MACHINE TOOL

User-Specific Parameterization of Process Monitoring Systems

B. Denkena[1] · **H. Klemme**[1] · **J. Becker**[1] · **H. Blech**[1]

Abstract

Errors in milling processes such as tool breakage or material inhomogeneities are a major risk to the quality of machined workpieces. Errors like a broken tool may also lead to damages to the machine tool. Process monitoring systems allow for autonomous detection of errors, therefore, promoting autonomous production. The parameterization of these systems is a trade-off between high robustness (low false alarm rate) and high sensitivity. Even though several monitoring systems have been introduced for single-item and series production, a universal parameterization technique that weighs off sensitivity and robustness does not exist. In this paper, a novel, model-independent and adjustable parameterization technique for monitoring systems is introduced. The basis for the parameterization is the material removal rate that indicates the temporal and quantitative impact of process errors (ground truth). The ground truth allows calculation of the established F_β-score, which is used to evaluate the monitoring system. An adjustment of the β-parameter influences the weighting of sensitivity and robustness. Accordingly, the β-parameter allows to easily control the sensitivity-robustness trade-off so that the monitoring system is economic for the company's specific situation. In this paper, a look-up table for hyper-parameters of the state-of-the-art tolerance range monitoring model is provided using the introduced parameterization approach. With this table companies and researchers can set the hyper-parameters of their monitoring models for 5-axis-milled single items user-specifically. To demonstrate, that introduced parameterization approach works for different kinds of monitoring models, a one-class support vector machine (SVM) is parameterized also.

Keywords Monitoring · Parameterization · SVM · Machining · Hyper-parameter · Process control

1 Introduction

To manufacture parts of high quality, it is important to detect machining errors. Permanent supervision by a human operator is costly. Moreover, the error detection rate is limited to what can be identified by a human operator. While the role of human operators in today's machining shop floor turns more and more to an administrative job, the autonomous collection of reliable production information is of high interest [1]. Monitoring systems automatically monitor machining processes by acquiring data from the machine control (e.g., spindle current or control data) or external sensors (e.g., acceleration or acoustic emission). The spindle current for

example has proven to be very sensitive for errors in milling processes [2]. Based on uni- or multivariate signals, the monitoring system indicates an alarm. Depending on the user preference, an alarm may cause an emergency stop of the machine tool or a flashing warning light.

The difficulty faced by systems relying on hard-coded knowledge suggests that machine learning (ML) systems such as monitoring systems need the ability to acquire knowledge from raw data [3, 4]. ML systems have two parameter types: Model parameters, which can be updated through a training process, and hyper-parameters that cannot be estimated via training [5]. In Fig. 1, a monitoring model and its parameters are shown. Based on machining data, the monitoring model predicts either an alarm or no alarm. While the hyper-parameters may apply to multiple machining processes, the model parameters are process specific.

Monitoring models that are only trained with input data (unsupervised learning) have already been studied (Fig. 1 green checkmarks) [6, 7]. Additional to the training, a

✉ J. Becker
becker@ifw.uni-hannover.de

1 Institute of Production Engineering and Machine Tools, Leibniz Universität Hannover, An der Universität 2, 30823 Garbsen, Germany

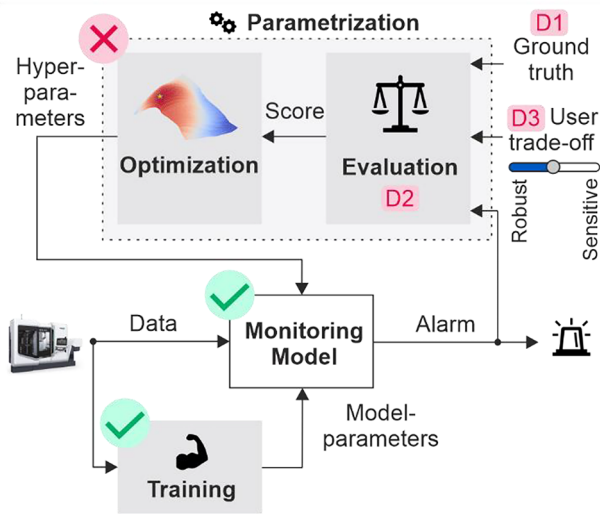set of hyper-parameters, therefore, does not exist. It depends on the preference of the production planer.

