermore, these pipelines are analyzed using 100 random splits of the initial dataset. The resulting boxplots presented in Fig. 9 show the statistical deviations of the cross-validation scores during training on the left and the test accuracy on the right side.

For each sensor subset, the classification algorithm predicts the occurrence of grinding burn with a very high performance of over 94%, considering both the accuracy and the balanced accuracy. Furthermore, a high accordance of the cross-validation score and the testing score is evident, which suggests no over- or underfitting happens. Additionally, the standard deviation of the CV-score is very low and symmetrically distributed, which indicates a low impact of the train/test split. In general, a high level of generalization capability is established for the classification of grinding burn.

Going further into detail, the complete feature matrix containing the features from all sensors performs best with a mean accuracy of over 99% applying an kNN-classifier on 87 selected features. The great performance of the classification algorithm is slightly reduced using only features from one individual sensor. For the $\mathbf{F}_{AE-S}$ dataset, a RFE (using a decision tree as estimator) is combined with a SVM (using a RBF-kernel), which results in a decrease of the mean accuracy by less than 1%. Using $\mathbf{F}_{AE-T}$ and $\mathbf{F}_{Irms}$, the performance is reduced to 95 to 97%. Consequently, a reliable classification of grinding burn is possible using only features from one of the three sensor

Table 6 Two-class prediction results of the optimized machine learning pipelines for the feature matrices $\mathbf{F}_i$ evaluated using the mean (balanced) accuracy during training (CV-score) and testing (Test-score)

Dataset	Accuracy		Balanced accuracy		Selected pipeline	No. selected features
	CV-score	Test-score	CV-score	Test-score		
F_{all}	0.9901	0.9904	0.9900	0.9908	`SelectPercentile, KNeighborsClassifier`	132
F_{AE-T}	0.9627	0.9658	0.9635	0.9675	`SelectPercentile, SVM`	39
F_{AE-S}	0.9799	0.9820	0.9809	0.9829	`RFE, SVM`	110
F_{Irms}	0.9567	0.9591	0.9424	0.9449	`SelectPercentile, ExtraTreesClassifier`	73
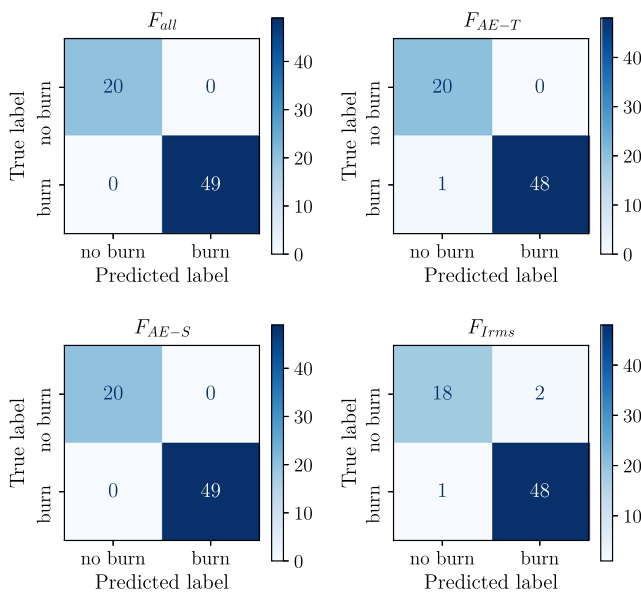
Fig. 8 Confusion matrices evaluated on the test dataset for the sensor-specific subsets considering the two-class problem for a fixed random state
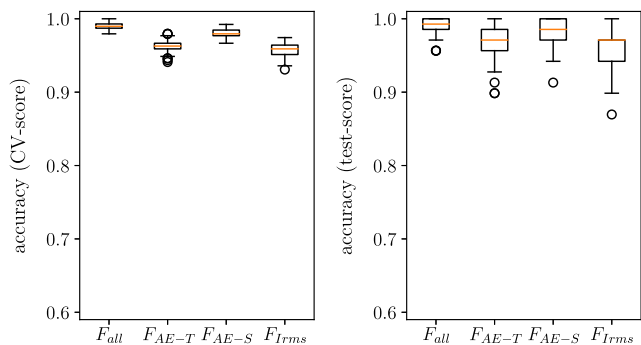
Fig. 9 Boxplots analyzing 100 random states for the sensor-specific feature subsets evaluated on the training (using CV) and testing dataset considering the two-class prediction

subsets. In addition, the multi-sensor approach performs even better than comparable methods from state of the art.

