Machine Learning for Cutting Tool Wear Detection: A Multi-Dataset Benchmark Study Toward Predictive Maintenance

Žiga Kolar
ziga.kolar@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Thibault Comte
thibault.comte@universite-paris-saclay.fr
Universite Paris-Saclay
Paris, France

Yanny Hassani
yanny.hassani@universite-paris-saclay.fr
Universite Paris-Saclay
Paris, France

Hugues Louvancour
hugues.louv@gmail.com
Universite Paris-Saclay
Paris, France

Jože Ravničan
joze.ravnican@unior.com
UNIOR Kovaška industrija d.d.
Zreče, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

This student paper investigates the use of machine learning techniques to automate the detection of tool wear in cutting machines, replacing manual monitoring with intelligent, data-driven solutions. Although the proposed ML methods are standard in predictive maintenance, our contribution lies in providing the systematic multi-dataset benchmark tailored for direct transfer to industrial environments. This establishes a reproducible baseline before deploying and validating on real UNIOR data. As part of the project, and in anticipation of collecting real-world accelerometer data from industrial machines, we conducted a series of benchmarking experiments using five publicly available datasets that include accelerometer and audio signals under various wear-related conditions. The datasets cover a variety of industrial contexts and labeling schemes, allowing us to assess different preprocessing strategies and classification models such as Random Forests, 1D Convolutional Neural Networks, and Long Short-Term Memory networks. Our best results—an F1-score of 0.9949—were achieved using an LSTM model on a vibration dataset simulating fault conditions. These findings highlight the strong potential of AI for predictive maintenance and lay the groundwork for transferring the developed pipelines to the system once real data become available. Future work will focus on real-time wear detection and model deployment within live production environments.

Keywords

accelerometer, neural networks, machine learning, cutting tool

1 Introduction

This student paper presents the work carried out by Thibault Comte, Hugues Louvancour, and Yanny Hassani on the UNIOR project, under the mentorship of Žiga Kolar, prof. dr. Matjaž Gams for Jozef Stefan Institute, and Joze Ravnican for Unior. The objective of the UNIOR project is to detect when a cutting machine becomes worn out by analyzing sensor signals, specifically accelerometer data along the x, y, and z axes. An accelerometer is mounted on the cutting machine to monitor vibrations occurring during the cutting process. Currently, the detection of wear is performed manually by a human operator. By leveraging artificial intelligence (AI) and machine learning (ML), this process can be automated, making it both easier and more efficient.

While awaiting the company to complete the necessary paperwork and acquire and install the accelerometer on the cutting machine, we identified similar publicly available datasets and conducted several machine learning experiments using them.

2 Related Work

This section briefly surveys recent research on the use of artificial intelligence (AI) techniques for tool wear monitoring in manufacturing processes such as milling, turning, and drilling. Munaro et al. [2] provide a systematic review of 77 studies, contrasting offline and online monitoring methods. Online approaches leveraging sensor data—such as force, vibration, acoustic emission, and power—are enhanced by AI models like SVMs, ANNs, CNNs, and LSTMs, offering accuracies above 90% and industrial relevance. Sieberg et al. [5] demonstrate CNN-based classification of wear mechanisms from SEM images, achieving 73% test accuracy. They emphasize dataset balance and magnification consistency as critical challenges. Colantonio et al. Shah et al. [4] argue for ML's superiority over physics-based models in wear prediction, underscoring ANN's predictive strength when supplied with high-quality data and standardized evaluation methods. Recent studies also explore multimodal sensor fusion, combining accelerometer, acoustic, and force signals to improve robustness [8]. Specifically, transfer learning has been shown effective for adapting models trained on laboratory data to industrial machines [8].

Unlike previous reviews such as Munaro et al. [2], which survey the field, our work provides a systematic multi-dataset experimental comparison across three different sensor modalities (accelerometer, vibration, audio) using standardized pipelines. This benchmarking is not only descriptive but forms the basis for industrial transfer to UNIOR's production line, bridging academic datasets with real machine applications.