Fig. 1 Introduced approach for parameterization of monitoring systems: The model-parameters are only applicable for a single machining process, while the hyper-parameters can be applicable for multiple processes

parameterization is necessary. Not all sensors are sensitive towards process errors. This is because the measured signals contain noise and the accuracy of sensors is generally limited [8]. Therefore, false predictions are unavoidable. Based on key performance indicators (KPI) from production planning, the user decides with the parameterization how to deal with false predictions. A system can either be optimized towards high sensitivity or high robustness. A single, optimal set of hyper-parameters, therefore, does not exist. It depends on the preference of the production planer.

Different monitoring approaches for machining have been introduced and compared over the last years [9–12]. However, their parameterization is either not covered sufficiently or is specific to an individual user preference. Therefore, this paper proposes a universal parameterization approach for single-item and series production. It is argued that the poorly addressed parameterization in literature is caused by three major deficits (D), that are shown in Fig. 1.

Monitoring systems reviewed in previous work use different methods for labeling of errors [11]. Therefore, a common procedure for labeling errors does not exist (D1). Previous studies show detection rates for individually generated process errors [12]. However, as the detected errors are labeled with varying methods, the comparability of the models is limited significantly. The labeling defines a ground truth that specifies, whether an error exists or not. This work introduces a material removal simulation for labeling process errors. Even though simulations are a well-known

approach for monitoring of machine tools [13], the simulation is usually not used as ground truth for process errors. As shown in Fig. 1, the definition of ground truth is necessary to execute the parameterization.

A common procedure for evaluating monitoring systems does not exist (D2). Metrics like accuracy are still used, even though they depend on the error density and parametrization [11]. Universal evaluation functions from fundamental statistics, that are widespread in areas like computer science or medicine are rarely considered. A metric is necessary to parametrize monitoring models with the approach from Fig. 1. In this work the F_β-score is proposed as a common evaluation metric.

A method for intuitive manipulation of a user-specific trade-off between sensitivity and robustness does not exist (D3) [14]. Therefore, setting the robustness sensitivity trade-off with a single parameter is currently not possible.

For the mentioned reasons, in this work, a parameterization of two monitoring systems is conducted using the approach shown in Fig. 1. To use this universal parameterization, the deficits of defining a ground truth (D1), evaluating a monitoring system (D2) and defining a user-specific trade-off (D3) are addressed sequentially. Afterward, the universal parameterization approach is demonstrated for a state-of-the-art tolerance range model and a one-class support vector machine.

2 State of the art

In Fig. 2 an established "tolerance range" monitoring system is shown. Exemplary the spindle current is used as the monitored signal. This monitoring system defines a tolerance range that should not be exceeded. The tolerance range (green) surrounds an estimated optimal signal $\hat{i}_{sp}(t)$ (blue). Without any errors, the monitoring system expects the estimated spindle current $\hat{i}_{sp}(t)$. The estimated spindle current $\hat{i}_{sp}(t)$ and the measured spindle current $i_{sp}(t)$ are the input signals for the monitoring system. To generate a tolerance range, the safety factor s_y (hyper-parameter 1) is multiplied with the certainty of the estimated signal $c(\hat{i}_{sp}(t))$ and added (upper bound) and subtracted (lower bound) from the estimated spindle current $\hat{i}_{sp}(t)$. Therefore, a high certainty leads to a narrow tolerance range, while a low certainty leads to a wide tolerance range. Additionally, to s_y, a safety factor s_t (hyper-parameter 2) is applied to guarantee a secure distance in the temporal direction. A variety of extensions of the described model exist, e.g. distinguishing s_y for the upper and lower boundary [15]. However, these more complex approaches are usually special solutions and require more hyper-parameters for parameterization.
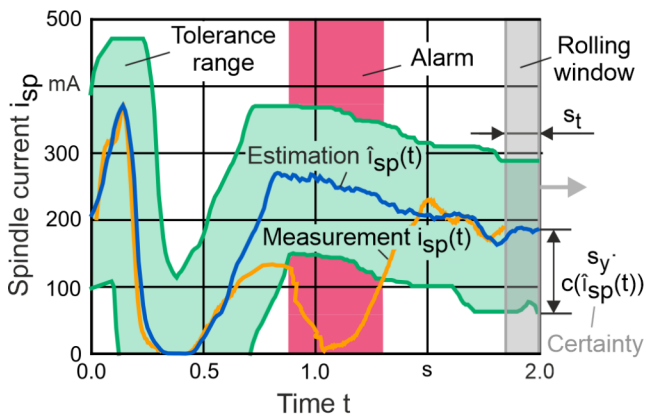
Fig. 2 Tolerance limits with constant distances

Fig. 3 Monitoring of serial and single item production

As shown in Fig. 3, the calculation of the estimated spindle current $\hat{i}_{sp}(t)$ and its certainty $c(\hat{i}_{sp}(t))$ differ for the production type. The certainty of the estimation is a number between zero and one that represents, how "certain" the system is about the predicted value. For example, might the certainty of the current prediction be lower at rapidly accelerating feeds then at a constant feed velocity. This has to be considered by the process monitoring system.