A deeper analysis of the wrong classified observations has shown that one experiment is classified wrong in each test run independently from the selected split. This indicates that this experiment is effected by an error in the grinding process, data acquisition, signal processing, feature extraction, or in the labeling of the etched workpiece. Removing this observation from the feature matrix would increase the overall performance towards 100%.

5.3 Multi-class prediction

For multi-class prediction, the classification algorithm is trained to distinguish between the four classes of grinding burn according to ISO14104 [48]. As visualized in Fig. 6, the class imbalance of the multi-class task is very low compared to the two-class problem. Consequently, no resampling method is applied.

The final pipelines with their scores are listed in Table 7 and the confusion matrix for a fixed random state is shown in Fig. 10. Furthermore, the boxplots in Fig. 11 visualize the statistical deviation of the training and testing accuracy.

Analyzing the complete feature matrix $\mathbf{F}_{all}$, the obtained pipeline consists of a kNN classifier and gives a high accuracy of about 98%. The low standard deviation and also the good agreement of the accuracy and the balanced one indicate a robust algorithm. In contrast to the two classes of grinding burn, the performance of the sensor subsets in multi-class prediction drops significantly. Focusing on the dataset of the AE-sensor in the spindle center, the RFE feature selector chooses 140 features using an extra trees classifier as estimator. The applied kNN algorithm predicts the four classes with about only 92%. Using different random states for the train/test split, the standard deviation increases to over 3%, especially considering the evaluation on the testing subset. Consequently, a robust classification result using only this sensor can not be expected. These results are also valid for the extra trees classifier applied on $\mathbf{F}_{Irms}$, where grinding burn is predicted with a mean accuracy of 88%. For both of them, a high accordance

Table 7 Multi-class prediction results of the optimized machine learning pipelines for the feature matrices $\mathbf{F}_i$ evaluated using the mean (balanced) accuracy during training (CV-score) and testing (Test-score)

Dataset	Accuracy		Balanced accuracy		Selected pipeline	No. selected features
	CV-score	Test-score	CV-score	Test-score		
$\mathbf{F}_{all}$	0.9747	0.9863	0.9811	0.9870	`SelectPercentile, KNeighborsClassifier`	87
$\mathbf{F}_{AE-T}$	0.7645	0.7660	0.8100	0.7123	`SelectFwe, SVM`	126
$\mathbf{F}_{AE-S}$	0.9166	0.9262	0.9351	0.9126	`RFE, KNeighborsClassifier`	140
$\mathbf{F}_{Irms}$	0.8732	0.8817	0.8875	0.8834	`SelectPercentile, ExtraTreesClassifier`	70

between the CV-score and the testing score as well as between the accuracy and the balanced accuracy is in evidence. Considering the confusion matrix, the wrong classified experiments are mainly shifted by one class. This could be effected by minimal deviations in certain feature values.

In contrast, the features of $\mathbf{F}_{AE-T}$ only achieve a classification accuracy lower than 80%, whereby the balanced accuracy is marginally better. The standard deviation of both metrics is very high; thus, the prediction on this dataset is not reliable. It can be assumed that the low information content of the AE-sensor on the tailstock is related to the sensor location on the workpiece tailstock. The transmission quality of the acoustic emission is somehow reduced due to the several mechanical transfer elements located between the source of AE creation and the signal detection. In addition, eigenfrequencies or structural vibration could influence the generalization ability of the AE sensor. Furthermore, the classes B, D, and E in the confusion matrix

become somehow indistinct, especially the decision boundary between the classes D and E is not built correctly. For the above reasons, the dataset obtained from AE sensor on the tailstock is not suitable for the multi-class prediction.

Again, two experimental observations are classified wrong in most of the test runs independently from the selected split. Hence, the classification performance could be slightly improved by removing these observations. As expected, the overall performance for all datasets of the multi-class prediction is inferior to the results obtained from the two-class task. So, a reliable multi-class prediction is only possible using all of the four signals in a multi-sensor approach.

