International Conference on Advanced and Competitive Manufacturing Technologies milling tool wear prediction using unsupervised machine learning

Thomas Gittler¹ · Magnus Glasder¹ · Elif Öztürk² · Michel Lüthi¹ · Lukas Weiss² · Konrad Wegener¹

Received: 5 January 2021 / Accepted: 12 May 2021 / Published online: 24 May 2021
© The Author(s) 2021

Abstract

Degraded or defect machine components and consumables negatively impact manufacturing quality and productivity. Diagnosing and predicting the wear or degradation status of critical machine components or parts are therefore of general interest. To tackle this challenge, data-driven approaches based on supervised machine learning principles have demonstrated promising results. However, supervised learning models capable of degradation identification require large quantities of data. In practice, run-to-failure data in large amounts is usually not available and expensive to obtain. To overcome this issue, this study proposes an unsupervised learning approach for degradation prognostics of machine tool components and consumables. It uses time series of multi-sensor signal data, which are transformed into a feature representation. The features consist of various characterizations of the time series, allowing to make different signal measurements comparable, and cluster them according to their feature values. The herewith obtained density-based clustering model is used to diagnose and predict the degradation states of components and parts in unknown conditions. The novelty in the proposed approach lies within the identification of continuous component and part degradation states based on unsupervised learning principles. The proposal is verified and demonstrated on an exemplary data set containing a small sample of run-to-failure multi-sensor signals of milling inserts and their corresponding wear state. By the application of the proposed procedure on the exemplary data set, we demonstrate that an unsupervised clustering approach is capable of separating wear data such that meaningful and accurate estimations of the part condition are possible. The advantages are its ability to cope with scarce data sets, its limited engineering and hyperparameter tuning effort, and its straightforward implementation to a multitude of degradation and wear diagnostics scenarios.

Keywords Condition monitoring · Machine learning · Prognostics and health monitoring · Unsupervised learning · Machine tools · Manufacturing · Milling · Tool wear

1 Introduction

Machine tools in manufacturing value chains are critical in view of amortization and smooth supply chain operation.

Peer-review statement: Peer review under the responsibility of the scientific committee of the International Conference on Advanced and Competitive Manufacturing Technologies

✉ Thomas Gittler
gittlert@ethz.ch

¹ ETH Zürich, Institute of Machine Tools and Manufacturing (IWF), LEE L, Leonhardstrasse, 21 8092 Zürich, Switzerland

² Inspire AG, Edison EG, Technoparkstrasse 1, 8005 Zürich, Switzerland

Degraded consumables and components affect both aspects, as they increase the risk for non-conformities and unplanned downtimes [1, 2]. Maintaining machines and tools in fully operational state is therefore of pivotal importance. The maintenance approaches can be divided into three main types: corrective, preventive, and predictive [3]. Corrective actions based on breakdowns or quality deteriorations fall short of assuring an optimal working regime. Preventive maintenance can be costly and nonetheless inhibit the risk of chance failures, which remain unnoticed without monitoring. Hence, wear status and conditions of machine components need to be continuously detected and assessed, to allow for predictive maintenance. Consequently, the diagnostic and prediction of wear on critical components and parts are primary research objects, according to Kusiak [4]. They are generally referred to as prognostics and health management (PHM). PHM “aims

at supporting maintenance operations” [2], and comprises any activity performed on a system to keep it in an operational state. To identify and preempt unwanted degradations or failures, data-driven approaches are increasingly used in manufacturing [5]. Especially under the paradigm digitization, smart factory, and biologicalization, machines in manufacturing systems are increasingly connected, measured, and equipped with intelligence and self-analyzing capabilities [6–8]. These advances especially support and accelerate the use of data-driven approaches to identify, diagnose, and predict machine part conditions and wear, which are unobservable or cumbersome to measure.

Responding to these challenges, three different approaches for PHM in machine tools can be distinguished (Table 1): anomaly detection, condition or wear identification, and remaining useful lifetime (RUL) estimation. These approaches are not mutually exclusive, wherefore combinations and hybrids exist. For anomaly detection (1), the optimum or initial working state of a machine is measured or modeled. Degradations and potential failures are identified as deviations or anomalies from the healthy state. Accordingly, the status of machine components or consumables is given as a qualitative indication. For condition or wear identification (2), the deviation from a healthy state is given as a quantitative indication, or a qualitative multiclassification. RUL estimations or predictions (3) put the identification of a condition in relation to prior observed degradation course physical wear models, or a combination thereof. Consequently, an indication of the expectable remaining usage time within thresholds of an a priori defined working state can be deducted.

Anomaly detection (1) requires to model merely the healthy state. As a consequence, breakdown reasons and the different failure types remain unknown. Additionally, the binary distinction between healthy and faulty states does not consider the various faults and their severity regarding the usability of the machine. In practice, not only the presence of anomalies but also the different kinds and their severity levels on machine tool components need to be captured, as described by Gittler et al. [9]. Condition, failure mode or wear identification (2), and RUL models (3) manage to indicate status and reasons for deteriorations. This allows to continuously assess the impact of current and potential future deteriorations, therefore offering additional possibilities to machine

users to ensure production uptime and consistent quality output. However, quantitative indications and multiclassification (2 and 3) require additional modeling, data, or knowledge of critical component degradation and failures [10]. Physical models are often not available or rely on specific experience or specialist domain knowledges. The design and training of condition identification and prediction models always involve domain knowledge, from sensor placement to signal parametrization all the way to model result validation. However, the incorporation of unformalized, abstract specialist knowledge into data-driven manufacturing PHM models does not follow a rigid framework, and is therefore subject to large variations. Compensating missing expertise with ubiquitous amounts of data, however, also faces limitations: The design and training of models capable of accurate degradation identifications and RUL predictions are difficult in practice due to data scarcity. Run-to-failure data is usually not available or expensive to obtain, as faults are often the result of multiple simultaneous degradations and therefore not unidimensional. Besides that, other challenges are the data gathering and modeling effort to reconstruct the different faulty states, as well as the transfer of the derived model to different machine and component types [9]. This holds especially true for multiple operating conditions, each requiring an individual and corresponding data set for training, which is generally not available and cumbersome to acquire in appropriate quantities.