The type of production is determined by the repetitiveness of certain workpieces and by the number of pieces produced [16]. In series production, the spindle current is recorded for numerous processes (Fig. 3 left: black curves) to train the algorithm. Calculating the mean values for every timestamp $\mu(t)$ generates the expected spindle current $\hat{i}_{sp}(t)$ (Fig. 3 left: red line) [15]. The standard deviation in every time step $\sigma(t)$ (Fig. 3: exemplary shown for time t_1 and t_2) represents the certainty for the expected signal. A best practice in series production is setting factor s_y to a value of 6 [15]. With this parameter for a standard distributed signal, only 0.00034% of the samples are classified falsely [17]. As the false alarm rate is defined per sample, the actual occurrence of false alarms in production also depends on the sampling rate. The described approach only parametrizes a single hyper-parameter for series processes. However, without considering the error distribution this parameterization could still lead to an unnecessarily sensitive system.

In the case of single-item production, previously machined parts do not exist. Therefore, the estimated spindle current $\hat{i}_{sp}(t)$ must be modeled based on a material removal simulation. The tool position is the input for the material removal simulation. In Fig. 3 (right), the position-dependent estimation $i_m(pos)$ and the actual measurement $i_{sp}(t)$ are shown for an exemplary cutting process. Even though the model estimates the spindle current, the model does not provide certainty for this estimation. Hence, a constant certainty is calculated that does not change with time. The certainty is the mean squared error (MSE) between the measured and
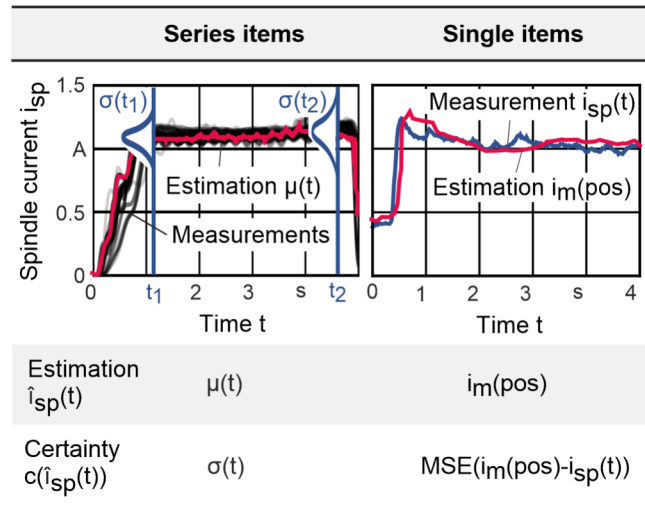
modeled spindle current for several parts (single items) that are used for the training of the model $MSE(i_m(pos)-i_{sp}(t))$.

3 Experimental setup and methodology

To examine the sensitivity and robustness of monitoring systems in a complex machining situation, a 5-axis-milling process is examined. An impeller geometry that comprises only free-form surfaces is used as an exemplary geometry. In Fig. 4, this impeller is shown. The milling process is performed using a ball-end milling tool with a diameter of 6 mm. In this manufacturing case, process monitoring is a challenge because the tool engagement conditions and the material removal rate change continuously. All experiments are performed with the turn-mill center DMG Mori NTX 1000. The machine is connected to a Beckhoff industrial computer that records positions, currents and control errors of all machine axes. The data acquisition rate is 10 ms.