3 Datasets

This section describes five different datasets that were identified—four containing accelerometer data and one featuring audio recordings.

Information Society 2025, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2025.skui.6224

3.1 Bosch CNC Machining Dataset

The Bosch CNC Machining dataset consists of real-world industrial vibration data collected from a brownfield CNC milling machine. Acceleration was measured using a tri-axial Bosch CISS sensor mounted inside the machine, recording the X, Y, and Z axes at a sampling rate of 2 kHz. Both normal and anomalous data were collected across six distinct timeframes, each spanning six months between October 2018 and August 2021, with appropriate labeling. Data were collected from three distinct CNC milling machines, each executing 15 processes [7]. A total of 1,702 samples were obtained, with each labeled as either "good" or "bad." The distribution of labels was 95.9% good and 4.1% bad.

3.2 Cutting Tool Wear Audio Dataset

This dataset comprises 1,488 ten-second .wav audio recordings of cutting tool wear collected at two spindle speeds: 520 RPM and 635 RPM. Each audio recording is labeled as either “BASE” (machine running without cutting), “FRESH” (sharp cutting tool), “MODERATE” (moderately worn tool), or “BROKEN” (broken or fully worn tool). The “FRESH,” “MODERATE,” and “BROKEN” labels were specifically chosen to simulate real cutting conditions, focusing on scenarios where the machine is actively engaged in material removal. In total, the dataset includes 400 “FRESH” samples, 376 “MODERATE” samples, and 362 “BROKEN” samples across both spindle speeds, offering a nearly balanced distribution well-suited for ML applications. Audio records had different lengths. No artificial background noise was added to the recordings. All cutting tools used were 16 mm end-mill cutters, and the workpiece material was mild steel [6].

3.3 Turning Dataset for Chatter

This dataset contains sensor signals collected from multiple cutting tests using a range of measurement devices, including two perpendicular single-axis accelerometers, a tri-axial accelerometer, a microphone, and a laser tachometer. Both raw sensor data and processed, labeled data from one channel of the tri-axial accelerometer are provided. There were four labels used: no-chatter, intermediate chatter, chatter, and unknown. The dataset contains a total of 117 signals, with the following label distribution: 51 labeled as no-chatter, 19 as intermediate chatter, 22 as chatter, and 25 as unknown. Data were collected under four distinct cutting configurations, defined by varying the stick out distance—the distance from the heel of the boring rod to the back face of the tool holder. The four stickout distances used were 5.08 cm (2 inches), 6.35 cm (2.5 inches), 8.89 cm (3.5 inches), and 11.43 cm (4.5 inches) [8].

3.4 UCI Accelerometer Dataset

To simulate motor vibrations, a 12 cm Akasa AK-FN059 Viper cooling fan was modified by attaching weights to its blades, and an MMA8452Q accelerometer was mounted to capture vibration data. An artificial neural network was then used to predict motor failure time based on this data. Three distinct vibration scenarios were generated by varying the placement of two weight pieces on the fan blades: (1) Red – normal configuration, with weights on neighboring blades; (2) Blue – perpendicular configuration, with weights on blades 90° apart; and (3) Green – opposite configuration, with weights on opposite blades. For each of the three weight configurations, vibration data was collected every 20 ms over a 1-minute interval per speed, resulting in 3,000 records per speed. In total, the dataset contains 153,000 vibration records from the simulation model [3].

3.5 Vibrations Dataset

This dataset contains vibrational data collected to support early fault diagnosis in machinery The data was gathered using an SG-Link tri-axial accelerometer sensor (by MICROSTRAIN Corporation) at a sampling rate of 679 samples per second for each of the three axes: axial (z), horizontal (x), and vertical (y). Experiments were conducted in the Mechanical Vibration Laboratory at the Mechanical Engineering Department of the University of Engineering and Technology (UET), Taxila. The setup simulated four distinct machine conditions: normal, cracking, offset pulley, and wear states, using a test rig designed for fault simulation [1].

