MM Science Journal | www.mmscience.eu
ISSN 1803-1269 (Print) | ISSN 1805-0476 (On-line)
Special Issue | HSM 2025
18th International Conference on High Speed Machining
October 15-16, 2025, Prague, Czech Republic
DOI: 10.17973/MMSJ.2025_11_2025142

HSM2025-44828

CONDITION MONITORING IN MACHINING USING A TRANSFORMER MODEL WITH IMPLICIT LABELING

T. Reeber[1]*, C. Neun[1], H.C. Mohring[1]
[1]University of Stuttgart, Institute for Machine Tools (IfW), Stuttgart, Germany
*Corresponding author; e-mail: tim.reeber@ifw.uni-stuttgart.de

Abstract

Tool wear represents a central challenge for manufacturing companies. The resulting workpiece rejects and machine downtimes cause significant costs. One difficulty lies in predicting the optimal tool change timing. In practice, two suboptimal scenarios occur: Either tools are changed too early, not fully utilizing their service life, or too late, which can result in quality losses or tool breakage. In the context of Industry 4.0 and manufacturing digitalization, large amounts of process data are continuously generated, enabling indirect process control of tool wear. The temporal dependence of process data and the multitude of influencing factors require the development of powerful analysis methods.

This paper examines the development of a concept for detecting tool condition using a Transformer-based approach in milling and drilling processes. The captured motor current of the machine axes is analysed. The concept uses implicit labelling of training data, utilizing only sensor signals from unworn tools. The Transformer encoder learns a representation of the unworn machining state, based on which a linear decoder performs time series prediction. The reconstruction error, i.e., the deviation between predicted and actual values, serves as an indicator of tool condition. Statistical parameters of the reconstruction error enable quantitative comparison between normal and worn tool behaviour.

Besides presenting the concept, the implementation, development of a suitable model architecture and determination of optimal hyperparameters are addressed.

Keywords:

Artificial Intelligence, Condition Monitoring, Machine Learning, Milling, Predictive Model, Process Control

1 INTRODUCTION

Cutting processes are an essential part of modern industrial manufacturing. The cutting tools used in this process are subject to continuous wear on the cutting edge, which significantly influences the quality and efficiency of the entire manufacturing process.

The technical impacts of tool wear are complex and consequential:

As the cutting edge becomes dull, the contact angle changes and the contact area between tool and workpiece increases. This leads to increased cutting forces and enhanced friction. The resulting increase in heat generation further accelerates the wear process, creating a self-reinforcing effect. Especially with thin-walled work pieces, the increased cutting forces can lead to deformations that compromise dimensional accuracy and induce unwanted residual stresses in the boundary zone. Additionally, the occurring flank wear deteriorates the surface quality of the workpiece [Klocke 2008], which, if the prescribed dimensions and surface values are not maintained, leads to the rejection of the affected workpieces.

A direct measurement of tool wear through continuous measurements is not practical in ongoing production, as the production flow would need to be interrupted for it. Therefore, possibilities for indirect process control are being researched, where tool wear is monitored based on process parameters that occur during machining.

In the context of Industry 4.0 and the increasing digitalization of manufacturing, large amounts of process data are continuously generated. These data contain valuable information about the condition of tools and machines but also present new challenges: The complexity and high volume of data require innovative approaches to data analysis and interpretation. Particularly the temporal dependency of process data and the multitude of influencing variables necessitate the development of powerful analytical methods [Yan 2017].

Deep learning approaches have established themselves as promising tools for tool condition monitoring. A significant breakthrough was achieved in 2017 with the Transformer architecture in the publication "Attention Is All You Need" by Vaswani et al. [Vaswani 2017]. This new type of deep learning model has fundamentally expanded the state of the art in various fields such as natural language processing, machine vision, and speech recognition, delivering remarkable results. The underlying attention mechanism enables learning complex relationships between dependent elements and recognizes patterns in data that might remain hidden with classical approaches. This capability makes the Transformer architecture a promising candidate for various industrial applications [Kamath 2022].