Altogether, data-driven approaches for PHM purposes in manufacturing need to be able to distinguish multiple degradation states. They have to cope with small amounts of data for training and incorporate different operational settings where data is especially scarce. Moreover, PHM approaches need to be robust and resilient towards noise introduced by alterations in these operational settings and noise caused by environmental influences and variances during data acquisition cycles.

In this study, we describe an unsupervised clustering approach capable of separating wear data in such a way that meaningful estimations of the part condition are possible. It extends an approach previously proposed for the qualitative diagnosis of actual failure states [9] to the quantitative estimation of intermediate degradations. It uses time series signals of multiple sensors as an input, decomposes the time series into characteristic features, and uses the multidimensional space

Table 1 Characterization of PHM approaches in manufacturing

	Approach	Type	Output	Input
(1)	Anomaly detection	Binary	Binary	Healthy state
(2)	Condition/wear identification	Diagnosis	Classification/quantitative	Healthy + failure state(s)
(3)	Remaining useful lifetime estimation	Prognosis	Quantitative	Healthy + failure state(s)/degradation

described by all possible feature values as coordinates for a point representing the test cycle. In this manner, the feature space contains multiple points, of which clusters of similar wear and degradation states are formed. For the diagnosis of parts in unknown wear states, test cycle time series data are also represented by their characteristic feature values, and compared with the points used for model training. The cluster that the test point is most likely to be a part of indicates the predicted condition or wear state. The approach uses very little data for model creation, requires little engineering effort for setup and hyperparameter tuning, and may incorporate data from test cycles in differing operating conditions. It has a comparably small number of hyperparameters to tune, and demonstrates encouraging results. The proposed solution is verified on an exemplary data set containing measurement data and resulting measurements of milling tool insert flank wear.

This study focuses on the steps related to data preprocessing, model building, and testing. It does not include the steps of sensor placement and parametrization based on domain knowledge, raw data collection, transmission and storage, or model deployment and software architecture. The written outline of our research is structured as follows: First, an overview of related work and comparable approaches is given, followed by a detailed description of the process and its pertaining steps. Thereafter, the application of the proposed solution to the reference data set as a validation with real data is outlined, and its performance examined and assessed. To conclude, the strong and weak points of the proposed solution are discussed, potential gaps and research potential for the future are described, and its application to other machine components are suggested.

2 Related work

As maintenance of a manufacturing system is paramount to maintain operations, quality, and ultimately amortization of machinery, it is a research and engineering topic receiving widespread attention. However, the reduction of system risk levels by identifying or predicting its condition or wear state is pursued in different approaches and paths. This related work section therefore focuses briefly on the current status of PHM in machining and manufacturing, before diving deeper into the currently used algorithms, with a specific focus on degradation and RUL estimation.

2.1 Prognostics, health, and monitoring in machining and manufacturing

PHM describes the entirety of functions and activities executed on or in interaction with a system to support and facilitate maintenance operations. The overarching goal is to retain a

system within boundaries of an operational state that ensures adequate availability, output quality, and performance. To fulfill these requirements, Sarazin et al. [2] stipulate that a PHM approach “is composed of a prognostic component [...] and a component able to give the health status of the system”. This allows to assess the health status of a system that is in operation as well as to predict its future state. Interventions based purely on the identification of a present default are referred to as condition-based maintenance (CBM). It represents an advantage over preventive approaches, as the maintenance of components in CBM is planned according to the actual condition of the equipment in contrast to breakdown or scheduled maintenance. With the prediction of a future degradation or failure state, downtimes caused by spare part lead times and unexpected supply chain ruptures can be further reduced. However, the prediction horizon needs to match or exceed these lead times, in order to provide the required benefits. If the prediction is smaller than the lead times or even zero, the prognostic component is identical to the identification. Hence, the forecast horizon of the prognostics component represents the difference between condition-based maintenance (CBM) and estimative or predictive maintenance (EM, PM). Based on increasing use and decreasing cost of communication technologies and computational power, manufacturing data is acquired and accumulated at a growing rate. Higher computational power enables fast data processing tasks, such as cleaning, model training, and iterative hyperparameter refinement. Affordable and smaller sensors connectable with standardized communication protocols allow to register a larger number of information sources. Consequently, the use of data and the application of analytics algorithms in PHM are growing rapidly, as pointed out by Tao et al. [11].

Wuest et al. [12] have provided a comprehensive review of machine learning in manufacturing, in which they describe the major steps for a PHM approach as preliminary analysis, monitoring, diagnostics, health assessment, and prognostics. As each of these steps is independent research objects on their own, the focus in this study lies within diagnostics, health assessment, and prognostics. Considerations for preliminary analysis and data acquisition for monitoring are exemplarily described e.g. by Gittler et al. in [13] and [14]. For the diagnostics and health assessment tasks, Wang et al. [15] give an overview of current PHM analytics with a focus of vibration signal-based health indicators.

