Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 107 (2022) 131–136

www.elsevier.com/locate/procedia

55th CIRP Conference on Manufacturing Systems

Smart Data Collection System for Brownfield CNC Milling Machines: A New Benchmark Dataset for Data-Driven Machine Monitoring

Mohamed-Ali Tnani*[a,b], Michael Feil[c], Klaus Diepold[b]

[a]Department of Factory of the Future, Bosch Rexroth AG, Lise-Meitner-Str. 4, 89081 Ulm, Germany
[b]Department of Electrical and Computer Engineering, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany
[c]Department of Informatics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany

* Corresponding author. Tel.: +49 899 2139-651. *E-mail address:* mohamed-ali.tnani@boschrexroth.de

Abstract

Manufacturing processes have undergone tremendous technological progress in recent decades. To meet the agile philosophy in industry, data-driven algorithms need to handle growing complexity, particularly in Computer Numerical Control machining. To enhance the scalability of machine learning in real-world applications, this paper presents a benchmark dataset for process monitoring of brownfield milling machines based on acceleration data. The data is collected from a real-world production plant using a smart data collection system over a two-years period. In this work, the edge-to-cloud setup is presented followed by an extensive description of the different normal and abnormal processes. An analysis of the dataset highlights the challenges of machine learning in industry caused by the environmental and industrial factors. The new dataset is published with this paper and available at: `https://github.com/boschresearch/CNC_Machining`.

Keywords: Smart manufacturing; Internet of Things; Machine learning; Process monitoring; Data mining; Industry 4.0; Industrial dataset

1. Introduction

Manufacturing equipment is characterized by its long-lasting capability of up to 20 or more years [1] and communicating with brownfield components and controllers can be a barrier to scalable real-time applications. Development towards Industry 4.0, and more specifically Cyber-Physical Systems and Internet of Things (IoT), paved the way for digitalization and retrofitting of old machinery with connected sensors, edge intelligence, cloud connection, etc. [2, 3]. One of the most robust and long-standing pillars of the production chain are Computer Numerical Control (CNC) machines.

Highly automated machining centers are characterized by their high-speed manufacturing but also by their complexity. The extreme environmental conditions and the high-speed processing engender operation failures such as tool breakage, improper tool clamping or chip jamming [4]. The high variety in tool types and tool operations (OP), in terms of shape, geometries, materials, coatings, surface finishing and physical changes over time, rises strong robustness and generalization challenges for traditional analytics [4]. The complexity increases with the discrepancy between the same processes caused by changes in the machining parameters and maintenance methods, such as lubrication of components.

Addressing these challenges, a wide variety of research in tool health monitoring [5, 6] and few in process quality [7, 8] has been performed. To enhance the research in the field, some machining datasets have been published. One of them is the SMART LAB Milling Dataset [9], which has been collected at the University of Michigan over 18 different experiments from direct measurements. The goal of the dataset is to investigate the tool wear detection as well as detection of inadequate clamping. A second dataset is from the NASA Milling Dataset [10], which studies the tool wear based on three different types of sensors, acoustic emission, vibration and current. However, both experiments were conducted in a laboratory during a short limited time frame. To the best of the authors knowledge, there exists no CNC research dataset from a real production environment collected over a long period of time and from different machines. These conditions are essential to build robust data-driven models and improve their generalization and thus their reliability in industry.

2212-8271 © 2022 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the International Programme committee of the 55th CIRP Conference on Manufacturing Systems
10.1016/j.procir.2022.04.022

This paper introduces a new dataset collected during real-life production. The data is collected from three brownfield milling CNC machines at different time frames in a two years interval. The first section of this paper describes the IoT system built to retrofit the old machinery, ease the data collection and enable parallel prediction and annotation. The second section provides in-depth description and analysis of the dataset that will be published with this paper. It is followed by an overview of the environmental and industrial challenges, which have been considered in a systematic way during dataset creation and annotation. This allows the scientific community to work on solutions for these real-world problems and provide comparable results for benchmarking. More specifically, the dataset has been designed to address the challenges of feature drifts between machines and over time, the high diversity of tool operations during production and the severe dataset imbalance in terms of number of samples per class. To overcome these challenges, we propose some data split scenarios which can be used in future work.

2. Experimental Set Up/ Data Acquisition System

2.1. Hardware components

To keep the research as close as possible to the industrial scenario, the data is collected from different 4-axis horizontal CNC machining centers during production. The machines are processing aluminum workpieces as depicted in Figure 1. For the data acquisition, we used an indirect method by collecting accelerometer data from Bosch CISS sensors [11] mounted to the rear end of the spindle housing. Other approaches opt for mounting the sensors in the machining area [12, 13, 5, 7]. This rear area remains unaffected by extreme machining environment, coolant or material chips and is available for retrofitting new sensors to brownfield machines. The sensor maintains a constant distance to the tool center point and the three axes of the accelerometer are in alignment with the linear motion axis of the machine. The sensor coordinate system is indicated in Figure 1.

