Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26095-cd

Reliability of a deep learning-based early warning system for table milling safety

Franziska Wolny[1], Philipp Busch[2], Nico Steckhan[1], Atmika Bhardwaj[1], Krutarth Prajapati[1], Muhammad Aziz Hassan[1], Hartmut Eigenbrod[2], Christoph Birenbaum[2], Sven Jacob[1], Silvia Vock[1]

[1] *Federal Institute for Occupational Safety and Health (BAuA), Germany.*
[2] *Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Germany.*

We present a proof of concept for a deep learning–based early warning system that forecasts kickback events in hand-fed table milling from multivariate sensor time series. To address the lack of woodworking datasets for acute safety events, we generate a dedicated dataset using a custom experimental setup that reproducibly induces kickback and records synchronized active power, spindle speed, and triaxial vibration data under multiple process configurations. A hybrid 1D CNN–attention–BiLSTM model is trained on standardized sliding windows with class-imbalance handling and threshold optimization to balance detection performance and false alarms. Evaluation across milling configurations shows second-scale early warning lead times with millisecond inference latency, but also highlights limited transferability between configurations and insufficient robustness for safety-critical deployment, motivating further validation and methodological improvements.

Keywords: table milling, machinery, anomaly detection, AI, machine learning

1. Introduction

The construction sector, particularly the wood-processing industry, exhibits elevated accident rates, with milling machines being a major source of injuries [1, 2]. Table milling machines are widely used due to their versatility in processing wood and wood-based materials, with workpieces fed either by power feeders or manually, especially in craft settings, by guiding them towards the rapidly rotating tool [3].

A major accident scenario for table milling machines is “kickback”. This is defined as the sudden, uncontrolled ejection of the workpiece in the opposite direction to the feed due to its engagement with the rotating milling tool. Kickback, particularly during manual plunge or test milling on spindle molders, poses a serious safety risk and can cause severe injuries and equipment damage [4]. Although passive safety measures exist, such as using clamping jigs for short workpieces, they are often not used in practice [3]. More recently, mechatronic safety systems have been introduced, primarily for circular sawing, using sensor-based detection to identify hazardous hand–tool proximity and trigger protective actions [5]. Altendorf Group presents an AI-based optical safety system with an extended pre-warning zone, illustrating the greater flexibility and reduced constraints of AI-based approaches for detecting hazardous situations [6].

In this paper, we present a customized table milling machine setup for generating training data for an AI-based monitoring system that can predict kickback. Section 2 discusses the phenomenon of kickback and current approaches to anomaly detection in time series. Section 3 introduces the experimental setup to record kickback events. The resulting data is described in Section 4. Section 5 describes the methods used to create the prediction model, followed by the results and discussion in Section 6. Finally, Section 7 suggests possible further research directions.

2. Background

2.1. *Kickback in Milling Processes*

When milling with stationary milling machines, common causes of kickback in hand-fed workpieces include feeding the workpiece in the wrong direction, excessive depth of cut, dull tooling, and inadequate hold-down or missing/poorly adjusted stops. Fig. 1 details a typcical sequence of a kickback event: (1) initial situation before milling; (2) continuous milling, whereby the workpiece

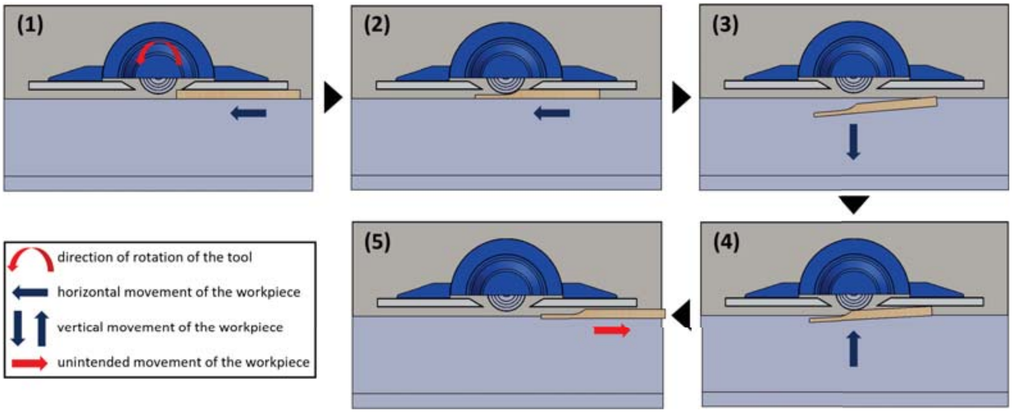

Fig. 1.: Sequence leading to kickback in handheld milling.