The machining data of five milled impeller pockets are used to demonstrate the parametrization approach. The data is divided into training, validation and test data to ensure valid results, as shown in Table 1. All algorithms considered in this work classify as unsupervised anomaly detection. Anomalies are patterns in data that do not conform to a well-defined notion of normal behavior [18]. The training process defines normal behavior and, therefore, the training data does not contain errors. However, setting hyper-parameters (parameterization) requires processes errors. The parameterization is conducted based on the validation dataset. The validation dataset contains labeled process errors. For each dataset, different milling areas on the workpiece are used as segments to minimize overfitting. In Fig. 4, the last 12 of 60 segments are shown (tool path segments). The
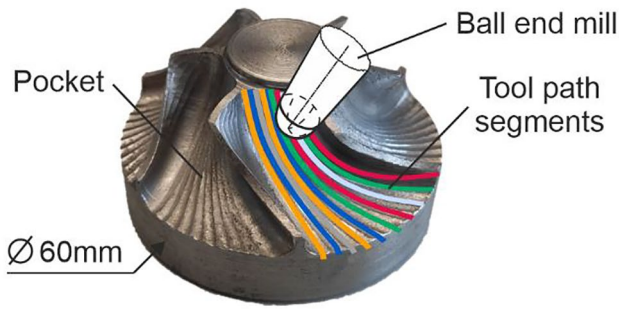
Fig. 4 Experimental workpiece and exemplary tool path

Table 1 Train-, validation- and test split

Dataset	Segments	Labels required	Contains errors
Training	0–30	no	no
Validation/ parameterization	31–45	yes	yes
Test	46–60	yes	yes

segments are spitted sequentially into training, validation and test data. Therefore, the 12 segments displayed in Fig. 4 belong to the test dataset.

4 Evaluation of monitoring systems

Process errors are required as a reference to evaluate the performance of a monitoring system. Therefore, process errors are introduced artificially in the experiments. In Fig. 5, a simulation of the unmachined part is shown. The part contains a notch that represents a material anomaly. Depending on the milling tool path, the notch has different impacts on the machining process. For certain positions, the tool only slightly "touches" the notch, whereas, for others, the tool is fully immersed into the notch, not cutting any material. Accordingly, the notch simulates a wide range of possible workpiece errors variants.

To determine optimal parameters for the process monitoring system and to evaluate the correctness of the predicted
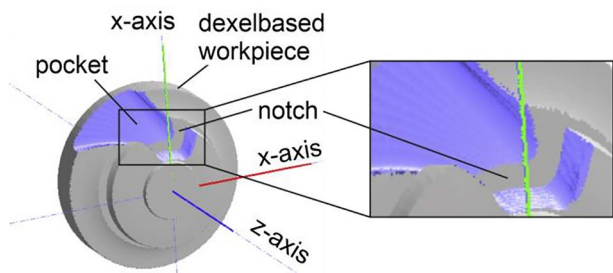
alarms, a metric is necessary. However, the verification of classified alarms requires knowledge about the quantitative and temporal impact of an error. This knowledge is referred to as ground truth. To generate this so-called ground truth, the simulation of the workpiece is executed with and without a notch. The difference in the material removal rate Q_w of both simulations constitutes the missing material due to the notch. In Fig. 6, the results of both calculations are shown. The differential material removal rate ΔQ_w identifies the milling affected by the material error. In the following evaluation and parameterization, this information is used as ground truth.

The differential material removal rate is quantitative ground truth for process errors. Therefore, the deficit D1 is addressed successfully. The approach labels the temporal occurrence and intensity of the errors. Also, the simulation of a material removal rate works for other cutting processes like turning or drilling [19].

The performance evaluation of a monitoring system is conducted through the comparison of the prediction with the simulation-based ground truth. In Fig. 7, the principle of evaluation is shown. The monitoring system uses the estimated data (spindle current $\hat{\imath}_{Sp}(t)$) and the measured data (spindle current $i_{sp}(t)$ and its change rate $\dot{i}_{sp}(t)$ considering the last five samples) to estimate if a process error is present. The prediction p is either an alarm (p = 1) or no alarm
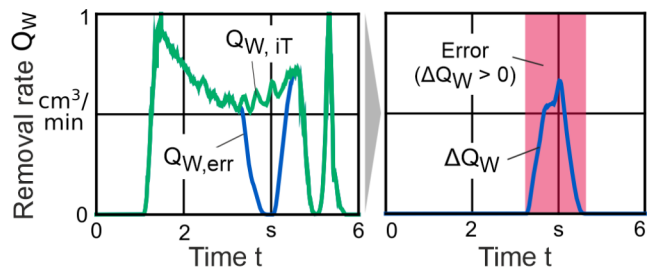