5.4 Feature analysis

Finally, the selected and most important features for the classification results are analyzed. Therefore, the number of selected features is shown in Tables 6 and 7 for each pipeline. At first, the number of selected features from each sensor for the $\mathbf{F}_{all}$ dataset is examined building the average again over 100 random states. As depicted in Fig. 12**b**, for multi-class prediction, features of all sensors are selected almost equally often. This indicates again that all sensors are relevant for the classification and no sensor can be omitted. In contrast, for two-class prediction (see Fig. 12**a**),
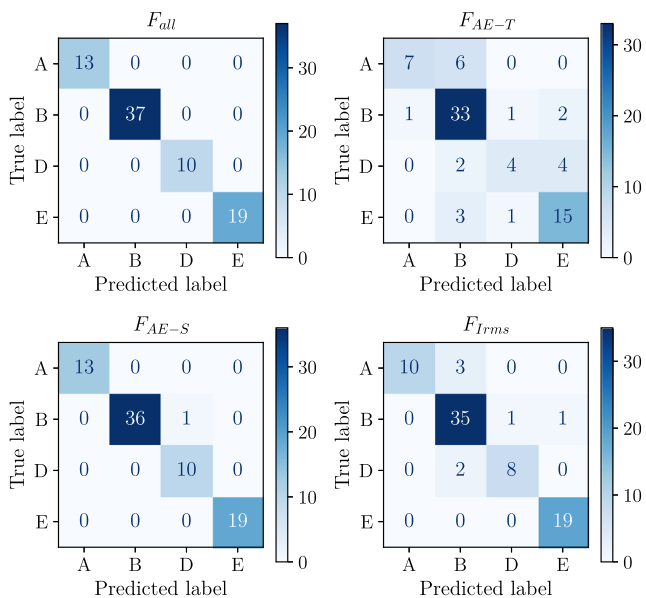
Fig. 10 Confusion matrices evaluated on the test dataset for the sensor-specific subsets considering the multi-class problem for a fixed random state
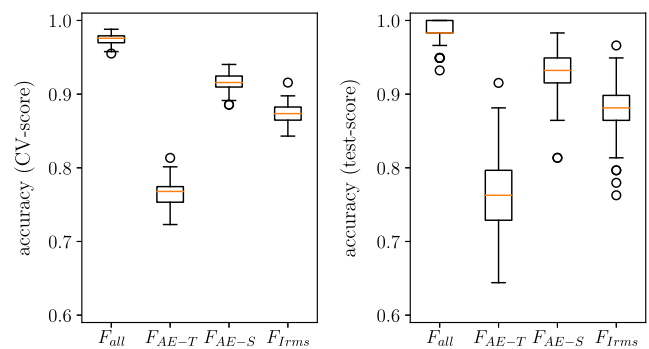

Fig. 11 Boxplots analyzing 100 different random states for the sensor-specific subsets evaluated on the test and training (using CV) dataset considering the multi-class problem

predominatly features from the AE sensor in spindle center are selected. In addition, the single power feature is selected for each random state. Hence, these results confirm the evaluation of the sensor-specific subsets as one sensor would be enough to differ between burn and no burn as well as multiple sensors are needed to classify multiple burn states.

In Fig. 13, the selected features are visualized with regard to the applied signal transformation methods, where TD represents the time-domain. The amount of features in each category is widely varying, e.g. the PSD comprises 585 features and the STFT only 9. To overcome this inequality during feature analysis, the number of selected features is related to the number of features in the specific category (green bars in Fig. 13). From all of the six time-frequency representations, features are selected, so each category contributes to the classification task.

Going further into detail, the selected features for the subset $\mathbf{F}_{AE-T}$ are analyzed. Although for the multi-class prediction features from all signal processing methods are picked, the STFT and WPT features are chosen most frequently related to the total number of features in the specific category. In two-class prediction, only features from WPT, PSD, and EEMD are selected. For both problems, many features are chosen from PSD-domain, whereby no characteristic frequency band is identified within 100 to 500kHz. Hence, the AE signal contains many relevant information in frequency-domain, but for multi-class prediction, the classifier is not able to learn the characteristics of grinding burn.