is pushed by the user against the direction of rotation of the tool and pressed against the stop; (3) sudden lift-off of the workpiece by the user reduces normal force and guidance; (4) renewed contact creates a large instantaneous chip thickness and sudden cutting engagement; (5) the tool's tangential force accelerates the workpiece against the milling direction, i.e., kickback occurs. The workpiece is torn from the operator's hands or hurled toward them. Kickback can cause severe injuries, including lacerations, fractures, crush injuries, and eye trauma, and can be fatal in extreme cases.

2.2. *Anomaly Detection in Industrial Settings*

An extensive survey of various anomaly detection methods including deep learning–based approaches is presented in [7], highlighting that deep learning outperforms traditional methods as the data scale increases due to its ability to learn hierarchical, high-level representations automatically. This automatic feature learning removes the need for manual domain-specific engineering and enables end-to-end solutions from raw data. Deep learning has become prominent for monitoring industrial machinery, particularly through the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to process time-series sensor data. Zhang *et al.* demonstrated the success of adaptive CNNs in extracting local features from raw vibration signals in bearing fault diagnosis, thereby establishing a baseline for modern sensor analysis [8]. More complex hybrid architectures have since emerged to capture temporal dependencies. For instance, Schlagenhauf *et al.* utilized a convolutional autoencoder to detect anomalies in the milling of steel, focusing on reconstruction errors to identify process deviations [9]. Similarly, Cooper *et al.* applied generative adversarial networks (GANs) to acoustic signals for detecting tool breakage [10]. However, these approaches predominantly focus on metal machining and gradual tool wear. Metal cutting is characterized by homogeneous materials and predictable force models. In contrast, woodworking introduces significant non-linearity due to anisotropic material properties (e.g., grain direction, knots). Furthermore, most existing models target maintenance metrics, such as remaining useful life (RUL), rather than acute safety events like kickback, which require real-time detection of rapid, transient precursors instead of long-term degradation trends.

The development of data-driven safety systems is currently limited by a lack of domain-specific datasets. Standard public benchmarks, such as the UC Berkeley Milling Data Set [11] or the IMS Bearing Data [12], focus exclusively on tool wear propagation and rotating machinery faults in metal processing. These datasets are well-suited for RUL prognosis but lack the acute, high-impact

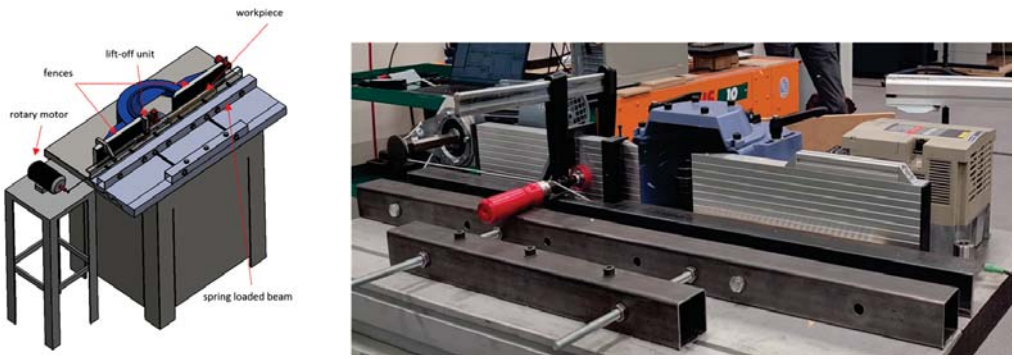

Fig. 2.: Left: Schematic of the experimental setup with traction unit, lateral guidance, and lift-off mechanism for the wooden slats. Right: Photograph of the setup. The vibration sensor is visible at the right edge.

kickback events characteristic of woodworking accidents. There is a distinct absence of datasets and systems that utilize process dynamics (milling speed, power, and vibration) to forecast hazardous machine states before they occur. Our work addresses this gap by providing a unique dataset focused on wood milling kickback and presenting a hybrid CNN-Attention-LSTM architecture for prediction.