4 Methodology and Results

This section outlines the methodology used for each dataset, focusing on multiclass classification. Various preprocessing techniques and machine learning algorithms were applied.

4.1 Bosch CNC Machining Dataset

The Bosch CNC Machining Dataset contains 95.9% good signals and 4.1% bad signals. The objective was to develop a binary classification model that outperforms a naive baseline, which achieves 95.9% accuracy simply by always predicting a signal as good.

Two approaches were tested on the Bosch CNC Machining dataset. The first approach applied random undersampling, which balances class distribution by randomly removing samples from the majority class while leaving the minority class unchanged. Since the majority class accounted for 95.9% of the data, this step was essential to prevent the model from defaulting to majority-class predictions. After applying the random undersampling, the Random forest model was used for binary classification. This method achieved 99% accuracy on 5-fold cross validation, providing a 3.1% improvement over the naive baseline model.

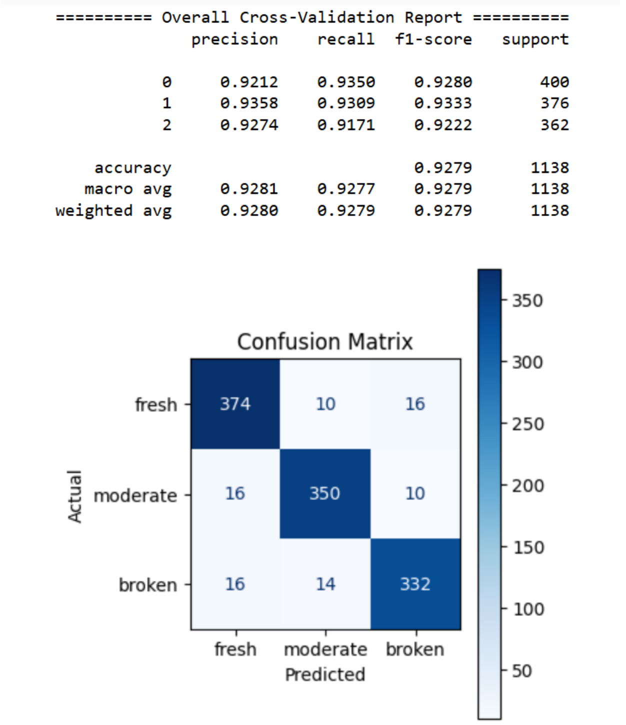
Different preprocessing strategies were necessary due to differences in data formats, sampling rates, and class balance across datasets. For example, in the Bosch dataset, random undersampling was applied only on the training folds during 5-fold CV to avoid information leakage.

In the second approach, features were initially extracted using two 1D Convolutional layers followed by two Max Pooling layers. To augment the data, random Gaussian noise was added to the signals, effectively doubling the size of the training set. A binary classification model using Random Forest was then trained on this augmented dataset. This model achieved a high accuracy of 0.996 under 5-fold cross-validation, outperforming the naive baseline by 3.7%.

McNemar’s test was applied between competing models on each dataset. Significant differences ($p < 0.05$) were observed between CNN and Random Forest on the Bosch dataset, confirming that improvements are not due to random variation.

4.2 Cutting Tool Wear Audio Dataset

The Cutting Tool Wear Audio Dataset contained 400 “FRESH”, 376 “MODERATE”, and 362 “BROKEN” samples across two spindle speeds, requiring a multi-class classification approach. Since the signals varied in length, we first identified the longest signal (48000 samples) and zero-padded shorter signals to match this length. To improve model accuracy, this maximum length was



```
========== Overall Cross-Validation Report ==========
              precision    recall  f1-score   support

           0     0.9212    0.9350    0.9280       400
           1     0.9358    0.9309    0.9333       376
           2     0.9274    0.9171    0.9222       362

    accuracy                         0.9279      1138
   macro avg     0.9281    0.9277    0.9279      1138
weighted avg     0.9280    0.9279    0.9279      1138
```

Figure 1: 5-Fold cross validation report and confusion matrix for Cutting Tool Wear Audio dataset.

later reduced. The model architecture included two 1D Convolutional layers and two 1D Max Pooling layers to reduce the dimensionality of the data while preserving essential features.