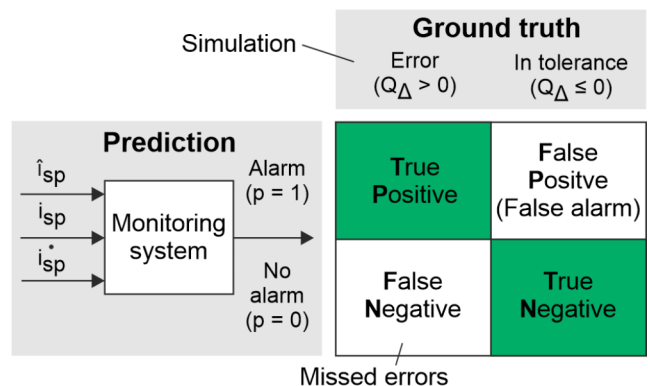**Fig. 6** Difference of material removal rate due to artificial error (ground truth)

Fig. 7 Monitoring system and confusion matrix

Fig. 5 Part with artificial material error (notch)

(p = 0). With the simulated ground truth, it is possible to evaluate whether a prediction is true or false. The confusion matrix shown in Fig. 7 differentiates the decisions of the monitoring system into four categories: True positive (TP), false positive (FP), false negative (FN), and true negative (TN).

An optimal monitoring system only returns true results, as seen in Fig. 7 marked green. However, given the signal noise and limited sensitivity of the sensors towards process errors, it is impossible to achieve only true predictions. Since a certain percentage of false alarms are not avoidable, it is necessary to decide how to cope with them. The values "recall" (1) and "precision" (2) constitute metrics that consider true and false classifications in a combined manner [20]. The recall (also called hit-rate or sensitivity) describes how many actual errors are detected. If a system only outputs alarms, the recall is one and the precision zero. The precision (robustness of the system) describes how many alarms are correct. If a system does not output any alarm the precision is one and the recall zero.

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

In the industrial context, this means that a high recall ensures a high detection rate of process errors but is accompanied by a higher percentage of false alarms. High precision, in turn, means that a high percentage of alarms is correct but also entails higher negligence of actual errors. An all-encompassing metric for the monitoring performance is calculated using the established F_β-score that combines precision and recall to give an overall evaluation [20].

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \quad (3)$$

Adjusting β weights off the formerly mentioned trade-off between precision and recall. The recall is β-times more important than the precision. Accordingly, if β is one, precision and recall are equally weighted [21]. The F_β-score is introduced as a general evaluation metric for monitoring systems. The parameter β allows setting the sensitivity-robustness trade-off with the single parameter. Therefore, the F_β-score solves the deficit of quantitative evaluation of monitoring systems (D2) and the deficit of a single parameter trade-off adjustment (D3) is solved with the parameter β.

5 Parameterization approach

Before parameterization, the model must be trained. In the case of the tolerance model, a single certainty value is optimized, which is trained by setting it to the root mean squared error (RMSE) between estimation and measurement for the training interval. For the considered training data, the certainty is 21.5 mA. To identify the ideal hyperparameters for the tolerance range model, the F_β-score is calculated for the validation data. With β set to a value of 1, F_β is calculated for a wide range of possible s_y- and s_t-values (Fig. 8). The simulated ground truth is used to evaluate the parameterization quality. A grid search for the $F_{\beta=1}$-score is conducted by varying the parameters s_y (0–8) and s_t (0–2 s). The grid search leads to global optimum at $s_y = 4.7$ and $s_t = 0.22$ s. This parameterization is the ideal balance between robustness and sensitivity of the system.

The described parameter optimization is repeated several times for a range of β values from 0.01 to 100. As the model has low complexity, the computation time is insignificant, less than 10 s. The resulting F_β-score is depicted in Fig. 9a. Additionally, the values for s_y and s_t that were identified during the optimization of the F_β-score are used to perform the classification and the calculation of precision and recall separately.