Next, the feature matrix $\mathbf{F}_{AE-S}$ is analyzed. Among the selected 140 features during multi-class prediction, more than 70% of the STFT, time-domain, WPT, and EEMD features are chosen from each category. In contrast, two-class prediction focuses mainly on EEMD and VMD features. These differences in feature selection might come
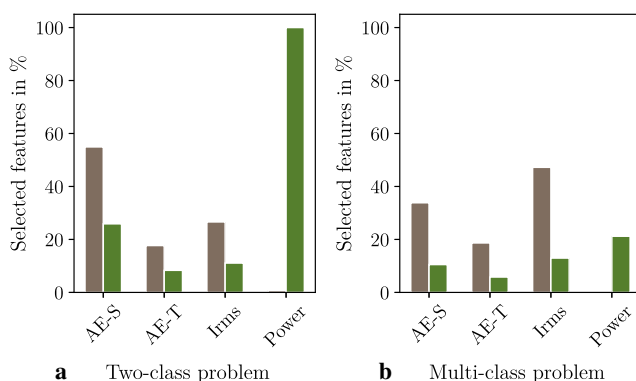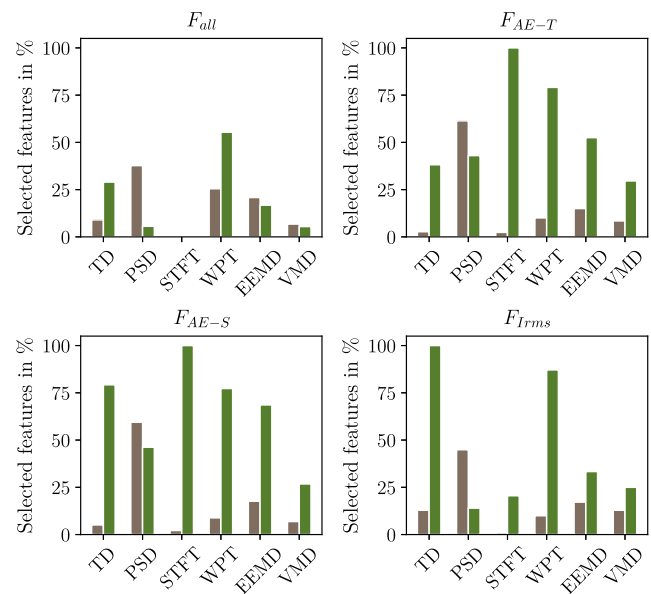
Fig. 13 Selected signal transformation methods for the multi-class prediction (**brown:** related to the total number of selected features, **green:** related to the total number of features in the specific category)

due to the different estimator inside the RFE-method. Although features from all frequencies are selected, two main frequency bands from 180 to 220kHz and from 100 to 140kHz are identified. This indicates that more than one characteristic frequency band correlates with the creation of grinding burn.

Assessing the I_{rms}-dataset, the feature selector mainly relies on the time-domain information. In addition, for multi-class prediction, the WPT features and, for two-class prediction, the STFT features are selected frequently. Going further into detail, the selected I_{rms}-frequency features for multi-class prediction are mainly located below 4kHz as well as between 16 and 20kHz and at about 32kHz. Besides the analysis of the feature selection, the most important features during classification can be evaluated for tree-based machine learning methods. Due to the known information content of each split, the mean decrease in impurity (MDI) can be interpreted as feature importance [42]. Following Rebala et al. [40], the more a feature contributes to decrease the impurity, the more important the feature is for the overall classification. In Fig. 14, the MDI values of the 10 most contributing features during classification using an extra trees classifier are visualized.