The output from the upper layers served as input to a feature selection algorithm, which identified the 96 most relevant features out of a total of 2048. These selected features were then used by a Random Forest classifier to predict the final label.

The best model for this dataset achieved 0.9279 (+/- 0.01) accuracy and 0.9279 F1 score on 5-fold cross validation. Results (precision, recall, F1-score and accuracy) are presented on Figure 1.

4.3 Turning Dataset for Chatter

Since each signal varies in length and can be quite long, an approach based on extracting time-domain and frequency-domain features was implemented. This method preserves essential information from the original signals while significantly reducing dimensionality, making the data more suitable for ML algorithms.

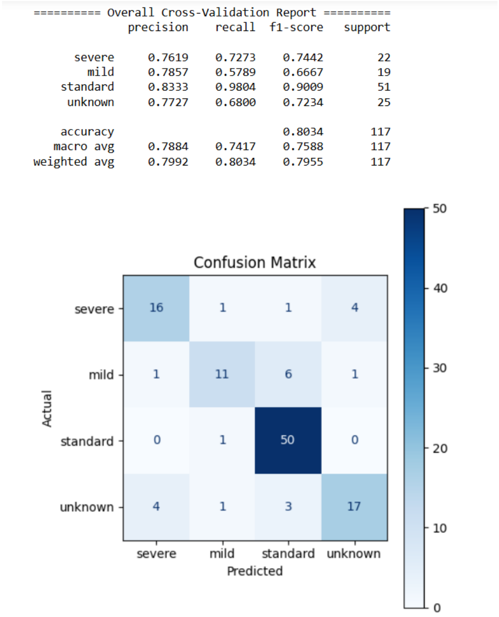
The following approach combines signal segmentation and frequency-domain feature extraction to summarize the spectral characteristics of a time-series signal. First, it divides the input signal into overlapping or non-overlapping fixed-size windows using a sliding window technique, where each segment is of 10000 windows length and the shift between consecutive segments is determined by step size, which in this case is 5000. This allows for localized analysis of signal dynamics over time.

Next, we applied the Fast Fourier Transform (FFT) to each segment, converting the time-domain signal into its frequency-domain representation. It computes the magnitude spectrum for each segment and then averages the spectral magnitudes across all segments to obtain a single, representative frequency-domain feature vector. This results in a compact yet informative summary that captures the dominant frequency components of the entire signal, while accounting for temporal variation through segmentation.

Furthermore, 11 additional features were extracted from the raw signal, including the mean, standard deviation, minimum, maximum, and median of the frequency values. These features capture the signal's central tendency and variability, providing a statistical summary of its frequency content. The 25th and 75th percentiles further quantify the signal's interquartile range, highlighting its variability and robustness to outliers. Root mean square (RMS) provides a measure of the signal's overall power. Skewness and kurtosis describe the asymmetry and peakedness of the distribution, respectively, offering insights into the signal's shape beyond basic statistics. Finally, zero crossings count the number of times the signal crosses the zero axis, serving as an indicator of frequency content and signal complexity. Together, these features form a rich representation for classification tasks involving time-frequency signals.

In total, there were 268 features (257 FFT features and 11 additional features) and 117 samples. A feature selection technique was applied to further reduce the number of features. 140 best features were selected and used as input for Random Forest classifier.

The best model for this dataset achieved 0.80 (+/-0.06) accuracy and 0.7588 F1-score on 5-fold cross validation. Results (precision, recall, F1-score and accuracy) are depicted on Figure 2.



```
========== Overall Cross-Validation Report ==========
              precision    recall  f1-score   support

      severe     0.7619    0.7273    0.7442        22
        mild     0.7857    0.5789    0.6667        19
    standard     0.8333    0.9804    0.9009        51
     unknown     0.7727    0.6800    0.7234        25

    accuracy                         0.8034       117
   macro avg     0.7884    0.7417    0.7588       117
weighted avg     0.7992    0.8034    0.7955       117
```

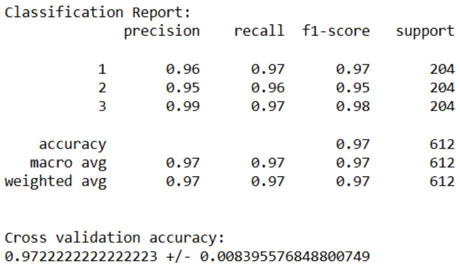
Figure 2: 5-Fold cross validation report and confusion matrix for Turning dataset for Chatter.