A strongly dominant weighting of either recall or precision in the optimization leads to an increase of the F_β-score. Regarding the corresponding precision and recall separately, a decrease of the lower weighted metric is observed. For a low β value, wide tolerance limits as depicted in Fig. 9b (left) are the result. In the context of a monitoring system for
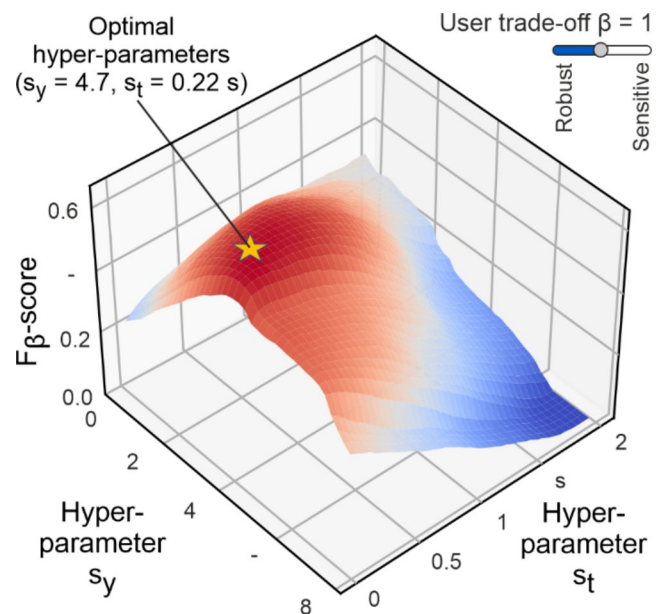

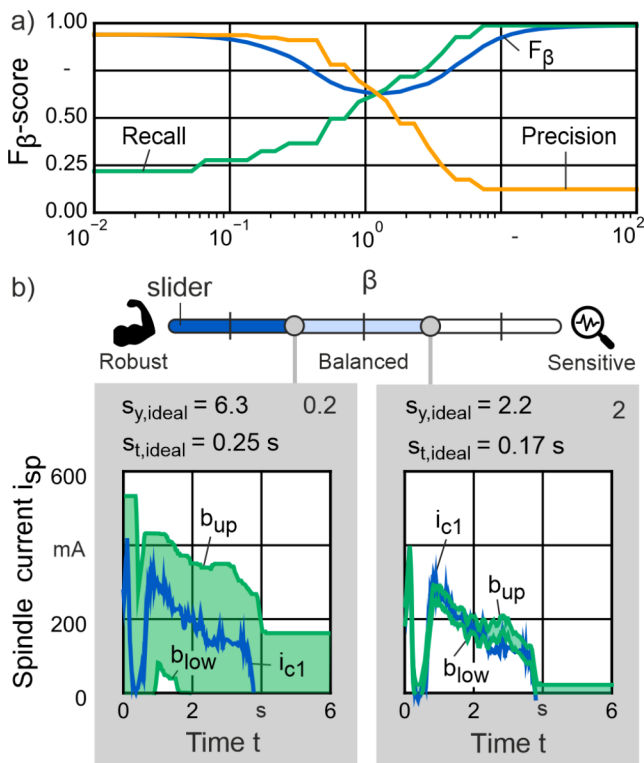**Fig. 8** Optimal parameters for the monitoring model

Fig. 9 Adjustment of the precision recall compromise based on the validation data

machining processes, this means fewer false alarms occur but the system is prone to miss actual errors. The case of high value for β is depicted in Fig. 9b (right). The high recall value contributes more to the F_β-score than the decreased precision. By introducing a slider for β in Fig. 9, the tolerance limits can either be pulled towards high robustness and few false alarms or to high sensitivity and few missed errors.

The evaluation of process monitoring systems using the F_β-score allows parameterizing process monitoring systems ideally. The introduced slider and the precision and recall graphs allow companies to set the trade-off between sensitivity and robustnesseasily and intuitively.

Table 2 is a look-up table for the optimal hyper-parameters that were identified in this paper for the tolerance range model. The look-up table has only proven validity for the introduced material anomalies and the impeller workpiece with the spindle current as the monitored signal. In future work the hyper-parameters should be compared with hyper parameters from different milling operations and defect types.

6 Generalization of the parameterization approach

In the preceding sections, a method is presented to identify an adjustable range of ideal hyper-parameters for a comparatively simple tolerance model. However, with the recent advances in machine learning, more general classifiers, like neural networks or support vector machines promise to increase performance. Therefore, the introduced parametrization approach is demonstrated on a more advanced and general classifier, the one-class support vector machine (SVM).

The considered one-class SVM uses the amplitude and the change rate of the spindle current as input parameters. Similar to the safety factors in the tolerance model, in a one-class SVM, the compromise of false alarms and not detected errors is defined by hyper-parameters. Unlike the traditional tolerance range model, the hyper-parameters of the one-class SVM do not correlate to the physical values. In Table 3 the major hyper-parameters of the one-class SVM in comparison to the hyper-parameters of the tolerance model are shown. The one-class SVM has a radial basis function (RBF) kernel. The parameter ν represents the margin of the one-class SVM. Intuitively, the γ-parameter represents how far the influence of a single training sample reaches. The γ-parameter can be regarded as the inverse of the radius of the influence of samples selected by the model as support vectors [22].