2 STATE OF THE ART

In the literature, there are different definitions for classifying a measurement method as direct or indirect. Ghani et al. define methods for direct wear detection as measurement techniques that capture the actual geometric changes caused by the wear of the tool [Ghani 2011]. Teti et al. define this more generally and refer to all metrological approaches for determining tool wear as direct, but not the evaluation of process variables that occur anyway, such as cutting forces or vibrations [Teti 2010]. In the state of the art, the definition of Teti et al. is followed when classifying. The commonality of both definitions is that wear measurements during machining are very difficult to perform due to the continuous contact between tool and workpiece are made almost impossible by the presence of cooling fluids [Ghani 2011].

Indirect methods of tool wear detection are based on the systematic recording of process variables using sensor systems. The sensor data are further processed and correlated or compared with tool wear conditions and forwarded to cognitive decision systems. These perform a comprehensive analysis and create a condition diagnosis. The result of this evaluation can be used in two ways: On one hand, it can be made available to the machine operator as a basis for decision-making. On the other hand, there is the possibility of transmitting the information directly to the numerical control of the machine tool. This enables automated initiation of adjustment and correction measures [Teti 2010].

The simple setup and good suitability for practical application are the essential advantages [Ghani 2011]. Wear detection during ongoing machining enables optimal determination of the tool change time - both when reaching the regular end of tool life and in case of sudden failure such as tool breakage. This increases productivity and reduces manufacturing costs [Giusti 1987].

The digital transformation of industry has led to an exponential increase in available process data in modern manufacturing systems. This development enables new approaches to data-driven machine monitoring that significantly improve both error detection and condition predictions.

The basic principle of these systems is based on training models with historical sensor data, from which characteristic features are extracted. During ongoing operations, the same features are extracted from the captured sensor signals, and the trained model makes automated decisions about the machine condition based on these.

A central challenge lies in identifying meaningful features for condition description. The available analysis methods - from classic time series analyses to complex time-frequency transformations - generate a broad spectrum of possible parameters. Advanced selection procedures are used to identify the most relevant features. These include statistical methods such as Fisher distribution as well as modern approaches of conditional density estimation using special decision trees (Conditional Density Estimation Trees - CDT).

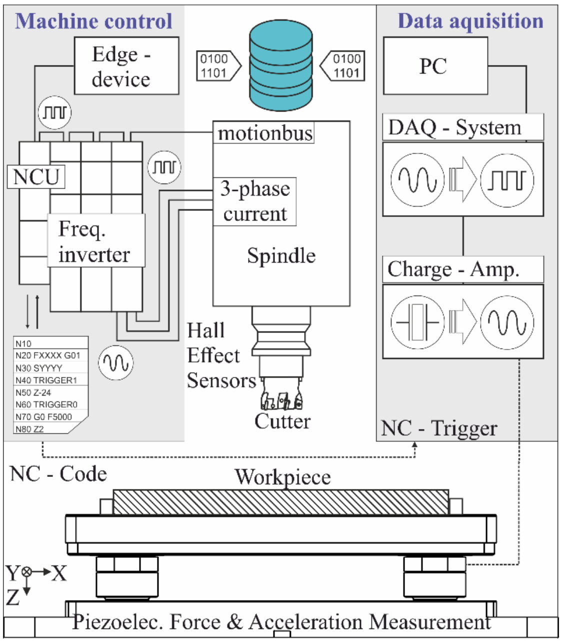
The selected features form the basis for various machine learning methods. In particular Support Vector Machines and linear regression models have established themselves as effective tools for condition prediction [Zhao 2018].

Deep Learning approaches have proven their high performance in various research areas - with significant breakthroughs in speech recognition, image classification, and natural language processing [Graves 2013, Krizhevsky 2017, Mnih 2009]. They are capable of learning high-quality feature representations directly from minimally processed or even unprocessed signals. Explicit feature extraction and selection before decision-making is not necessary, as the network implicitly extracts and filters relevant features. The network focuses on the information necessary for the model and ignores superfluous data. This end-to-end learning model significantly reduces the need for expert knowledge [Liu 2022].