The abovementioned frequency clusters and also the importance of the time-domain features are confirmed by the MDI values. In contrast to feature selection, none of the WPT features is listed under the top ten MDI values. This indicates that for the I_{rms}, the feature selector and the classifier are not perfectly aligned to each other and more features can be omitted. Furthermore, the formation

Fig. 12 Selected features for the F_{all}-dataset (**brown:** related to the total number of selected features, **green:** related to the total number of features in the specific category)
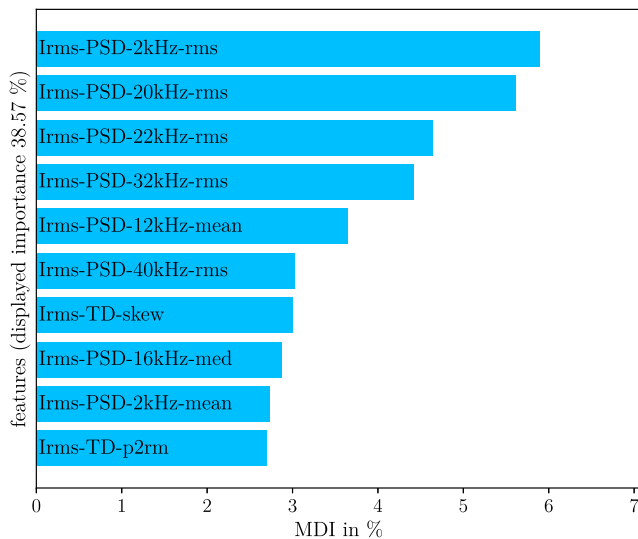
Fig. 14 Feature importance (MDI) of extra trees classifier on Irms dataset for mulit-class prediction

of frequency clusters might point out that a shift of these feature values leads to another level of workpiece damage.

It can be concluded that all of the sensors and signal transformation methods are necessary to achieve the high classification accuracy for the detection of grinding burn. Nonetheless, the calculation time of the signal transformation methods strongly varies from few milliseconds for PSD to many minutes for VMD. Therefore, it should be analyzed in future work if the classification result deteriorates when the time-consuming features from VMD and EEMD are omitted.

6 Conclusion

A new method for the in-process detection of grinding burn using machine learning is presented. First, an experimental procedure to generate controlled grinding burn using touch dressing with a low dressing depth is proposed, which enables the systematic creation of the four classes of grinding burn (according to ISO14104 [48]). The touch dressing procedure allows to approximately emulate the wear progress of the abrasive grains in a controlled manner. Due to the broad variation of the radial infeed velocity and the grinding wheel speed, a classification method can be derived, which is independent from the selected process parameters and also valid for high specific material removal rates. For each experiment, different time-frequency transforms are applied to acoustic emission, spindle electric current, and spindle power signals to calculate statistical features. Optimized pipelines, consisting of a feature selection and a classification method, are created for the sensor-specific feature subsets using

genetic programming for hyperparameter optimization. The results show a high classification accuracy for both two-class and multi-class prediction, whereby the applied multi-sensor approach gives better results than taking only features from one sensor. This work outperforms state of the art approaches by obtaining a mean classification accuracy of 98.6% for multi-class prediction and 99% for two-class prediction, respectively. Moreover, the post-mortem detection of grinding burn using nital etching can be replaced by the in situ method presented in this work. As a consequence in industrial manufacturing, all of the produced parts can be checked and the costly inspection process is replaced by an intelligent algorithm with a high accuracy.

In future work, the robustness of the proposed method can be improved further, whereby the experiments could be extended to different grain sizes, bond types, and also workpiece materials. Furthermore, the training of the ML algorithm is based on the results of the nital test, so the subjectivity in the class labels cannot be avoided. Therefore, in the future, other post-mortem detection methods like hardness testing can be applied on some selected workpieces to improve the quality of the labels and as a consequence the robustness of the algorithm. In this work, the hyperparameter optimization using genetic programming requires for each pipeline approximately 240 CPU hours parallelized on six cores. To improve the computation time without loosing classification performance, the use of GPUs or more efficient methods for hyperparameter optimization are needed in the future. In addition, new machine learning algorithms and feature selection methods can be explored further to strive for an error-free detection of grinding burn.