2.2 PHM algorithms and approaches

Wu et al. confirm the effectiveness of data-driven approaches for PHM purposes in manufacturing, which not only complement, but also outperform model-based solutions [16]. Yet, they point out that most current studies are based on “classical machine learning techniques, such as artificial neural networks (ANNs) and support vector regression (SVR)”. As a

more elaborate solution, they propose a random forest (RF) for the diagnostic of tool wear, which performs superior to various ANN approaches on their data and tested setup. Zhao et al. [17] have examined deep learning and its application to machine health monitoring in a thorough and comprehensive study. Depending on the chosen network type, the underlying data quality, and the data quantity, promising results can be observed. Fawaz et al. [18] have particularly examined time series classification (TSC) via deep learning in a more global context. They distinguish between generative models and discriminative models. Generative models contain an unsupervised training step before classification, in order to obtain a time series representation that allows for a classification. Discriminative models use the raw time series as input, and directly emit an output classification probability.

Overall, supervised learning algorithms are the state of the art in most machine monitoring and diagnostics applications. Diagnosis serves an important role in pursuing the relationship between the monitoring data and health states of machine. On the negative side, supervised learning approaches require large amounts of data for model training, especially in the presence of noise or inaccuracies. Moreover, supervised approaches struggle with the detection and labeling of component behavior outside of the learned cases, as long as the model is not used exclusively for anomaly detection. Hence, noise, outliers, and inaccurate data have an unfavorable impact due to the inherent input–output relationship of supervised models [9], raising the requirements either for model engineering or for training data set size.

To counter these negatives, unsupervised algorithms have a number of favorable aspects, as Zhang et al. [19] have demonstrated. Their approach, which they named “AnomDB,” is an anomaly detection of machined serial parts using density-based clustering. They use principal component analysis (PCA) on time series feature representations of machining signals. The resulting points are clustered in a 2D space, in which outliers are designated as anomalies. Given the underlying clustering approach, the model can be trained with very little data points, whereas inaccurate and noisy data does not interfere with model building. In contrary, it can even be incorporated in the model in order to represent the variation due to process and signal noise. The underlying principle used is density-based spatial clustering of applications with noise (DBSCAN), respectively a modification called hierarchical DBSCAN (HDBSCAN). Despite the encouraging results by Zhang et al. [9], there are some inherent limitations. Their solution is capable of detecting deviations from a pre-defined state only. Although the detection of outliers or the classification of well-delimited clusters is a straightforward procedure, the prediction of a degradation as a continuous distribution remains a challenge.

2.3 Degradation and remaining useful lifetime predictions

As the challenges vary greatly with different PHM tasks, there are manifold approaches to degradation and RUL predictions. A main determinant of adequate approaches is the quantity, resolution, and quality of signal time series available for model training and testing. As there are different data sets in use, some publicly available and some proprietary to the authors, comparisons are qualitative based on the approximate data set attributes that are revealed in the respective publications.

An approach to determine the degradation state of filters in a machine tool oil mist filter was published by Gittler et al. in 2019. They proposed the use of an ANN to model exemplary and environmental variances on an oil mist filter, allowing to observe the degradation states of filters. It models the power consumption based on process and environmental input variables, and uses the deviation from the actual measured power consumption as an indicator for the degradation. In this case, specific domain knowledge allowed to compare the difference in power consumption between healthy and degraded filters with previous measurements, to deduct the need for replacement. As the underlying had a particularly low sample rate, the authors deemed a TSC not suitable, wherefore raw data was directly fed to an ANN to model the input–output relationship.

With a focus on tool wear, Sun et al. [20] used time series representation of acoustic emission (AE) signals, to apply an SVM for a multiclassification of tool wear in 2004. They divided tool wear into three regimes, and subsequently applied a binary SVM classifier to each. As a result, risk for losses incurred by premature tool change and tool breakage could be reduced. Losses due to quality deteriorations following excessive tool wear are not considered. They fed 9 features, of which three represent cutting conditions and 6 are extracted from the AE signal, to the prediction algorithm. As the approach is a double-layer binary classification, it does not actually predict a degradation or RUL.

The beforementioned study of Wu et al. [16] recently proposed a random forest (RF)-based tool wear prediction approach, in which they extracted 12 features from cutting force, vibration, and AE signals. They benchmarked the RF algorithm results against support vector regression (SVR) and ANN models by mean squared error (MSE), R^2 value, and required training time. The RF proved to be superior in all accuracy metrics, and inferior in required training times. The overall data volume they acquired on a proprietary setup amounted to 8.67 GB worth of signal time series.

In general, degradation models and RUL models are uniformly constructed via supervised approaches, incorporating the previously described drawbacks of sensitivity to noise and requiring large amounts of training data. In the case of degradation and RUL models, this translates to numerous run-to-

failure tests that need to be conducted to acquire the necessary data. To cope with noisy and very small data sets, Gittler et al. recently suggested to use unsupervised clustering for failure recognition in machine tool components [9]. They extend the principle of Zhang et al. not only to allow for the detection of anomalies, but also to designate the type of anomaly, i.e. failure states on machine components. It allows to extract large numbers of features from multi-sensor time series, and selecting only the relevant ones for TSC-based condition identification. The challenges associated herewith are the representation of time series that allow for clustering, the identification of non-convex clusters, the designation of outliers as unclassified points or noise, and the ability to cope with highly variable cluster densities. The proposal was verified and tested on a machine axis, in which real data was acquired from components on which failures and degradations were recreated artificially. Their approach has demonstrated satisfactory results on a real-world data set with very little samples, correctly classifying both a priori known and unknown error states of a machine axis. It also demonstrates that unsupervised approaches, though mostly disregarded for manufacturing in the past, have proven useful for challenges that cannot be overcome with supervised learning. However, Gittler et al. do not provide an actual wear or condition degradation. In general, unsupervised approaches for degradation and RUL identification in machining have not yet been evaluated to our knowledge. The characteristics of unsupervised learning might be useful; however, there are also imminent challenges associated: Unsupervised algorithms manage to distinguish well discriminable states, but encounter difficulties with continuous distribution characteristic to wear or degradation processes. This is due to the fact that continuous distributions of feature representations make clustering difficult. Table 2 describes the caveats of both supervised and unsupervised approaches, which in short are as follows: Supervised methods attain high accuracy while requiring large data volumes for training, whereas unsupervised methods require little data but lack accuracy of discrimination.