3. Experimental Setup

Experiments were performed on a vertical spindle molder machine of type Bernardo T 800 F to investigate kickback during hand-fed milling. The machine offers three spindle speeds (e. g. 3090 or 6080 rotations per minute) selectable via a belt drive with stepped pulleys. Rebating cutterheads of type Falzmesserkopf 125 x 50 x 30 mm - Z4 / V4 from Bernardo were used. Spruce wooden slats with the dimensions 400 x 60 x 37 mm were milled in the experiments. The cutter engagement can be adjusted continuously, both in height (vertical tool position) and in depth of cut (lateral infeed at the fence).

A dedicated fixture was developed to simulate manual milling. Forward feed of the slats in the cutting direction is generated by an integrated traction unit, see Fig. 2: a hook mounted on the end face of each slat serves as the attachment point for a rope that is wound by a rotary motor with a take-up reel, pulling the workpiece through the cut. Lateral guidance is maintained by a spring-loaded beam that presses the workpiece against the fence.

In routine workplace operations, brief operator inattention can cause kickback. In our experimental setup, this is simulated by a lift-off mechanism. The essential components are a wedge to lift the workpiece and a retraction unit that removes the wedge, see Fig. 3. The wedge is mounted in contact with the left fence. As the motor pulls the workpiece up the wedge, the workpiece first departs the fence and then separates from the cutterhead. The lift-off increases until, at the wedge's end, the workpiece trips a retaining element that abruptly retracts the wedge. Concurrently, the pull rope disengages from the workpiece hook. The spring-loaded beam then presses the workpiece back against the fence, suddenly restoring contact with the cutterhead and producing kickback.

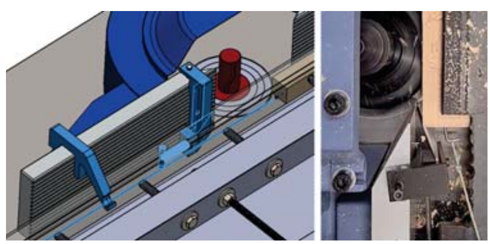

Fig. 3.: Left: Schematic setup of the lift-off unit. Right: Lift-off process during milling.

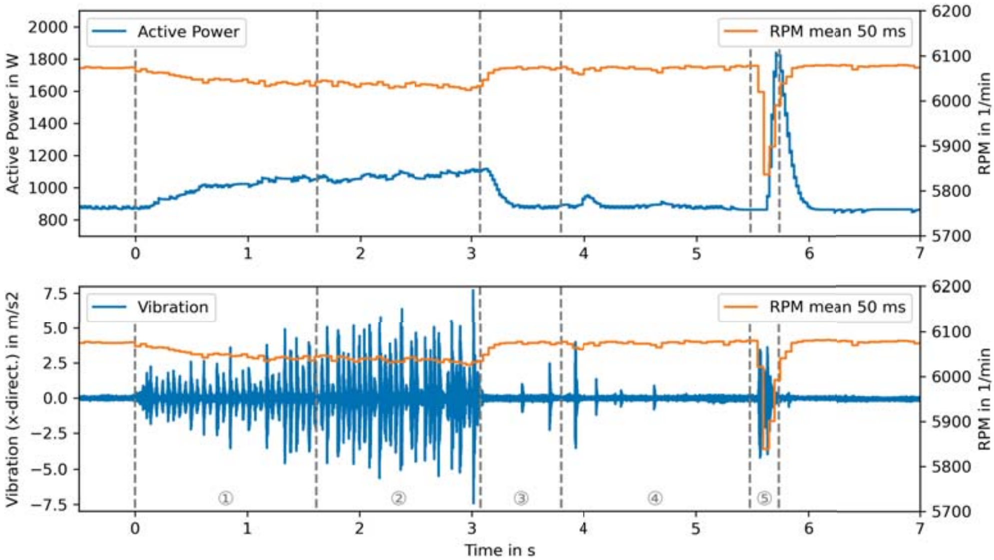

Fig. 4.: Typical measurement data from a single milling pass.

Our experimental setup reflects a typical real-world scenario for kickback. Nevertheless, other combinations of uneven feed motion and contact pressure can also lead to kickback.