Declarations

References

1. Wegener K, Baumgart C (2018) Grinding burn. In: CIRP Encyclopedia of production engineering, pp 800–806. Springer Berlin
2. He B, Wei C, Ding S, Shi Z (2019) A survey of methods for detecting metallic grinding burn. Measurement 134:426–439. https://doi.org/10.1016/j.measurement.2018.10.093
3. Gholizadeh S, Leman Z, Baharudin BT (2015) A review of the application of acoustic emission technique in engineering. Struct Eng Mechan 54:1075–1095. https://doi.org/10.12989/sem.2015.54.6.1075
4. Lee DE, Hwang I, Valente CMO, Oliveira JFG, Dornfeld DA (2006) Precision manufacturing process monitoring with acoustic emission. Int J Mach Tools Manuf 46(2):176–188. https://doi.org/10.1016/j.ijmachtools.2005.04.001
5. Terchi A, Au YHJ (2001) Acoustic emission signal processing. Measurement and Control 34(8):240–244. https://doi.org/10.1177/002029400103400804
6. Griffin JM, Chen X (2009) Multiple classification of the acoustic emission signals extracted during burn and chatter anomalies using genetic programming. Int J Adv Manufact Technol 45(11-12):1152. https://doi.org/10.1007/s00170-009-2026-7
7. Kim BS, Lee SH, Lee MG, Ni J, Song JY, Lee CW (2007) A comparative study on damage detection in speed-up and coast-down process of grinding spindle-typed rotor-bearing system. J Mater Process Technol 187-188:30–36. https://doi.org/10.1016/j.jmatprotec.2006.11.222. 3rd International Conference on Advanced Forming and Die Manufacturing Technology
8. Aguiar PR, Cruz CED, Paula WCF, Bianchi EC (2008) Predicting surface roughness in grinding using neural networks. In: Aramburo J, Trevino AR (eds) Advances in Robotics, Automation and Control. Rijeka. https://doi.org/10.5772/5535
9. Eda H, Kishi K, Usiu N, Kakino Y, Fujiwara A (1983) In-process detection of grinding burn by means of utilizing acoustic emission. J Japan Soc Precis Eng 49(9):1257–1262. https://doi.org/10.2493/jjspe1933.49.1257
10. Saxler W (1997) Erkennung von Schleifbrand durch Schallemissionsanalyse. Ph.D. Thesis, RWTH-Aachen, Duesseldorf
11. Wang Z, Willett P, DeAguiar PR, Webster J (2001) Neural network detection of grinding burn from acoustic emission. Int J Mach Tools Manuf 41(2):283–309. https://doi.org/10.1016/S0890-6955(00)00057-2
12. Kwak J-S, Ha M-K (2004) Neural network approach for diagnosis of grinding operation by acoustic emission and power signals. J Mater Process Technol 147(1):65–71. https://doi.org/10.1016/j.jmatprotec.2003.11.016
13. de Aguiar PR, Bianchi EC, Canarim RC (2012) Monitoring of grinding burn by acoustic emission. In: Acoustic emission, 2nd edn, IntechOpen, pp 341–364, https://doi.org/10.5772/31339
14. Neto RFG, Marchi M, Martins CHR, de Aguiar PR, Bianchi EC (2014) Monitoring of grinding burn by ae and vibration signals. In: ICAART
15. Liu Q, Chen X, Gindy N (2005) Fuzzy pattern recognition of ae signals for grinding burn. Int J Mach Tools Manuf 45(7):811–818. https://doi.org/10.1016/j.ijmachtools.2004.11.002
16. Chen X, Mohamed A, Oluwajobi A (2012) Investigation of AE features in grinding. J Phys Conf Ser 364:012090. https://doi.org/10.1088/1742-6596/364/1/012090
17. Griffin JM (2015) Traceability of acoustic emission measurements for micro and macro grinding phenomena—characteristics and identification through classification of micro mechanics with regression to burn using signal analysis. Int J Adv Manufact Technol 81(9-12):1463–1474
18. Lajmert P (2010) An application of hilbert-huang transform and principal component analysis for diagnostics of cylindrical plunge grinding process. J Machine Eng 10(1):39–49
19. Lajmert P, Sikora M, Kruszynski B, Ostrowski D (2018) Application of principal component analysis and decision trees in diagnostics of cylindrical plunge grinding process. In: Hamrol A, Ciszak O, Legutko S, Jurczyk M (eds) Advances in manufacturing, pp 707–716. Springer International Publishing, Cham
20. Yang Z, Wu H, Yu Z, Huang Y (2014) A non-destructive surface burn detection method for ferrous metals based on acoustic emission and ensemble empirical mode decomposition: from laser simulation to grinding process. Meas Sci Technol 25(3):035602. https://doi.org/10.1088/0957-0233/25/3/035602
21. Guo W, Li B, Shen S, Zhou Q (2019) An intelligent grinding burn detection system based on two-stage feature selection and stacked sparse autoencoder. Int J Adv Manufact Technol 103(5):2837–2847. https://doi.org/10.1007/s00170-019-03748-5
22. Gao Z, Wang X, Lin J, Liao Y (2017) Online evaluation of metal burn degrees based on acoustic emission and variational mode decomposition. Measurement 103:302–310. https://doi.org/10.1016/j.measurement.2017.02.049
23. Webster J, Marinescu I, Bennett R, Lindsay R (1994) Acoustic emission for process control and monitoring of surface integrity during grinding. CIRP Ann 43(1):299–304. https://doi.org/10.1016/S0007-8506(07)62218-5
24. Huang N (2005) Introduction to the hilbert-huang transform and its related mathematical problems. Interdiscip Math, 5. https://doi.org/10.1142/5862
25. Isermann R (2006) Fault diagnosis systems an introduction from fault detection to fault tolerance. Springer
26. Hahn SL (1996) Hilbert transforms in signal processing. Artech house publishers
27. Huang N, Shen Z, Long SR, Wu MLC, Shih HH, Zheng Q, Yen NC, Tung C-C, Liu HH (1998) The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 454:903–995. https://doi.org/10.1098/rspa.1998.0193
28. Quek ST, Tua PS, Wang Q (2003) Detecting anomalies in beams and plate based on the hilbert–huang transform of real signals. Smart Mater Struct 12(3):447–460. https://doi.org/10.1088/0964-1726/12/3/316
29. Wu Z, Huang N (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in Adaptive Data Analysis 1:1–41. https://doi.org/10.1142/S1793536909000047
30. Zosso D, Dragomiretskiy K (2013) Variational mode decomposition. IEEE Trans Signal Process, 62. https://doi.org/10.1109/TSP.2013.2288675
31. Grimmett G (2001) Probability and random processes. Oxford University Press, Oxford
32. Tönshoff HK, Friemuth T, Becker JC (2002) Process monitoring in grinding. CIRP Ann 51(2):551–571. https://doi.org/10.1016/S0007-8506(07)61700-4
33. Sharma V, Parey A (2016) A review of gear fault diagnosis using various condition indicators. Procedia Eng 144:253–263. https://doi.org/10.1016/j.proeng.2016.05.131. International Conference on Vibration Problems 2015
34. Teti R, Jemielniak K, O'Donnell GE, Dornfeld D (2010) Advanced monitoring of machining operations. CIRP Ann Manuf Technol 59:717–739. https://doi.org/10.1016/j.cirp.2010.05.010
35. Han J, Kamber M, Pei J (2012) Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann Publishers, Waltham, Mass.