3 METHODOLOGY

3.1. Machine tool and drive current measurement setup

In Fig. 1, the experimental setup for the data acquisition on the tool wear experiments can be seen. Machine data was gathered using an edge computer connected to the main control of the NC – control. Force data was collected simultaneously to identify the usefulness of the data for machine learning. The acquisition was triggered using the NC-Code. For a comparison of internal and external drive current data and the setup, see [Reeber 2024].



```
N10
N20 FXXXX G01
N30 SYYYY
N40 TRIGGER1
N50 Z-24
N60 TRIGGER0
N70 G0 F5000
N80 Z2
```

Fig. 1: Experimental setup for recording the machine data during the tool wear experiments.

3.2 Conceptual Approach

To use the historical sensor measurement data captured by the edge computer for wear detection, a Transformer model is trained for time series prediction of axis currents. The model learns the characteristic current patterns of non-worn tools. This is based on the physical relationship between wear and current consumption: A worn cutting edge requires higher cutting forces, which leads to increased power consumption and thus higher motor currents. This relationship is also evident in the cascade control, where the control deviation of the axis current is compensated by adjusting the control voltage of the electric motor via pulse width modulation. The electrical power as a product of voltage and current directly correlates with the mechanical power (product of cutting force and cutting speed).

As input, the model receives the currents of the X, Y, Z, and spindle axes, as these together represent the relevant system states of the milling process. The multivariate approach allows the model to recognize correlations and dependencies between the axes. This is particularly important since changes in one axis can have immediate effects on others. By considering all relevant axis currents, more complex patterns can be recognized, and the prediction accuracy increases. For training, only the first 5 meters of tool cutting path length from each test series are used, representing non-worn tool behaviour. It should be noted that apart from using a sliding window for sequence generation and scaling the model inputs, no additional feature engineering was employed, allowing the model to train directly with the raw sensor data.

In the following Fig. 2 the current of the spindle axis for different wear states is shown. This measurement comes from the DMG Mori Lasertec 65 DED hybrid. Three characteristic phases can be identified. At the beginning, the current increases significantly, reflecting the acceleration phase of the motor. The negative sign of the current indicates the flow direction and depends on the type of measurement. After this acceleration phase, the current pattern transitions into the milling process. Subsequently, the stable operating phase follows, in which the load current level remains constant. In the Zoom illustration, the entry of each cutting edge is clearly recognizable. At the beginning of the cutting-edge entry, the tool is loaded by the interaction with the workpiece. This sudden additional load requires higher power from the electric motor, which is reflected in an increase in current. The final phase is characterized by the braking phase of the electric motor. The relationship between the higher average current level of the considered axes and the wear is clearly recognizable.

The blue curve represents the spindle current measured with a new, unworn tool. The orange curve demonstrates the current pattern observed with a tool experiencing normal wear, while the gray curve illustrates the current signature of a highly worn cutting tool. The legend box indicates the cutting distance traveled by each tool.

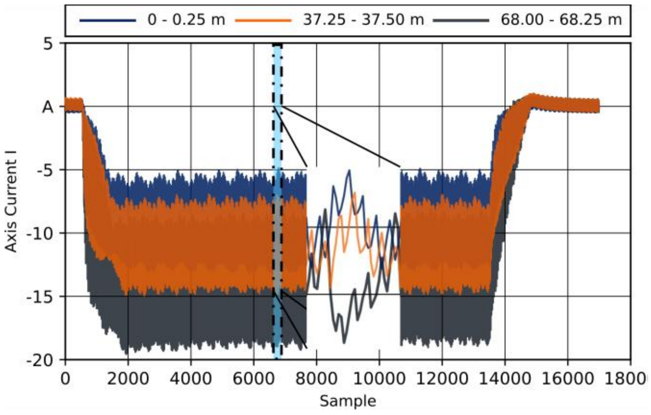


Fig. 2: Comparison of Spindle Current Patterns with Different Tool Conditions.