Several signals were acquired with a Delphin data acquisition system to characterize the milling process: spindle speed was derived from an inductive sensor that produces three pulses per shaft revolution, sampled at 50 kHz. Table vibration was recorded with a industrial triaxial accelerometer, also sampled at 50 kHz. The total active machine power was measured using a Beckhoff EL3403, which provides a new power value every 20 ms (50 Hz). The active power value was forwarded to the Delphin system for time-synchronized logging. This configuration enables the correlation of spindle dynamics, structural vibration, and electrical load with the onset and evolution of kickback events. As a result, each dataset contains the following synchronized sensor signals: active power of the molder machine in watts, motor rotational speed in rotations per minute (RPM), and table vibration in x, y, z in m/s^2.

A typical dataset is shown in Fig. 4. The graph also shows the relevant process phases: (1) initial milling with reduced contact between the workpiece and cutterhead, (2) stationary milling, (3) the workpiece contacts the lift-off wedge, which induces a twist in the motor's holding fixture and slows down the workpiece's motion in this phase, (4) lift-off, and (5) kickback. While every dataset contains a kickback event, the duration of these events in each dataset is less than 5% of the total measurement time, resulting in a highly imbalanced data set.

4. Data Preparation

Table 1 summarizes the parameter values used in the experiments. All possible combinations of the three parameters were tested. For each of the 18 parameter combinations, eight experimental runs were performed, yielding a total of 144 datasets.

Table 1.: Parameters and parameter values

Milling speed (RPM)	Milling height (mm)	Milling depth (mm)
6080	5	5
3090	10	10
	15	15

Categorization enables evaluation of geometric effects, speed effects and overall performance across the complete dataset. For the experiments (Tables 2, 3), the datasets were categorized as follows:
(1) Small Milling: Height and depth = 5 mm, evaluated at both speeds.
(2) Large Milling: Height and depth = 15 mm, evaluated at both speeds.
(3) Fast Milling: All height–depth combinations at 6080 RPM only.
(4) Slow Milling: All height–depth combinations at 3090 RPM only.
(5) Total Milling: All speeds and all height–depth combinations combined.

A custom graphical user interface was used to manually annotate kickback events by visually inspecting the motor power signal and marking event start and end times for each recording. The resulting annotations were exported and used in a pre-processing pipeline that reloads the corresponding raw sensor data and maps annotated event times to sample indices. Based on these indices, a three-state temporal label sequence was generated, consisting of normal operation, a warning state extended backward by a fixed temporal buffer, and a fixed-duration event state. All sensor channels were standardized using a global z-score normalization to ensure consistent scaling across recordings. The normalized signals were then padded at the beginning and segmented into overlapping sliding windows, each assigned a binary label depending on whether a warning state occurred within a predefined future prediction window, yielding samples for predictive model training. As not all files could be annotated, the number of files used for model training differs from the total number of available files.

5. Model Construction and Metrics

A hybrid 1D CNN–attention–BiLSTM architecture was used for multivariate time-series event prediction (Fig. 5). Three 1D convolutional layers (64/128/256 filters; kernel sizes 7/5/3), each with batch normalization and ReLU, extracted local temporal features, with max pooling after the first two layers. A temporal attention module (1×1 convolution + sigmoid) reweighted salient time steps before two stacked BiLSTM layers (128 units per direction, dropout 0.4) captured long-range dependencies. The final representation was passed through a 128-unit fully connected layer with ReLU and dropout, followed by a linear output layer producing a single logit (sigmoid at inference).

The model was trained for binary prediction of events within a future window using sliding windows over normalized signals. Class imbalance was handled with weighted binary cross-entropy (positive weight equal to the negative-to-positive ratio). Optimization used AdamW (learning rate and weight decay 1×10^{-4}), gradient clipping (1.0), a validation-loss scheduler (factor 0.5), and early stopping based on validation AUC, with the best AUC checkpoint retained.

During inference, window-level probabilities were optionally smoothed using a moving average, and warnings were triggered when probabilities exceeded a predefined threshold for a sustained period.

The hybrid model was evaluated using metrics designed to capture both classification accuracy and early warning capability. Given the imbalanced nature of rare event data and the safety critical context, threshold dependent and independent metrics were jointly considered.