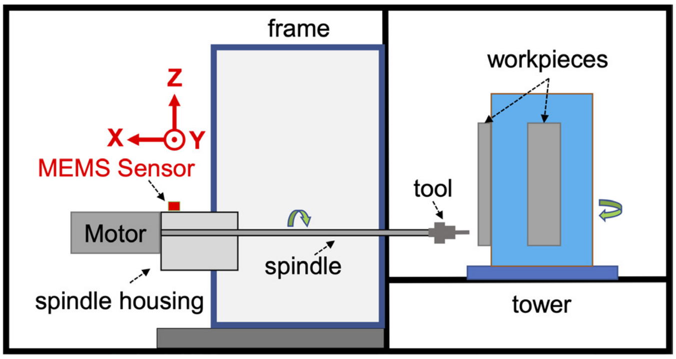

Fig. 1: Schematic sketch of the experimental setup: 4-axis machining center with mounted sensor.

Using the low-cost tri-axial CISS sensor, acceleration data is collected with a sampling rate of 2 kHz. As mentioned in Section 3.2.2, most relevant frequencies to monitor the machining processes are low integer multiples (1..4) of the spindle speed. For tool operations present in this dataset (see Table 1), these frequencies will be in the range of 75 Hz to 1 kHz. According to the Nyquist-Shannon theorem [14], a minimum sampling rate of 2 kHz is sufficient to detect machine anomalies. Sampling with this rate along the 3-axes produces an amount of 4.14 GB per day. Such volumes of data cannot be fully stored and processed in on-premise solutions. It demands a smart data mining system to collect, store, annotate, process and learn from the gathered data.

2.2. Software Architecture for Data Collection

To have reliable annotation, continuous data collection and simultaneous Machine Learning (ML) evaluation, we require an IoT architecture which enables:

1. central aggregation of selected anomalies and processes across different machining centers and locations,
2. local storage and processing of raw sensor data including event annotation by product experts,
3. aggregation of annotated data in a central database,
4. centralized training of ML models, and
5. management and deployment of models and modules from the cloud to the edge device.

Sun et al. [15] proposes the offloading of the ML inferencing to on-premise servers to improve the communication effort and latency. In similiar fashion, Yigitoglu et al. [16] proposed a framework for Fog computing. Motivated by both works [16, 15], the data collection system presented in this work is characterized in an edge-to-cloud architecture. The main goal of this architecture is the simplification of data annotation, the use of expert knowledge in the shop floor, and the centralized storage of annotated data in the cloud. Through an anomaly detector module, potential events and anomalies are pre-selected for annotation. In this section, we outline the edge-to-cloud data collection system.

2.2.1. Edge stack

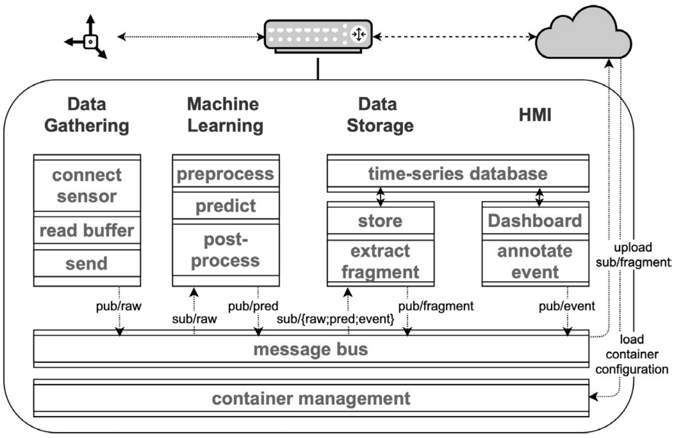

Fig. 2: Concept and interaction of containers in the edge stack.

This is a resupply of March 2023 as the template used in the publication of the original article contained errors. The content of the article has remained unaffected.

The edge stack represented in Figure 2 describes the modules running in the production line on site. The modules are managed from the cloud side by an orchestration client running on the edge device. A messaging bus using the Message Queuing Telemetry Transport (MQTT) protocol provides a standardized interface for local inter-application communication. The data gathering and annotation system involves multiple modules. Firstly, a data gathering module establishes a connection to the accelerometer sensor and triggers the read. The data stream is afterwards published on the message bus. Secondly, the data stream is subscribed by a ML module, which with predictions on the stream, supports the quality check process by pre-selecting the correct time frame for anomalies. This allows time-delayed annotations to be entered by the end-of-line quality check, while retaining the majority of data only in the edge time-series database. Ultimately, a dashboard allows the visualization of the ML pseudo-labels and manual annotation via the user interface. Once an event is validated by the experts, the corresponding data segment gets acquired and queried for upload to the cloud. The major benefit of the architecture is the collaboration of data science and domain expertise. It allows additionally in-place distribution of updated ML modules, which support and improve data annotation.