3.3 Model Implementation and Training

The chosen architecture of the Transformer model is shown in the following Fig. 3.

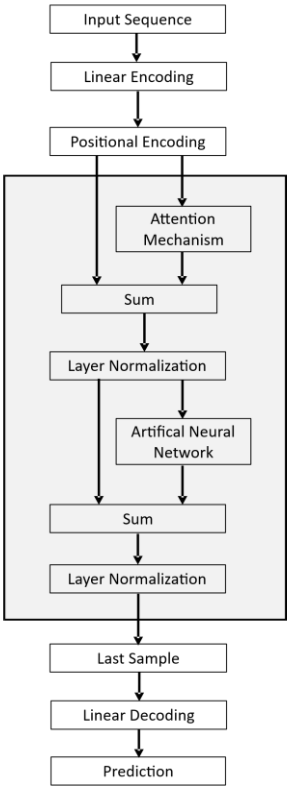


Fig. 3: Chosen Architecture for the Transformer Model.

The features must be normalized before they can be passed to the transformer model, as otherwise an unequal weighting would occur due to the irregular value ranges of the different features. Furthermore, this is intended to reduce the risk of exploding or vanishing gradients. The features are scaled to a value range between 0 and 1.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

Here, the maximum and minimum values are determined in advance to ensure the model already covers the axis currents predominantly used in industry.

The values for x_{max} and x_{min} for all axes are shown in the following Tab. 1.

Feature	x_{min} [A]	x_{max} [A]
Current Spindle-Axis	-200	200
Current X-Axis	-50	50
Current Y-Axis	-50	50
Current Z-Axis	-75	75

Tab. 1: Scaling Values

The generation of training and validation data for the transformer model is carried out using a sliding window approach. From each measurement record, overlapping time sequences of a defined length are extracted. The selected offset t_{offset} of four time steps increases the temporal distance between the input sequence and the value to be predicted. This increases the complexity of the prediction task and prevents the model from relying too heavily on immediately preceding values of the sequence. This approach maximizes the number of training examples per measurement record.

This formula defines how the individual sequences are extracted from the original dataset and linked to their corresponding target values.

$$f(X_i(t)) \rightarrow y(t) \qquad (2)$$

$$X_i(t) = [x(t_{start} + t_{shift}), \dots, x(t + t_{shift} + l_{seq})] \quad (3)$$

$$y(t) = x(t + t_{shift} + l_{seq} + t_{offset}) \quad (4)$$

Where:

t_{start}: Starting point from which sequences are generated, here t = 0.

t_{shift}: Sliding window for sequence generation

l_{seq}: Sequence length

y: Target Value, $t_{offset} = 1$ for the subsequent value, here $t_{offset} = 4$.

The same positional encoding as in the publication "Attention Is All You Need" by Vaswani et al. [Vaswani 2017] is used, which introduced the transformer architecture with attention mechanism. This encoding adds deterministic information about the sequence order to each timestep and its features, as the transformer architecture itself has no inherent way of understanding sequential relationships. The encoding involves sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (5)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (6)$$

Here, *pos* refers to the position of the token or sample within the sequence. The variable *i* is the respective dimension index and d_{model} is the chosen dimension of the transformer encoder. By using these trigonometric functions, each position in the sequence receives a unique encoding pattern that the model can learn to interpret, allowing it to understand the temporal relationships between different timesteps in the input data sequence.

The model's hyperparameters were established through comprehensive optimization using the Python library Optuna. This tool systematically evaluated various hyperparameter combinations by measuring their impact on error metrics. Parameters that couldn't be determined through Optuna's automated process—such as sequence size and loss function—underwent cross-validation to identify their optimal values. In the following Table 2, the evaluated hyperparameters are presented with the selected optimal values highlighted in bold.