3 Materials and methods

The methodology resembles a conventional data science approach, wherefore we proposed the following structure:

- (1) Data acquisition: sensor and signal requirements, test cycle setup, and recording of raw data,
- (2) Data preprocessing: data parsing, cleaning, and preparation for model training,
- (3) Model creation: construction of a model via the deployment of suitable learning algorithms,
- (4) Model deployment: test, evaluate, and update model with test data set or future test cycle data.

The overall procedure is depicted with exemplary and simplified excerpts in Fig. 1.

3.1 Data acquisition

Data acquisition is the first step of the proposed method, which provides the fundamental input for the information extraction and processing. It requires a number of prepositions to be met. Recorded data for PHM purposes comes as monitoring data, i.e. raw sensor signal time series, or as event meta data, i.e. the stable operating and environmental conditions during data acquisition. As the resulting data sets need to comply with the constraints set by the downstream analytics, the following principles are crucial to obtain the desired outcome and results.

As the acquired time series will be represented by features, both high resolution and sample rate are necessary to permit the calculation of all pertinent features. Hence, the sampling rate needs to be $\gg 0.1$ kHz, and as high as possible—as long as the measurement principle does not introduce high pitch noise signals for high sampling rates, or frequency filters are used. Some differences to detect in the time series are only minor, wherefore the obtainable resolution has the second highest impact on the result. For the signal and sensor selection, there needs to be a direct link with the process. In general, force or representative current or input power measurements are always helpful, accompanied by AE or vibration signals. For the test cycle preparation, during which the sample data is

Table 2 Comparison of prediction approaches and their limitations

Publication	Category	Method	Limitation
Sun et al. [20]	Supervised	SVM	Classification of only 3 regimes
Wu et al. [16]	Supervised	Random forest	Large data volume needed
Fawaz et al. [18]	Supervised	Deep learning	Explorative study only
Gittler et al. [14]	Supervised	ANN	Quantitative anomaly only
Zhang et al. [19]	Unsupervised	Clustering	Binary anomaly only
Gittler et al. [9]	Unsupervised	Clustering	Categorical classification

Fig. 1 Methodology and process overview with exemplary depictions along the process chain

acquired, the following conditions need to be met: In a best-case scenario, data should always be acquired in a reproducible stationary regime. This can be inside or outside machining or processing, as long as a stable and constant condition persists during data acquisition. In case there are changing conditions over the sample measurements, e.g. varying feed or cutting depth, inside and outside of all test cycles, this information should be measured alongside as event meta data. Timewise, the data needs to be captured either with a timestamp or with equidistant sample points. Each sample should be at least the equivalent of roughly 1000 data points. However, the longer the sampling period, the less susceptible to noise and outlier values the feature extraction. Sample measurements should be repeated or measured for longer time spans, in order to cross-check the data consistency within a sample measurement.

3.2 Data preprocessing

From the test cycle measurements, the initial record needs to be divided into training set for model creation, and a test set to evaluate the performance. With degradation models, it should be ensured that both training and test set contain an approximately equal amount of samples all along the degradation curve. All sample measurements need to be labeled with its quantitative wear or degradation indication, in order to allow for comparability. The samples need to be of equal length in order to allow for comparability of the extracted features later

during processing. The longer the samples, the less noisy are the extracted features. However, feature extraction is computationally expensive, wherefore computing times increase with sample duration and quantity. If computing power or time is a constraint for model construction, this trade-off needs to be weighed. If longer time spans are recorded instead of singular sample measurements, these spans can be divided into shorter samples; e.g., a 9000 ms measurement can be divided into 3 samples of 3000 ms each. In case data is sparse, a longer time sample measurement of n time units can be divided into $n-m$ samples of length m by means of windowing, as shown exemplarily in Fig. 2. As a result, a larger number of samples with corresponding degradation values are obtained, even from sparse data sets. For the next steps, it is necessary to turn the continuous degradation values into categorical labels by means of binning. The entire span of degradation values is divided into equidistant bins, in order to allow for significance testing and clustering. It is imperative to verify that all bins are populated with at least a minimum amount of samples. The so-prepared and binned samples can now undergo feature representation, in order to allow for subsequent clustering.

3.3 Time series feature representation

As time series are themselves difficult to interpret and compare, the transformation into a feature representation allows for such operations. The overall process of feature representation comprises two main steps: feature extraction, and the

Fig. 2 Data preprocessing flowchart with detailed examples

following selection of useful, i.e. significant features with pertinent distributions. The process is exemplarily illustrated in Fig. 3, with a special focus on the selection of features based on their distribution, as shown in Fig. 4.

There are numerous approaches to feature representation, some using a small number, and others all the way to multiple hundreds per recorded time series. Each sensor signal is considered a different time series, wherefore the number of extracted features increases by a multiple with additional data sources. The proposed approach is purely data-driven for the feature extraction and selection, wherefore no specific domain knowledge is invested at this point. There are a total of more than 700 features extracted per sample signal, which are e.g. autocorrelation values with varying lags, absolute energy values, fast Fourier transformations (FFT), complexity values, or just simple means and quantile distributions. However, not

all features are significant or even useful, wherefore a multi-step filtering process is required.