To achieve comparability between the tolerance range and the one-class SVM, the hyper-parameters of the SVM where optimized equivalent to the values s_y and s_t in the previous method. The computing time for a grid search

Table 2 Look-up table for user specific parameterization of tolerance range monitoring models for 5-axis milling

β	s_y	s_t	Interpretation
0.1	7.53	0.62	Robust
0.16	7.53	0.42	
0.25	7.53	0.22	
0.40	7.53	0.22	
0.63	4.74	0.22	
1	4.74	0.22	Balanced
1.58	3.81	0.22	
2.51	2.88	0.22	
3.98	1.02	0.02	
6.31	0.10	0.02	
10	0.09	0.2	Sensitive

Table 3 Hyper-parameters of the models

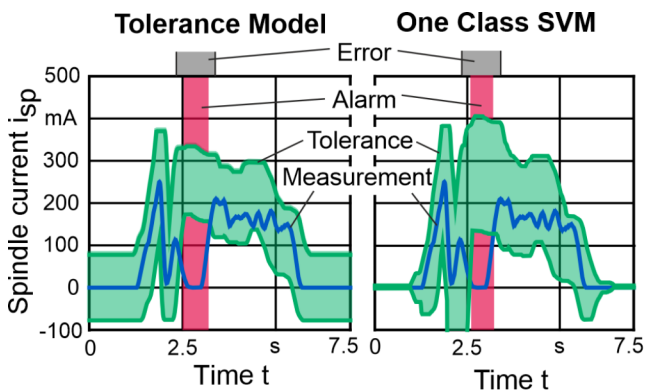| Model	Hyper-parameters
Tolerance model	S_y, S_t
One-class SVM	γ, ν, S_t

Fig. 10 Comparison of the monitoring boundaries of a traditional tolerance model and the one-class SVM

increases exponentially with the number of hyper-parameters. Therefore, for the one-class SVM, the grid search is replaced with a successive halving algorithm [23]. For every step during the successive halving algorithm, a training of the SVM based on the simulated ground truth is conducted. The classification is evaluated using the F_β-score. For both approaches the hyper-parameters that lead to the best F_β-score were used to calculate tolerance limits based on the estimated spindle current.

As shown in Fig. 10, the one-class SVM recognizes the artificial error zones where the material is missing. However, using the test dataset, the performance of the one-class SVM does not improve. Both monitoring algorithms detect 12 out of 15 errors and produce three false alarms in one milled pocket. In all cases, false alarms and undetected errors are caused by inaccurate current predictions.

The introduced parametrization approach works for both the existing, specially designed monitoring models and more advanced machine learning models like a one-class SVM. With the new parameterization method, it is possible to identify hyper-parameters for any kind of monitoring model and user preference. With the look-up table Table 2, a state-of-the-art tolerance range model can be parametrized for 5-axis milling without process errors.

7 Conclusion and outlook

This work presents an overall approach to parametrize any type of process monitoring system with a user specific sensitivity-robustness trade-off. A material removal simulation determines the exact amount of missing material in a workpiece. The simulated, missing material was used as a quantitative ground truth for process errors. This ground truth is the basis for the calculation of the established F_β-score, which evaluates the monitoring system. Since the trade-off between high sensitivity and high robustness towards false

alerts depends on the specific use case of a process monitoring system, the adjustment of the β value was introduced to set the desired weighing. The tailored evaluation of the monitoring output was used to optimize the hyperparameters of a tolerance range model and a one-class support vector machine. The influence of a varying the β value on the corresponding best set of hyperparameters and the results of the applied classification was presented in details for a tolerance range model. By using the same methodology with a one class SVM as monitoring model the generic applicability of the presented approach could be proven. Both classification algorithms exhibit similar results.

In future work the application on different types of machine tools and process errors is suggested. Especially artificial errors that lead to an increase of the spindle current, such as inserts of harder material should be considered as well as the transferability of hyperparameter sets between process configurations. Moreover, the approach could be advanced by replacing the one-class support vector machine with an even more advanced technique, like a one-class neural networks.