Parameter	Value
Sequence Size	16, **32**, 64, 128
Loss Function	**Mean Absolute Error**, Mean Squared Error
Dimension Transformer-Encoder	32, 64 ,**128**, 256
Number Attention Heads	**1**, 2, 4
Number Attention Layers	1, **2**, 3
Learning rate at the Beginning of Training	1e-6 to 1e-3: **0,0009**
Dropout Rate	0,1 to 0,3: **0,15**
Weight Decay	**0,00001**
Learning Rate Reduction	0,1 to 0,9: **0,33**
Batch Size	**128**

Tab. 2: Hyperparameter Evaluation and Selection.

The Transformer model was trained with the programming language Python 3.11.9 using the PyTorch framework utilizing the Graphical Processing Unit with the CUDA 12.5 interface. In detail, Table 3 displays the specific hardware components that were employed in this study.

Component	Specification
Operating System	Windows 10 Enterprise Edition 64-bit
RAM	128 GB
Processor	Intel Core i9-10920X 3,50 GHz
Graphical Processing Unit	NVIDIA Quadro RTX 4000 8 GB VRAM
Hard Drive	Toshiba SSD 2 TB

Tab. 3: Employed Hardware Components

The model underwent a comprehensive training process spanning 93 epochs, with each epoch requiring approximately 182 seconds to complete, resulting in a total training duration of nearly 4,7 hours.

3.4 Tool State Classification Logic

For classifying the tool as worn, reference distributions were created for both unworn and worn tool inserts. A higher tool cutting path length of 30 meters for normal tool behaviour was selected to ensure that slightly worn tools are also classified as normal. Additionally, a reference distribution for worn tool behaviour was created from every measurement file with a tool cutting path length exceeding 60 meters. This 60-meter threshold was established because all tool inserts surpassed a flank wear of 150 micrometres at this point. The highest measured flank wear of a tool insert after 60 meters exceeded 1700 micrometres.

For all distributions, only data points between samples 2000 and 8000 were utilized. This selective sampling was implemented because the model exhibits higher reconstruction errors during the accelerating and braking phases. The model couldn't learn these transition phases as effectively as the in-process patterns since each measurement file contains only one accelerating and one braking phase, providing limited training examples for these states compared to the abundant in-process data. The obtained reference distributions are shown in the following Fig. 4.

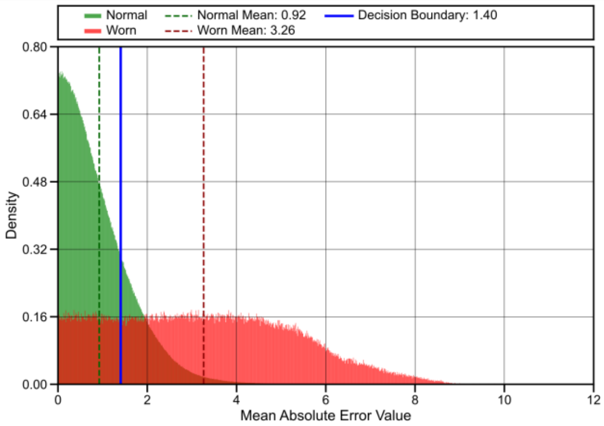


Fig. 4: Reference Distributions for Classifying the Tool State.

During processing, measurement data are continuously transmitted while the model simultaneously generates predictions, allowing reconstruction errors to be calculated in parallel. To reduce the influence of short-term fluctuations and enable a reliable comparison with the reference distributions, the mean of the last 200 reconstruction errors is used as the basis for condition assessment. This mean absolute error is then classified relative to the reference distributions, and its corresponding percentiles are determined. These percentiles are subsequently converted into probabilities using sigmoid functions, which—unlike linear mappings—offer tuneable parameters for task-specific calibration.

The functions for the sigmoid curves are show in Equation 7 and 8.

$$P_{normal}(x) = 1 - \frac{1}{1+ e^{(-\alpha(x-50))}} \quad (7)$$

$$P_{worn}(x) = \frac{1}{1+ e^{(-\alpha(x-50))}} \quad (8)$$

In the following Fig. 5, the sigmoid functions for probability conversion for normal and worn tool behaviour are displayed. The tuneable parameter alpha is set to $\alpha = 0{,}1$, which controls the steepness of the transition between the two states in the probability curves.