36. Guyon I, Gunn S, Nikravesh M, Zadeh LA (2006) Feature extraction: Foundations and applications (studies in fuzziness and soft computing). Springer, Berlin

37. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2015) Feature selection for high-dimensional data. Springer, Berlin

38. Pilnenskiy N, Smetannikov I (2020) Feature selection algorithms as one of the python data analytical tools. Future Internet 12(3):54

39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Édouard Duchesnay (2011) Scikit-learn: Machine learning in python. J Mach Learn Res 12(85):2825–2830

40. Rebala G, Ravi A, Churiwala S (2019) An introduction to machine learning. Springer, Cham

41. Zhou Z-H (2012) Ensemble methods - foundations and algorithms, 1st edn. CRC-Press, Boca Raton

42. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

43. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1

44. Le TT, Fu W, Moore JH (2020) Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics 36(1):250–256

45. Klocke F (2017) Fertigungsverfahren 2: Zerspanung mit geometrisch unbestimmter schneide. Springer, Berlin Heidelberg

46. Nazarchuk Z, Skalskyi V, Serhiyenko O (2017) Acoustic emission. Springer, Berlin

47. Jemielniak K (2001) Some aspects of acoustic emission signal preprocessing. J Mater Process Technol 109:242–247. https://doi.org/10.1016/S0924-0136(00)00805-0

48. ISO14104:2017(E) (2017) Gears — surface temper etch inspection after grinding, chemical method. Standard, International Organization for Standardization

49. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357