To make features comparable, they need to be scaled or normalized, preferably with a robust scaler, as some distributions are very skewed or prone to dominant outliers. To counter this effect, the feature sets of all samples should be filtered for outliers, in order to discard samples affected by noise, transitional zones, or non-stationary regimes. This can be done by means of filtering by z -score, or by introducing a confidence interval bounded by n standard deviations δ around the mean μ : $CI = \mu \pm n * \delta$. The confidence interval is calculated per bin. For the first filter step, a feature significance hypothesis test needs to be carried out. The respective p -values for all samples are calculated according to the categorical label that was assigned by binning during preprocessing. Due to the high number of bins to represent the degradation curve, the

Fig. 3 Flowchart detail feature extraction, features are exemplary and non-exhaustive

Fig. 4 **a** Qualitative example for a feature with low significance, different colors for different bins (almost complete overlap and identical distribution densities). **b** Example for a significant feature with smaller overlaps, with a higher likelihood of passing filters and use for clustering

maximum tolerable p -value should be rather conservative, i.e. $p < 0.5\%$. Higher p -values risk overlapping point clouds, which are impractical to cluster. A visual reference for a non-significant and a significant feature density distribution is shown in Fig. 4 (a) and (b), respectively. Following the significance test, filtering for the distribution of features within a bin is necessary. The sample distribution of a single bin (i.e., samples of the same degradation state) should be densely packed in each feature. Therefore, standard deviation, kurtosis, span of quantile limits, or a combination of these indications can serve as a filter. By these means, the potential to cluster the resulting feature representations can achieve the necessary threshold to obtain meaningful results. As a last filter, the correlation among the remaining features needs to be checked, in order to avoid bias or dominance of single characteristics. As a general rule, the number of remaining features should be in the range of 10–100 and always smaller than the number of samples, to allow for a straightforward clustering without overfitting on single features. As an outcome, all samples are now described by a set of features, representing different time series characteristics. These features can be perceived as a position vector of the sample as a coordinate in a multidimensional space, in which the number of dimensions n equals the number of remaining features m after filtering.

3.4 Model creation

After feature extraction and filtering, the feature representations of the training set should now form point clouds, in which each cloud is composed of points from identical bins. These clouds can now be turned into clusters, which serve as the prediction delimitation for points of unknown degradation states for the deployed model. However, as the number of retained features m is identical to the dimensions n of the space in which the point clouds lie, the clusters will be in a high-dimensional space. This translates to the necessary assumption that the clusters will not be of convex shapes. It is therefore mandatory to select a clustering algorithm capable of clustering based on density rather than distributions. In conventional clustering or partitioning approaches, the distances from the mean (e.g., for k -means) or Gaussian probability

distributions (e.g., for Gaussian mixture models) implicitly assume a hyper-ellipsoid cluster shape for $n > 3$. Hence, Gaussian distributions, even in a multidimensional space, may be cumbersome to turn into adequate clustering results. As a remedy, DBSCAN or HDBSCAN yield better results, as outlined in [9], on which this approach builds. The advantage of HDBSCAN over DBSCAN is the auto-inference of another hyperparameter ϵ , which reduces the engineering and optimization effort. As a result, HDBSCAN should return a number of clusters almost identical to the number of bins contained in the training data, and very few or zero outliers that are not members of any clusters. The latter are also called noise points, as their characteristics do not belong to any of the other found clusters. Depending on the input data, this may hold true, or could be an artifact of regional high-density areas due to overlapping point clouds. However, in recent developments of HDBSCAN, there are further possibilities to include points with lower probability of being assigned, which is referred to as soft-clustering (Fig. 5). For data sets with very few samples, it is possible and in some cases even desirable to have more clusters than the originally number of bins contained in the training set. This is due to variances in the samples and features, which may be due to noise or varying operating conditions during the acquisition of the test cycle data. The quality of the clustering result can be measured by one or more indicators; e.g., the number of clusters is equal to or higher than the number of bins in the training data set, a small number of noise points, an even number of points across all clusters (or according to the training set if the distribution is skewed), the standard deviation of all wear indications within a cluster is small, and the span of all degradation indications within a cluster is small compared to the overall degradation span of the training set. As soon as an adequate model is established, it can be verified with the test data set, and subsequently deployed.

3.5 Model deployment

The validation of the created model is identical to its use in a deployed state. Raw sensor signal time series should first be filtered for outliers by comparison with the samples used for model training; e.g., feature values of the test samples should

Fig. 5 Model training procedure demonstrating the application of HDBSCAN to identify clusters of samples with similar VB values

lie within the min-max boundaries of the training samples. Subsequently, the features used in the trained model are extracted for the test samples, and then scaled to match the training data set format.

For the prediction of new points, there are two possibilities: (i) The *approximate predict* function of HDBSCAN can be used, which assumes clusters to be fixed and assigns a probability of belonging to a certain cluster to each new point. (ii) The new point is added to the training set, and an entirely new clustering via HDBSCAN is performed. In both cases, the test sample is either assigned to an existing cluster label or identified as noise. As noise classifications are obtained frequently, it is useful to have multiple samples from the same test cycle, to obtain a significant and robust result. As noise point classifications are also frequent with points that lie just between two or more clusters, or in areas of cluster overlap, these points can be attributed to an existing cluster in a further step, in order to enhance prediction accuracy. Together with the attribution to a distinct cluster, the attribution probability, also referred to as membership strength, can be calculated. It is an indication of similarity to existing clusters. For predictions that held true in hindsight, points with high membership probabilities in sparsely populated areas and points with low membership probabilities in densely populated areas are optimal candidates for model update and refinement. The outcome of a trained model, as well as how the deployed model is used for prediction, is depicted in the following result section in Fig. 9 and Fig. 10, respectively. It should be noted that the clusters resulting from model training are transformed into the 2D plane for the sake of visualization.