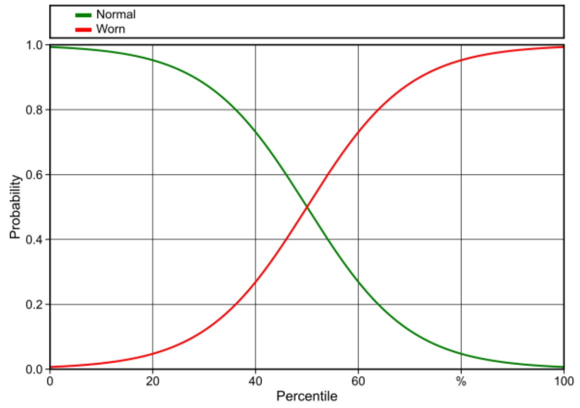


Fig. 5: Sigmoid functions for probability conversion.

The calculated probabilities must then be normalized with the Equations 9 to 11.

$$P_{total} = P_{normal} + P_{worn} \quad (9)$$

$$P_{normal,normalized} = \frac{P_{normal}}{P_{total}} \quad (10)$$

$$P_{worn,normalized} = \frac{P_{worn}}{P_{total}} \quad (11)$$

A mean absolute error value of 1.4 results in the 77.63 percentile for the normal unworn distribution and a 22.63 percentile of the worn distribution. After probability conversion and normalization this evaluates approximately to a 50% probability for both tool states and marks the decision boundary between both distributions.

4 VALIDATION

4.1 Model Evaluation

The achieved model quality of the model trained with data from the DMG Mori Lasertec 65 DED is examined. For this purpose, the result of the time series prediction is considered.

In the following Fig. 6 and Fig. 8, we present the transformer model's predictions on previously unseen test data, which was measured on the same machine alongside the true measured values for the spindle axis. The resulting absolute reconstruction error is displayed for both normal unworn (Fig. 7) and worn tools (Fig. 9), providing a clear visualization of the model's performance in distinguishing between these two states.

As mentioned earlier the model exhibits higher reconstruction errors during the accelerating and braking phases. But it is evident that the reconstruction error increases as expected.

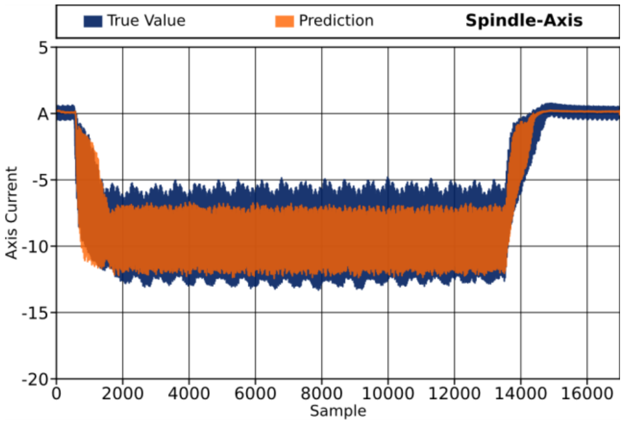


Fig. 6: Transformer Prediction and real measured values for a normal unworn tool.

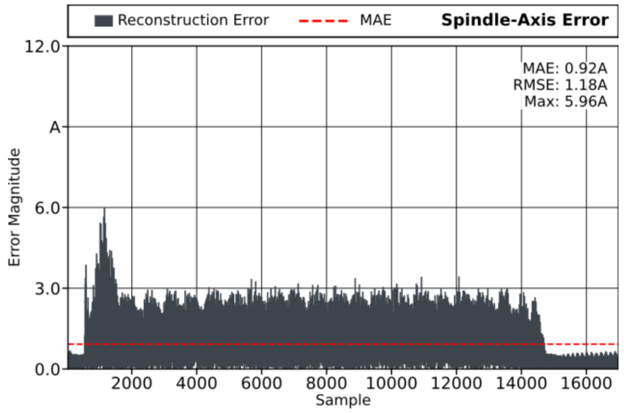


Fig. 7: Absolute Reconstruction Error for a normal unworn tool.