4.4 UCI Accelerometer Dataset

This method implements a complete machine learning pipeline for classifying time-series accelerometer data using features extracted from both the time and frequency domains. Data is first loaded from a CSV file, where each row contains an activity label and raw X, Y, and Z accelerometer readings. The signal is segmented into non-overlapping windows of fixed size (50 samples, corresponding to 1 second at 50 Hz), and only windows with consistent activity labels are retained for supervised learning.

Next, time-domain and frequency-domain features were extracted from a signal. Time-domain features include basic statistics (mean, standard deviation, min, max, median), RMS, peak-to-peak range, skewness, kurtosis, zero-crossing rate, signal energy, and crest factor. Frequency-domain features are extracted via FFT and include spectral centroid, spectral spread, peak frequency, and energy in predefined low (0–5 Hz) and high (10–25 Hz) frequency bands.

This feature-rich representation is passed through a machine learning pipeline that includes feature scaling, univariate feature selection (Select K Best ANOVA F-statistical method), and classification using a Random Forest classifier. The best model for this dataset achieved 0.972 (+/-0.008) accuracy and 0.97 F1-score on 5-fold cross validation. Results (precision, recall, F1-score and accuracy) are depicted on Figure 3.



```
Classification Report:
              precision    recall  f1-score   support

           1       0.96      0.97      0.97       204
           2       0.95      0.96      0.95       204
           3       0.99      0.97      0.98       204

    accuracy                           0.97       612
   macro avg       0.97      0.97      0.97       612
weighted avg       0.97      0.97      0.97       612


Cross validation accuracy:
0.9722222222222223 +/- 0.008395576848800749
```

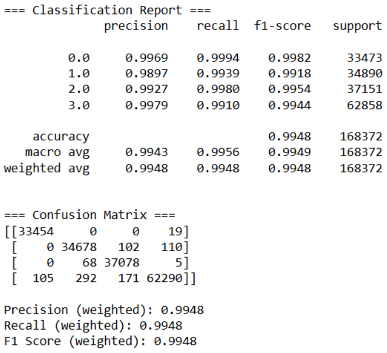
Figure 3: 5-Fold cross validation report and confusion matrix for UCI Accelerometer dataset.

4.5 Vibrations Dataset

In this method the time series data was effectively segmented into overlapping windows of fixed length 226. A total of 168,372 samples were generated, providing a sufficient amount of data for training deep learning models. A Long Short-Term Memory (LSTM) neural network was chosen due to its effectiveness in handling sequential data. The network architecture consisted of two LSTM layers with 128 and 64 units, respectively, along with two Dropout layers incorporated to reduce the risk of overfitting and improve generalization. This method achieved the best performance to date, reaching an accuracy of 0.9948 (+/-0.005) in 5-fold cross-validation and an F1-score of 0.9949. The results are presented on Figure 4.