4 Results

For the validation of results, an existing and publicly available data set is used. The data set was created and distributed by Agogino and Goebel [21] of NASA AMES und UC Berkeley lab, respectively. It has served as a foundation for other approaches and studies, and has ideal properties to demonstrate the strengths of the proposed solutions: It contains multiple operating conditions with very small amounts of samples each, in which run-to-failure recordings during intermediate data acquisition sequences are recorded with high sample rates (250 Hz).

4.1 Data set properties and description

Tool wear requires more than just a binary classification between healthy and faulty, as pointed out by Sun et al. [12]. First of all, maximum admissible tool wear may depend on material or machining parameters, and second, tool wear is a highly dynamic process. Initially, the cutting edge of the insert experiences a rather rapid degradation before attaining a uniform but slow wear progress. After a varying period of time, the degradation gradient increases and sudden ruptures or chippings cause the insert to fail. These inherent dynamics require a more granular approach than a binary classification, in order to allow for timely and preventive interaction before failure, quality non-conformities, or breakdowns. Moreover, there are different feed rates, depths of cut (DOC), and materials used, which allows to test the abilities of the proposed approach. The data set contains signals of six sensors, with vibrations and AE measured on both spindle and table (Fig. 6 a), as well as AC and DC motor currents with high resolution

and sampling rate. The wear is given as a quantitative indication of the VB value, which represents the width of the damaged surface in reference to its original state (Fig. 6b). The VB value is the target prediction variable. Its degradation curve is exemplarily depicted for two of the overall 16 cases in Fig. 7. The provided sample data set is based on real-world data, representing challenges that correspond to the ones one would encounter in an actual shop floor setting and data acquisition scenario.

4.2 Data preprocessing

In a first step, missing values for VB in the data set are filled by means of interpolation. For missing VB indications during the first run, zero flank wear is assumed. Runs at the end of the life cycle with no indication are discarded from the data set, as no meaningful replacement is possible. From the remaining samples, a selection of cases is isolated: All runs with identical material and feed rate are selected as a subset, with varying DOCs as only variable parameter. This is a reasonable limitation, as the DOC is the only parameter difficult to isolate in a real-world setting, while feed and material can be set and therefore monitored without additional effort. The given study is performed on the subset of all *cast iron* milling samples, with a feed rate of 0.25 mm, and DOCs of 0.75–1.50 mm. As a result, there are 4 runs-to-failure in the subset in total, which

Fig. 6 **a** Depiction of clamping and sensor setup on machine table, with which the data was acquired. **b** Close-up of cutting insert edge and its respective flank wear VB (ger. Verschleissmarkenbreite), from [21]

Fig. 7 Exemplary degradation curves of the used data set for two cases: cast iron cut with a feed rate of 0.25 mm and a depth of 0.75 mm (orange), and steel cut with a feed rate of 0.25 mm and a depth of 0.75 mm (blue)

are of identical setting except for the variations in DOC. From the remaining samples in the subset, a separation of 80% for model training and 20% for model testing is conducted arbitrarily. The training samples are now divided into bins by means of their VB values. For the demonstration case, VB values are in the range of 0.00–0.76 mm, for which a total of 9 bins each spanning 0.0844 mm provide a sufficient resolution for later predictions. However, this also indicates that the maximum obtainable precision in case of unfavorable data distribution is half the span of a bin, i.e. 0.0422 mm.

4.3 Time series feature representation

The resulting bins serve as the common denominator for the selection of features. The samples will be examined bin by bin, in order to determine which features allow for an accurate separation and representation of bins. For each of the 6 sensor time signals, 763 features are extracted, amounting to a total of 4578 feature values per sample. These features normalized with a robust scaler. Contrary to a min-max-scaler, it allows to accommodate extreme distributions due to noise and outliers. On the downside, a robust scaler does not result in equal ranges for all features, which has adverse effects on the subsequent filtering process. Depending on the range of feature values prior to scaling, the normalized feature ranges differ by an order of magnitude to the power of ten. Applying standardized filters, e.g. a maximum standard deviation $\delta_{max} \leq 0.7$ across all samples of one bin of a given feature, may risk to unnecessarily discard vital features with wider distributions. However, as long as the number of remaining features allows for reasonable clustering results, this consideration is left for further optimization. Throughout the filtering process, the retained features are reduced to 1962 features after the relevance filter, 547 features after the maximum standard

deviation filter, 203 features after the maximum quantile distance filter, and 37 after the correlation filter. An overview of all retained features, their parameters, and ranges is provided in the [Supplementary information](#). Due to the large number of bins, there is eventually an overlap between feature values of two or more bins in all remaining features, as indicated in Fig. 4 (b), and demonstrated in Fig. 8. Nonetheless, when plotting the distribution of feature values grouped by bins in a histogram, it becomes evident that the feature values of identical bins tend to accumulate in the same ranges. This is the necessary condition for the following clustering, as it ensures that sample points of the same bin accumulate in the same region. The overview over all features and their distributions is provided in the [Supplementary information](#).

4.4 Model creation and deployment

The samples are now be used for clustering by the remaining 37 feature values. To verify whether the samples are well-

Fig. 8 Exemplary depiction of two selected features, and their respective occurrence distributions of samples by bin. **a** Change quantiles of the acoustic emission (AE) sensor of the table. **b** C3 time series non-linearity indicator for the AC loop of the spindle motor current loop

represented by their feature values and form cluster-like shapes accordingly, a visual representation via t-distributed stochastic neighbor embedding (t-SNE) is helpful (Fig. 9). It can be seen that some bins clearly accumulate in singular clusters (e.g., blue and yellow), while the others tend to have more overlap. t-SNE visualizations help to gain an insight of how distinct the samples differ in the feature representation. The result shown in Fig. 9 indicates difficulties in the regions of higher VB values, as there is an overlap of multiple bins. However, due to data scarcity, improvements can only be made by reducing the number of samples with the risk of overfitting, or by acquiring additional data samples. With the given data set, the number of samples is limited by definition, wherefore the model is trained on the basis shown.