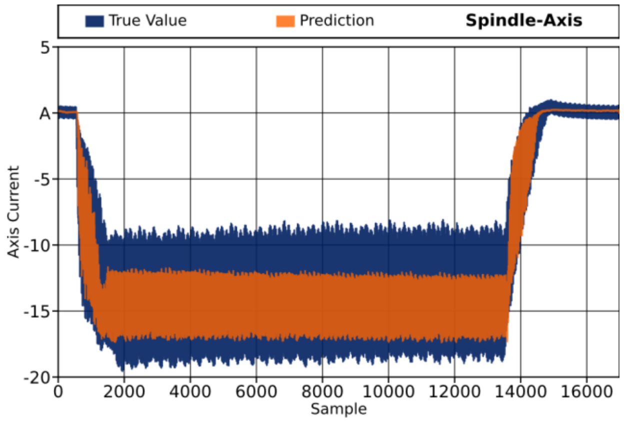


Fig. 8: Transformer Prediction and real measured values for a worn tool.

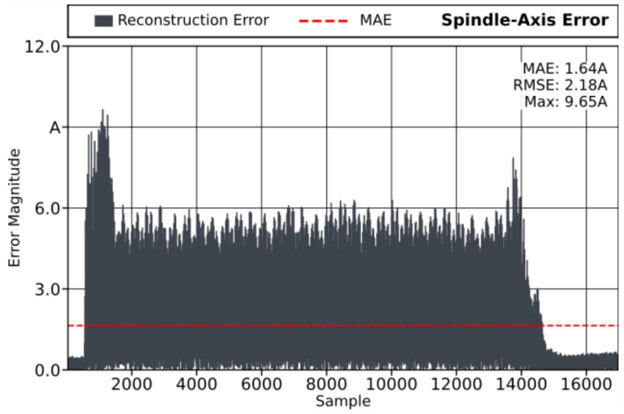


Fig. 9: Absolute Reconstruction Error for a worn tool.

4.2 Classification Logic Evaluation

The classification logic was validated by simulating the cutting process using all current measurements from the test dataset, for which the corresponding tool cutting path was known. Due to reduced model performance during acceleration and deceleration phases, the evaluation was limited to the steady state cutting phase.

Fig. 10 shows the proportions of the classification results based on the mean absolute reconstruction errors calculated using a sliding window of 200 points for each measurement file, alongside the corresponding mean spindle current. The results demonstrate a clear transition pattern in tool condition. Throughout the initial segment (0-30 meters of tool cutting path length), the classification predominantly indicates normal tool condition, except for a brief deviation around the 20-meter mark where a small proportion of worn tool status is detected. Beyond approximately 35 meters, the classification results show a significant shift towards the worn tool status, which becomes increasingly dominant for the remainder of the cutting path. The progressive reduction of the mean spindle current correlates strongly with the transition from normal to worn tool condition, providing additional evidence for the validity of the classification logic.

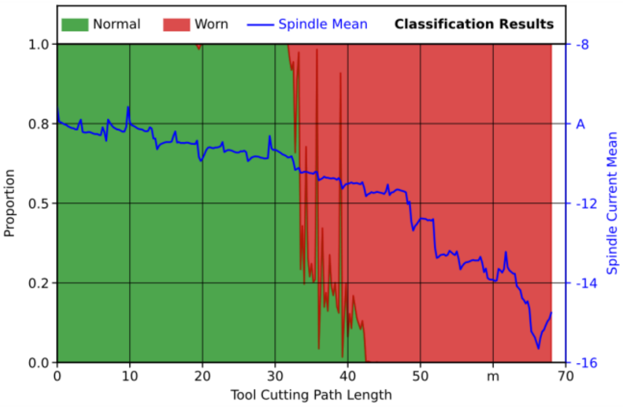


Fig. 10: Results of the Classification Logic Evaluation.