5 Conclusion

This student paper explored machine learning for automated cutting tool wear detection. Using five public datasets and models such as Random Forests, CNNs, and LSTMs, we achieved strong performance, notably 0.9949 F1 on the Vibrations dataset. These benchmarks highlight ML's potential for predictive maintenance and provide ready-to-deploy pipelines for future industrial data. Future work will focus on validating the model on industrial machines, optimizing its performance, and deploying it in real-time. Additionally, for ordered domains like the Cutting Tool Wear Audio dataset, misclassifications should not be penalized equally (e.g., "FRESH" -> "MODERATE" vs. "FRESH" -> "BROKEN"). Thus, future research will explore ordinal metrics, such as weighted accuracy or quadratic weighted kappa.



```
=== Classification Report ===
              precision    recall  f1-score   support

         0.0     0.9969    0.9994    0.9982     33473
         1.0     0.9897    0.9939    0.9918     34890
         2.0     0.9927    0.9980    0.9954     37151
         3.0     0.9979    0.9910    0.9944     62858

    accuracy                         0.9948    168372
   macro avg     0.9943    0.9956    0.9949    168372
weighted avg     0.9948    0.9948    0.9948    168372


=== Confusion Matrix ===
[[33454     0     0    19]
 [    0 34678   102   110]
 [    0    68 37078     5]
 [  105   292   171 62290]]

Precision (weighted): 0.9948
Recall (weighted): 0.9948
F1 Score (weighted): 0.9948
```

Figure 4: 5-Fold cross validation report and confusion matrix for Vibration dataset.

This study has several limitations. First, the datasets used are publicly available and may not fully capture the variability of industrial machining environments. Second, in some cases class balance was artificially enforced via undersampling, which could affect generalizability. Third, we recognize that the lack of direct industrial validation is a current limitation. However, our pipelines were designed for immediate deployment once the company's accelerometers are installed, ensuring direct continuity from these benchmark studies to industrial application. This study therefore serves as a reproducible foundation rather than a final industrial deployment. Partial validation experiments with UNIOR's machines are planned as the next project stage.

Acknowledgements

The authors acknowledge funding support from the company UNIOR for the GREMO LIGHTWEIGHT project. The authors also acknowledge the funding from the Slovenian Research and Innovation Agency (ARIS), Grant (PR-10495) and Basic core funding P2-0209.

References

[1] Muhammad Umar Khan, Muhammad Atif Imtiaz, Sumair Aziz, Zeeshan Kareem, Athar Waseem, and Muhammad Ammar Akram. 2019. System design for early fault diagnosis of machines using vibration features. In *2019 International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET)*. IEEE, 1–6.

[2] Roberto Munaro, Aldo Attanasio, and Antonio Del Prete. 2023. Tool wear monitoring with artificial intelligence methods: a review. *Journal of Manufacturing and Materials Processing*, 7, 4, 129. DOI: 10.3390/jmmp7040129.

[3] Gustavo Scalabrini Sampaio, Arnaldo Rabello de Aguiar Vallim Filho, Leilton Santos da Silva, and Leandro Augusto da Silva. 2019. Prediction of motor failure time using an artificial neural network. *Sensors*, 19, 19, 4342.

[4] Raj Shah, Nikhil Pai, Gavin Thomas, Swarn Jha, Vikram Mittal, Khosro Shirvni, and Hong Liang. 2024. Machine learning in wear prediction. *Journal of Tribology*, 147, 4, (Nov. 2024), 040801. eprint: https://asmedigitalcollection.asme.org/tribology/article-pdf/147/4/040801/7400649/trib_147_4_040801.pdf. DOI: 10.1115/1.4066865.

[5] Philipp Maximilian Sieberg, Dzhem Kurtulan, and Stefanie Hanke. 2022. Wear mechanism classification using artificial intelligence. *Materials*, 15, 7, 2358. DOI: 10.3390/ma15072358.

[6] Nachiket Soni, Amit Kumar, and Hardik Patel. 2023. Acoustic analysis of cutting tool vibrations of machines for anomaly detection and predictive maintenance. In *2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC)*. IEEE, 43–46.

[7] Mohamed-Ali Tnani, Michael Feil, and Klaus Diepold. 2022. Smart data collection system for brownfield cnc milling machines: a new benchmark dataset for data-driven machine monitoring. *Procedia CIRP*, 107, 131–136.

[8] Melih C Yesilli, Firas A Khasawneh, and Andreas Otto. 2020. On transfer learning for chatter detection in turning using wavelet packet transform and ensemble empirical mode decomposition. *CIRP Journal of Manufacturing Science and Technology*, 28, 118–135.