As a next step, a clustering model is trained on the selected samples and their respective feature distributions with HDBSCAN. The quality of the clustering result can be measured via the number of clusters, the number of noise points, and the standard deviation or span of VB values of all points per cluster. In a best-case scenario, there is a cluster for each bin, containing only samples from that respective bin, and zero noise points. For the given training case, HDBSCAN identified 8 clusters with 3 noise points, with a standard deviation δ ranging from 0.009 to 0.118 mm. As identified in Fig. 9, the overlapping samples in the bottom-right corner (high VB values) have proven difficult for clustering. Also, the cloud of yellow-green points between the center and the bottom-right corner have resulted in clusters containing samples of more than one bin. With the so-trained model, a validation can now be carried out on the reserved test samples. Figure 10 shows the training and test samples, visualized with t-SNE in their feature representation (*n.b.* t-SNE is a stochastic process and depicts only the spatial relation of points to one another, wherefore the unused inserts (blue markers) are now in the bottom-right corner).

Fig. 9 t-SNE representation of all samples used for model training, color ranges from new (blue) to worn-out inserts (red)

Fig. 10 t-SNE visualization of training samples (O) and test samples (x), color ranges from new (blue) to worn-out inserts (red)

Points of sparsely populated areas are usually impossible to classify. They indicate either a noisy sample or an outlier—or simply a sample from an intermediate state between two clusters. As Fig. 10 indicates, most test samples of the unused inserts (blue X) coincide with the training samples of the same bin (blue O), whereas others, especially the green and yellow shaded test samples, are trailing on the boundaries of the training samples. Furthermore, the overlapping clusters of the worn-out samples may impair the prediction accuracy. Both circumstances may provoke a large number of noise point classifications when applying the prediction with HDBSCAN. However, the remedy for especially scarce data sets is to calculate the membership vector. It indicates whether a sample is ambiguous (almost equal membership probabilities to multiple clusters), or just of low probability of belonging to one cluster. Both phenomena can be solved by using the highest probability of the membership vector in case of a noise point classification. For further model updates, and consequently denser clusters of the trained model, this additional step is not necessary any longer. In practice, if there is only one measurement to the sample, there is a strong risk of classifying a degraded tool as an outlier, and therefore as neither new nor degraded. However, in practice, multiple measurements should be taken when assessing tool conditions, resulting in a number of usually $n > 10$ samples, with which multiple classifications can be undertaken. This limits the risk of receiving a false positive indication, while allowing to obtain better a more accurate distinction between new and worn tools.

4.5 Test results and performance evaluation

For the given study, both approaches of HDBSCAN were tested: the *approximate predict* function, and the possibility to perform a new clustering with the point to predict as an

additional sample. The re-clustering with HDBSCAN showed consistently slightly better performances. The results of the clustering are visualized in Fig. 11. There are three main observations to be made: (i) The overall prediction precision and accuracy are satisfactory while some deviations and uncertainties persist, (ii) the binning introduces a staircase-like structure in the prediction values, and (iii) the higher VB values have a poorer prediction performance due to the overlap of multiple bins.

The prediction achieved a mean average error (MAE) of -0.0067 mm, and a root mean square error (RSME) of 0.0822 mm.

4.6 Discussion

The results are overall satisfactory, especially since the prediction of flank wear has rather low accuracy requirements in practice. However, especially the higher flank wear regions would require a more accurate prediction result, in order to optimize for tool changeover cost and quality cost avoidance. As the RMSE shows significant room for improvement versus the boundaries introduced by the binning procedure, the algorithm can be further tuned for higher performance. As this study is however a proof of concept encouraging further research in the field, it can be considered a starting ground. Furthermore, in a real-world setting, we would disregard and not train the healthy setting, and instead focus on the end of the degradation cycle. This significantly helps the density-based clustering to converge, as the denser areas of healthy samples do not distort the clustering tree. This could improve resolution, accuracy, and therefore significance of the predictions in the critical phases of the tool life cycle, to reduce losses incurred by premature tool change, quality non-conformities, and tool breakage. However, the demonstrated results indicate the strong capabilities of the proposed approach to

Fig. 11 Results of the prediction on the test set, actual flank wear VB values (blue) versus predicted VB values (orange)

distinguish densities of sparse distributions of sample points from milling in-process sensor recordings.

4.7 Conclusion

The proposed study has shown the effectiveness of unsupervised learning approaches for the prediction of milling tool wear, with a convincing accuracy of the prediction values over the entire degradation span. Moreover, it has shown favorable results with a very small data set, allowing to train and test a model of as few as 4 runs to failure with intermediary measurements. With the underlying density-based clustering approach, it is able to accommodate any foundation from very scarce to abundant data sets, allowing for continuous updates and improvements over time in use. The proposed approach is straightforward to implement, requires little tuning of hyperparameters, and is transferable to and from other approaches in process and component monitoring. It can therefore serve as a baseline to future time series-based condition monitoring, diagnosis, and prediction efforts.

4.8 Outlook

The given results combined with the findings of [9] demonstrate the promising applicability of unsupervised approaches for PHM purposes in manufacturing. It is clear that some improvements can and should be made to allow for a use of degradation predictions in production. However, it can be considered a first step. In future research efforts, this approach will be examined for in-process workpiece quality predictions. Besides the extension to other applications, there are a number of optimizations and refinements to be undertaken. A promising step is to develop a quality indicator for input data. The underlying data set used in this study has some clear drawbacks with very few samples in higher wear ranges, and also rather noisy and non-stationary sample recordings. It requires a large amount of additional data preprocessing, up to the point where a large number of sub-samples were discarded. Hence, an indication of minimum data quality required to obtain appropriate clustering results in conjunction with the number of bins is deemed helpful.