The optimal threshold was determined by maximizing the F1-score across thresholds ranging from 0.1 to 0.9 on test predictions, selecting the probability cutoff that achieves the best precision–recall balance. In highly imbalanced scenarios, the Precision–Recall (PR) curve is more informative than the ROC curve, as the latter can present an overly optimistic view when the negative class dominates. The area under the PR curve (PR-AUC) was therefore used as a primary evaluation metric, emphasizing correct detection of rare events while penalizing false alarms.

Inference speed is defined as the mean latency per sliding window, averaged across all windows and excluding data loading and processing, providing an indicator of suitability for near real-time applications. Lead time is defined as the temporal difference between the onset of a predicted warn-

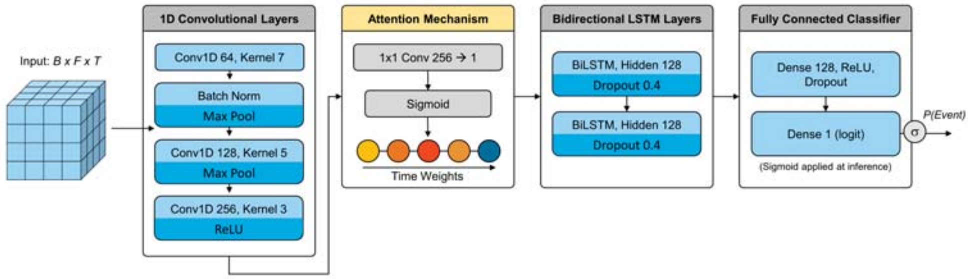

Fig. 5.: Architecture of the proposed CNN–BiLSTM–Attention model for time-series event prediction.

ing and the start of the corresponding labelled warning period, quantifying the advance notice available for initiating protective action.

6. Results and Discussion

We assessed the cross-configuration performance of five independently trained models across all milling configurations (total set, fast and slow speed, small and large milling depth). Table 2 presents results from a 15% holdout test split, whereas Table 3 reports performance on the full dataset. In most cases, models achieved their highest performance when evaluated on the same configuration used for training, suggesting limited transferability of configuration-specific features.

The model trained on slow milling speed obtained the highest F1-score (0.73), indicating a favorable balance between precision and recall under this operating regime. In contrast, the model trained on all configurations (Total milling) demonstrated the strongest overall generalization capability, yielding stable performance across different evaluation settings.

Across all models and configurations, events were detected with a median lead time of approximately 6.2 seconds prior to occurrence. In several cross-configuration evaluations, models operate at highly conservative decision thresholds, achieving near-perfect specificity while failing to capture a substantial fraction of positive events. While the slow-milling model achieves the highest peak F1-score, the model trained on all configurations provides more stable performance across heterogeneous operating conditions, highlighting a trade-off between specialization and robustness. These results suggest that in heterogeneous environments, either configuration-specific models or a jointly trained model with adaptive thresholding may be preferable to single-configuration training.

Occlusion importance quantifies how strongly the model relies on each input channel by measuring the drop in performance (here, PR-AUC) when that channel is occluded (set to zero) while all others remain unchanged. A larger PR-AUC decrease indicates that the removed signal contains information the model heavily depends on for warning detection. Fig. 6 shows that rotational speed and vibration in the X/Z axes dominate the model's predictive capability in this setting.

Inference latency on our GPU cluster is under 3 ms, which is acceptable given that the prediction lead time is significantly longer. This makes the system suitable, as a safety state can be activated in time.

In its current state, the proposed prediction model does not demonstrate sufficient accuracy

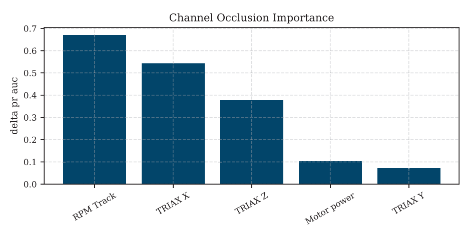

Fig. 6.: Occlusion importance for the total model

Table 2.: Test-split performance comparison with best values per metric highlighted in bold. Inference latency is reported as mean $\pm$ standard deviation in milliseconds. Confusion matrix entries are shown as globally normalized percentages.