References

1. Prinz C, Morlock F, Freith S, Kreggenfeld N, Kreimeier D, Kuhlenkötter B (2016) Learning Factory Modules for Smart Factories in Industrie 4.0. Procedia CIRP 54:113–118. https://doi.org/10.1016/j.procir.2016.05.105.
2. Denkena B, Bergmann B, Becker J, Blech H (2021) Sensorlose Prozessüberwachung für die Einzelteilfertigung: Künstliche Intelligenz zur strombasierten Prozessüberwachung komplexer Bauteilgeometrien. wt Werkstattstechnik online 111:305–308
3. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press
4. Kißkalt D, Fleischmann H, Kreitlein S, Knott M, Franke J (2018) A novel approach for data-driven process and condition monitoring

systems on the example of mill-turn centers. Prod Eng Res Devel 12:525–533. https://doi.org/10.1007/s11740-018-0797-0.

5. Yang L, Shami A (2020) On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. Neurocomputing 415:295–316. https://doi.org/10.1016/j.neucom.2020.07.061.

6. Denkena B, Dittrich M-A, Noske H, Witt M (2020) Statistical approaches for semi-supervised anomaly detection in machining. Prod Eng Res Devel 14:385–393. https://doi.org/10.1007/s11740-020-00958-9.

7. Liang YC, Wang S, Li WD, Lu X (2019) Data-Driven Anomaly Diagnosis for Machining Processes. Engineering 5:646–652. https://doi.org/10.1016/j.eng.2019.03.012.

8. Lauro CH, Brandão LC, Baldo D, Reis RA, Davim JP (2014) Monitoring and processing signal applied in machining processes – A review. Measurement 58:73–86. https://doi.org/10.1016/j.measurement.2014.08.035.

9. Jemielniak K (1999) Commercial Tool Condition Monitoring Systems. Int J Adv Manuf Technol 15:711–721. https://doi.org/10.1007/S001700050123.

10. Teti R, Jemielniak K, O'Donnell G, Dornfeld D (2010) Advanced monitoring of machining operations. CIRP Annals - Manufacturing Technology 59:717–739. https://doi.org/10.1016/j.cirp.2010.05.010.

11. Kim D-H, Kim T, Wang X, Kim M, Quan Y-J, Oh JW, Min S-H, Kim H, Bhandari B, Yang I, Ahn S-H (2018) Smart Machining Process Using Machine Learning: A Review and Perspective on Machining Industry. Int J of Precis Eng and Manuf -Green Tech 5:555–568. https://doi.org/10.1007/s40684-018-0057-y.

12. Abellan-Nebot JV, Romero Subirón F (2010) A review of machining monitoring systems based on artificial intelligence process models. Int J Adv Manuf Technol 47:237–257. https://doi.org/10.1007/s00170-009-2191-8.

13. Altintas Y, Aslan D (2017) Integration of virtual and on-line machining process control and monitoring. CIRP Ann 66:349–352. https://doi.org/10.1016/j.cirp.2017.04.047.

14. Denkena B, Dahlmann D, Damm J (2015) Self-adjusting process monitoring system in series production. Procedia CIRP 33:233–238. https://doi.org/10.1016/j.procir.2015.06.042.

15. Brinkhaus J-W (2009) Statistische Verfahren zur selbstlernenden Überwachung spanender Bearbeitungen in Werkzeugmaschinen. PZH, Produktionstechn. Zentrum, Garbsen.

16. Mathieu L (2004) Wörterbuch der Fertigungstechnik III. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-18960-9.

17. Morgan J, Brenig-Jones M (2015) () Lean Six Sigma For Dummies, 3. Aufl. John Wiley & Sons, Hoboken

18. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A Survey. ACM-CSUR 41:1–58. https://doi.org/10.1145/1541880.1541882.

19. Böß V, Denkena B, Breidenstein B, Dittrich M-A, Nguyen HN (2019) Improving technological machining simulation by tailored workpiece models and kinematics. Procedia CIRP 82:224–230. https://doi.org/10.1016/j.procir.2019.04.157.

20. Chinchor N (1992) () MUC-4 evaluation metrics. In: Unknown (ed) Proceedings of the 4th conference on Message understanding - MUC4 '92. Association for Computational Linguistics, Morristown, NJ, USA, p 22. https://doi.org/10.3115/1072064.1072067

21. Dalianis H (2018) Clinical Text Mining. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-78503-5.

22. 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, Duchesnay E (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830

23. Hertel L, Collado J, Sadowski P, Ott J, Baldi P (2020) Sherpa: Robust hyperparameter optimization for machine learning. SoftwareX 12:100591. https://doi.org/10.1016/j.softx.2020.100591.

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