The selection of features is also crucial to obtain proper results. For the selection of filters that lead to favorable clustering results, e.g. a Bayesian hyperparameter optimization to find a global optimum is a large leap ahead. The combination of both factors would turn the suggested algorithm into a deterministic approach, without any need of adjustment of parameters. This would subsequently allow to incorporate more

operating conditions in a single model while maintaining a high level of precision and accuracy.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00170-021-07281-2>.

Funding Open Access funding provided by ETH Zurich.

Data availability The underlying data repository is publicly available, and link and descriptions are provided in reference [11].

Code availability The code is not publicly available, but will be provided upon request.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Tsui KL, Chen N, Zhou Q, Hai Y, Wang W (2015) 2015, Prognostics and health management: a review on data driven approaches. *Math Probl Eng*. <https://doi.org/10.1155/2015/793161>
2. Sarazin A, Truptil S, Montamal A, Lamothe J (2019) Toward information system architecture to support predictive maintenance approach. In: Popplewell K, Thoben K-D, Knothe T, Poler R (eds) *Enterprise Interoperability VIII*. Springer International Publishing, Cham, pp 297–306
3. Jimenez-Cortadi A, Irigoien I, Boto F, Sierra B, Rodriguez G (2020) Predictive maintenance on the machining process and machine tool. *Applied Sciences (Switzerland)*, 10/1, <https://doi.org/10.3390/app10010224>.
4. Kusiak A (2018) Smart manufacturing. *Int J Prod Res* 56(1–2): 508–517. <https://doi.org/10.1080/00207543.2017.1351644>
5. Wang L, Alexander CA (2015) Big data in design and manufacturing engineering. *Am J Eng Appl Sci* 8(2):223–232. <https://doi.org/10.3844/ajeassp.2015.223.232>
6. Wegener K, Gittler T, Weiss L (2018) Dawn of new machining concepts: compensated, intelligent, bioinspired, *Procedia CIRP - 8th CIRP Conference on High Performance Cutting (HPC 2018)*, 77:1–17.
7. Liao Y, Deschamps F, de Loures E, FR, Ramos LFP (2017) Past, present and future of Industry 4.0-a systematic literature review and research agenda proposal. *Int J Prod Res* 55(12):3609–3629. <https://doi.org/10.1080/00207543.2017.1308576>

8. Da Xu L, Xu EL, Li L (2018) Industry 4.0: State of the art and future trends. *Int J Prod Res* 56(8):2941–2962. <https://doi.org/10.1080/00207543.2018.1444806>
9. Gittler T, Scholze S, Rupenyan A, Wegener K (2020) Machine tool component health identification with unsupervised learning, *Journal of Manufacturing and Materials Processing*, 4/3:86, <https://doi.org/10.3390/jmmp4030086>.
10. Li X, Ding Q, Sun JQ (2018) Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliab Eng Syst Saf* 172:1–11. <https://doi.org/10.1016/j.ress.2017.11.021>
11. Tao F, Zhang M, Liu Y, Nee AYC (2018) Digital twin driven prognostics and health management for complex equipment. *CIRP Ann* 67(1):169–172
12. Wuest T, Weimer D, Irgens C, Thoben K-D (2016) Machine learning in manufacturing: advantages, challenges, and applications. *Prod Manuf Res* 4(1):23–45. <https://doi.org/10.1080/21693277.2016.1192517>
13. Gittler T, Gontarz A, Weiss L, Wegener K (2019) A fundamental approach for data acquisition on machine tools as enabler for analytical Industrie 4.0 applications, *Procedia CIRP - 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering (ICME 2018)*, 79/July:586–591. <https://doi.org/10.1016/j.procir.2019.02.088>
14. Gittler T, Stoop F, Kryscio D, Weiss L, Wegener K (2020) Condition monitoring system for machine tool auxiliaries. *Procedia CIRP* 88:358–363. <https://doi.org/10.1016/j.procir.2020.05.062>
15. Wang D, Tsui KL, Miao Q (2017) Prognostics and health management: a review of vibration based bearing and gear health indicators. *IEEE Access* 6:665–676. <https://doi.org/10.1109/ACCESS.2017.2774261>
16. Wu D, Jennings C, Terpenney J, Gao RX, Kumara S (2017) A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests, *Journal of Manufacturing Science and Engineering. Transactions of the ASME* 139(7):1–9. <https://doi.org/10.1115/1.4036350>
17. Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., et al., 2016, Deep learning and its applications to machine health monitoring: a survey, 14/8:1–14.
18. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A (2019) Deep learning for time series classification: a review. *Data Min Knowl Disc* 33(4):917–963. <https://doi.org/10.1007/s10618-019-00619-1>
19. Zhang L, Elghazoly S, Tweedie B (2019) AnomDB: unsupervised anomaly detection method for CNC machine control data. *PHM* 2019:1–12
20. Sun J, Rahman M, Wong YS, Hong GS (2004) Multiclassification of tool wear with support vector machine by manufacturing loss consideration. *Int J Mach Tools Manuf* 44(11):1179–1187. <https://doi.org/10.1016/j.ijmachtools.2004.04.003>
21. Agogino A, Goebel K (2007) Milling data set, NASA Ames Prognostics Data Repository [Online] Available: <https://ti.arc.nasa.gov/c/4/>. [Accessed: 01-Feb-2020].

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.