Experiment	PR AUC	F1	Threshold	Latency (ms)	TN (%)	FP (%)	FN (%)	TP (%)
Total milling	0.6095	0.6109	0.7367	1.744 ± 0.367	79.55	9.27	2.18	8.99
Small milling	**0.8059**	0.6750	0.9000	1.770 ± 0.332	80.19	8.66	**1.05**	**10.10**
Slow milling	0.7599	**0.7210**	0.5735	2.505 ± 1.100	**82.52**	**5.13**	2.50	9.86
Large milling	0.5911	0.6012	0.1327	1.732 ± 0.291	81.97	7.44	2.84	7.75
Fast milling	0.6547	0.6239	0.8837	1.816 ± 0.319	82.25	7.70	2.01	8.05

Table 3.: Cross-configuration performance grouped by training regime. Best values per metric within each training group are highlighted in bold (higher is better; FPR lower is better). Median lead time is reported in seconds.

Trained on	Evaluated on	n_{files}	Prec.	Rec.	Spec.	FPR	PR-AUC	Best F1	Lead$_{50}$	TP (%)	TN (%)	FP (%)	FN (%)
	Large	12	0.112	0.004	**0.995**	**0.005**	0.327	0.008	7.37	0.05	86.98	0.40	12.57
	Small	15	0.485	0.325	0.948	0.052	0.501	0.389	6.20	4.24	82.45	4.50	8.81
Fast	Fast	59	**0.529**	**0.717**	0.914	0.086	**0.632**	**0.609**		8.46	80.66	7.54	3.34
	Slow	67	0.449	0.546	0.886	0.114	0.432	0.493	6.79	7.92	75.80	9.71	6.57
	Total	126	0.377	0.716	0.821	0.179	0.341	0.494	6.62	9.37	71.39	15.52	3.72
	Large	12	**0.545**	0.674	**0.919**	**0.081**	**0.598**	**0.602**	7.59	8.50	80.28	7.10	4.12
	Small	15	0.264	0.601	0.748	0.252	0.246	0.367	8.24	7.84	65.07	21.88	5.21
Large	Fast	59	0.269	0.893	0.676	0.324	0.213	0.414	7.66	10.54	59.62	28.58	1.26
	Slow	67	0.310	**0.961**	0.638	0.362	0.232	0.469	7.29	13.93	54.56	30.94	0.56
	Total	126	0.290	0.538	0.802	0.198	0.242	0.377	7.23	7.04	69.71	17.20	6.05
	Large	12	0.116	0.000	**1.000**	**0.000**	0.207	0.001	6.31	0.00	87.35	0.03	12.62
	Small	15	0.330	0.139	0.958	0.042	0.327	0.195	5.11	1.81	83.29	3.66	11.24
Slow	Fast	59	0.385	0.714	0.847	0.153	0.501	0.500	3.82	8.43	74.74	13.46	3.37
	Slow	67	**0.710**	**0.750**	0.948	0.052	**0.775**	**0.729**	3.68	10.87	81.07	4.44	3.63
	Total	126	0.464	0.509	0.911	0.089	0.532	0.485	5.04	6.66	79.20	7.71	6.43
	Large	12	0.464	0.518	**0.914**	**0.086**	0.493	0.490	7.45	6.54	79.84	7.54	6.08
	Small	15	**0.573**	**0.819**	0.909	0.091	**0.768**	**0.674**	4.03	10.68	79.00	7.95	2.37
Small	Fast	59	0.001	0.001	0.783	0.217	0.110	0.001	14.35	0.02	69.07	19.13	11.78
	Slow	67	0.396	0.714	0.816	0.184	0.358	0.510	6.91	10.34	69.74	15.77	4.15
	Total	126	0.403	0.566	0.874	0.126	0.371	0.471	6.91	7.41	75.95	10.96	5.69
	Large	12	0.414	0.231	**0.953**	**0.047**	0.394	0.296	6.24	2.91	83.26	4.12	9.71
	Small	15	0.400	0.751	0.831	0.169	0.570	0.522	7.20	9.80	72.25	14.70	3.25
Total	Fast	59	0.456	0.543	0.913	0.087	0.505	0.496	3.46	6.41	80.55	7.65	5.39
	Slow	67	**0.543**	**0.817**	0.883	0.117	**0.660**	**0.652**	3.02	11.84	75.54	9.97	2.65
	Total	126	0.522	0.767	0.894	0.106	0.621	0.621	5.66	10.04	77.70	9.21	3.05