5 SUMMARY AND OUTLOOK

This study evaluated a transformer-based approach for tool wear classification using an implicit labelling concept. The model was trained exclusively on historical motor current data from unworn milling tools, eliminating the need for explicitly labelled worn tool data.

The transformer encoder learned to represent the normal unworn machining state, enabling a linear decoder to perform time series prediction. The reconstruction error-the difference between predicted and actual values-served as key indicator of tool condition.

Beyond presenting the conceptual framework, this research provided comprehensive details on implementation, model architecture and optimized hyperparameters. Evaluation demonstrated strong overall model performance with limitations primarily in acceleration and deceleration phases due to their limited representation in the training dataset.

For practical application, we developed and validated a classification algorithm based on reference distributions that converts reconstruction error percentiles to probabilities using a sigmoid function, testing it through simulations with independent test data.

6 ACKNOWLEDGMENTS

This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within ”Learning Production Technology - use of Artificial Intelligence (AI) in Production (ProLern)” and managed by the Project Management Agency Karlsruhe (PTKA). The author is responsible for the content of this publication.

7 REFERENCES

[Klocke 2008] Klocke, F. and König, W. Fertigungsverfahren – Drehen, Fräsen, Bohren, Springer Berlin, Heidelberg 2008. doi: 10.1007/978-3-540-35834-3.

[Yan 2017] Yan, J. et al. Industrial Big Data in an Industry 4.0 Environment: Challenges, Schemes, and Applications for Predictive Maintenance. IEEE Access, October 2017, Vol. 5, pp. 23484-23491. doi: 10.1109/ACCESS.2017.2765544.

[Vaswani 2017] Vaswani, A. et al. Attention Is All You Need. In: Von Luxburg, U. et al., eds. Proceedings of the 31st International Conference on Neural Information Processing Systems. doi: 10.5555/3295222.3295349.

[Kamath 2022] Kamath, U. et al. Transformers for Machine Learning: A Deep Dive. Boca Raton: CRC Press, 2022. doi: 10.1201/9781003170082.

[Ghani 2011] Ghani, J.A. et al. Monitoring online cutting tool wear using low-cost technique and user-friendly GUI. Wear, 2011, Vol. 271, Issues 9-10, pp 2619-2624. doi: 10.1016/j.wear.2011.01.038.

[Teti 2010] Teti, R. et al. Advanced monitoring of machining operations. CIRP Annals, 2010, Vol. 59, Issue 2, pp 717-739. doi: 10.1016/j.cirp.2010.05.010.

[Giusti 1987] Giusti, F. et al. On-Line Sensing of Flank and Crater Wear of Cutting Tools. CIRP Annals, 1987, Vol. 36, pp 41-44. doi: 10.1016/S0007-8506(07)62549-9.

[Zhao 2018] Zhao, R. et al. Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks. IEEE Transactions on Industrial Electronics, Feb. 2018, Vol. 65, No. 2, pp 1539-1548. doi: 10.1109/TIE.2017.2733438.

[Graves 2013] Graves, A. et al. Speech recognition with deep recurrent neural networks. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. doi: 10.1109/ICASSP.2013.6638947.

[Krizhevsky 2017] Krizhevsky, A. et al. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, May 2017, Vol. 60, pp 84-90. doi: 10.1145/3065386.

[Mnih 2009] Mnih, A. and Hinton, G. A scalable hierarchical distributed language model. In: Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference, September 2009, pp 1081-1088.

[Liu 2022] Liu, R. et al. Multi-sensor Data Fusion and Feature Extraction for Cutting Tool Condition Monitoring: A Review. In: Proceedings of the 2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN). doi: 10.1109/ICICN56848.2022.10006480.

[Reeber 2024] Reeber, T. et al. A Comparative Analysis of Axis Drive Current Measurements in CNC Machine Tools for Machine Learning Assisted Process Monitoring. Procedia CIRP, 2024, doi: 10.1016/j.procir.2024.10.078.