and robustness to be deployed as a safety-critical component in milling machines. However, it shows potential as an assistance system. With further methodological improvements, future iterations may achieve the reliability required for safety-related applications. To certify a machine-learning-based safety component of machinery in the European Union, compliance must extend beyond traditional machinery safety to encompass emerging AI-specific regulation and harmonized standards. We discussed some of these aspects in [13]. The forthcoming AI Act's risk-based classification and extensive conformity assessment procedures are intended to safeguard health and safety but currently lack detailed technical guidance, which creates challenges for developers, particularly SMEs with limited resources. Several standardization activities are currently underway to address these challenges. Notably, ISO/IEC AWI TS 22440 (Functional safety and AI systems) is being developed to provide requirements, guidance, and examples of application for integrating

AI into safety-related functions.

7. Outlook

Given the physical danger and cost of generating real kickback events, synthetic data generation represents a viable pathway for performance enhancement. Generative adversarial networks (GANs) could be trained to generate realistic artificial kickback signatures. This would allow for the artificial enrichment of the minority kickback labels. Moreover, validating the model on different materials is necessary to examine wether the learned features are universal to the milling process or specific to the properties of spruce. Further tests are currently being carried out with a different type of wood to extend the available training data. Furthermore, we will benchmark the current CNN-Attention-LSTM architecture against temporal convolutional networks (TCNs), which look at the whole history at once using dilated convolutions.

Acknowledgment

This study and the acquisition of the data set was financed by the German Federal Ministry of Labor and Social Affairs. The research was in part conducted by the junior research group "Artificial Intelligence in a Safe and Healthy Working Environment" at the Federal Institute for Occupational Safety and Health (BAuA). Upon completion, the full dataset will be made publicly available on Zenodo.

References

1. DGUV. *Arbeitsunfallgeschehen 2024*. Deutsche Gesetzliche Unfallversicherung e.V. (DGUV), Berlin., 2024.
2. G. Feihle. Tischlereien und Schreinereien: Arbeitsunfall-Analyse zeigt Präventionspotenziale. *BGHM Magazin der Berufsgenossenschaft Holz- und Metall*, (04/2023), 2023.
3. DGUV. *Tischfräsmaschinen – Bau und Ausrüstung zum sicheren Verwenden*. FBHM-129. Deutsche Gesetzliche Unfallversicherung e.V. (DGUV), Berlin., 2022.
4. BGHM. *Holzbearbeitungsmaschinen. TSM – Handhabung und sicheres Arbeiten.* BG-Nr. 96.18. Berufsgenossenschaft Holz- und Metall, Mainz., 2020.
5. Felder Group. Pcs® Safety system. URL: `https://www.felder-group.com/en-sa/pcs`. Retrieved 2026-01-19.
6. Altendorf Group. Arbeitsunfälle an Formatkreissägen – Prävention durch HAND GUARD ist wirtschaftliche und gesellschaftliche Pflicht. URL: `https://www.altendorf.com/`. Retrieved 2026-01-19.
7. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41:15, 2009.
8. Wei Zhang, Chuanhao Li, Gaoliang Peng, Yuanhang Chen, and Zhujun Zhang. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical Systems and Signal Processing*, 100:439–453, 2018.
9. Tobias Schlagenhauf, Jan Wolf, and Alexander Puchta. Convolutional-based encoder–decoder network for time series anomaly detection during the milling of 16mncr5. *Data*, 7(12), 2022.
10. Clayton Cooper, Jianjing Zhang, Robert X. Gao, Peng Wang, and Ihab Ragai. Anomaly detection in milling tools using acoustic signals and generative adversarial networks. *Procedia Manufacturing*, 48:372–378, 2020.
11. A. Agogino and K. Goebel. Milling data set. *BEST lab, UC Berkeley. NASA Ames prognostics data repository*, 2007.
12. J. Lee, H. Qiu, G. Yu, J. Lin, and Rexnord Technical Services. Bearing data set. *IMS,University of Cincinnatiy. NASA Ames prognostics data repository*, 2007.
13. Franziska Wolny, Silvia Vock, Rasmus Adler, and Taras Holayad. The need for systematic approaches in risk assessment of safety-critical ai-applications in machinery. *Proceedings of the 35th European Safety and Reliability Conference (ESREL2025)*, 41:15, 3648-3655.