2.2.2. *Cloud Stack*

Publishing large-scale dataset, training ML models centrally or aggregating data from multiple edge devices require a cloud stack. Figure 3 presents the data flow and the main components required in the cloud. Annotated vibration fragments from edge devices get streamlined to the central data storage. Using the vibration fragments from the central data storage, the data is segmented and preprocessed. Since annotating high frequency data can be very challenging for human experts, verification of correctness of the labels is essential. After verification, the ML model is (re-)trained, gets build and registered in the Container Registry. The edge device management communicates the new model version to the local server. Using this paradigm, we successfully improve our models at the edge through the continuous collection of anomalies.

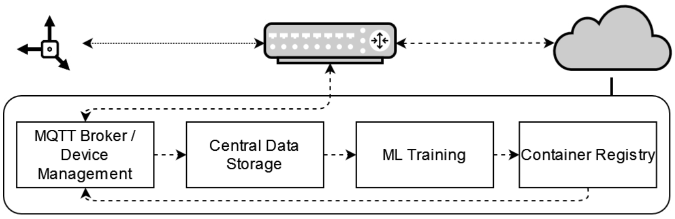

Fig. 3: Services and building blocks in cloud stack.

3. New CNC Machining Dataset

The overall goal of this paper is to enhance the scalability of machine learning in real-world applications by presenting a dataset containing the main challenges that hinder the reliability of ML algorithms in the manufacturing environment. The challenges are caused on the one hand by the variation of material components (spindle, machining tools, raw material produced, etc.) due to wear or discrepancies in the physical structure of parts across machines, and on the other hand by the frequent changes in the production flow as a result of customer requirements and technological progress.

The following section presents the dataset and the various process operations. We introduce how we systematically embed the real-world challenges into the collected data.

3.1. *Data Description*

The data is collected in a production plant from 3 different CNC machines (M01, M02 and M03) on a regular basis during the time interval of October 2018 to August 2021. The time frame is tagged as ”Month_Year” and represents the 6-month interval before the label. For example, ”Aug_2019” would refer to the period between February 2019 and August 2019.

The machine performs a sequence of several operations using different tools on aluminium parts to work the specified design. It is important to mention that the machines produce different parts and the process flow changes over time. To study the drift between machines and over time, the dataset is built with 15 different tool operations that run on all 3 machines at different time frames. Table 1 gives an overview on the characteristics of the different operations.

Table 1 Tools operations collected from M01, M02 and M03.

Tool operation	**Description**	**speed** [Hz]	**feed** [$mm\ s^{-1}$]	**duration** [s]
OP00	Step Drill	250	≈ 100	≈ 132
OP01	Step Drill	250	≈ 100	≈ 29
OP02	Drill	200	≈ 50	≈ 42
OP03	Step Drill	250	≈ 330	≈ 77
OP04	Step Drill	250	≈ 100	≈ 64
OP05	Step Drill	200	≈ 50	≈ 18
OP06	Step Drill	250	≈ 50	≈ 91
OP07	Step Drill	200	≈ 50	≈ 24
OP08	Step Drill	250	≈ 50	≈ 37
OP09	Straight Flute	250	≈ 50	≈ 102
OP10	Step Drill	250	≈ 50	≈ 45
OP11	Step Drill	250	≈ 50	≈ 59
OP12	Step Drill	250	≈ 50	≈ 46
OP13	T-Slot Cutter	75	≈ 25	≈ 32
OP14	Step Drill	250	≈ 100	≈ 34

For sake of confidentiality the tool operations order has been shuffled and only a part of the production flow is present in the dataset. Each operation in the table represents a specific process performed by a different tool with unique parameters.

As described in the experimental set-up, the data has been collected from the accelerometer with no further information from the machine's controller. Figure 4 gives an overview on

This is a resupply of March 2023 as the template used in the publication of the original article contained errors. The content of the article has remained unaffected.

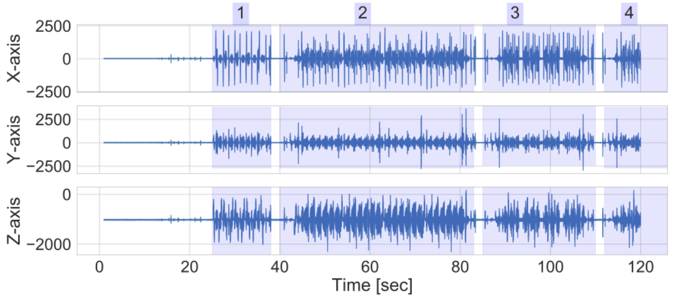

Fig. 4: Overview of the segmentation step of the different tool operations. The X, Y, Z acceleration axes of 4 sequential tool operations are illustrated. The tool number is mentioned in the upper border.

the collected acceleration data from a machining sequence. The data is then manually segmented and structured in the research database. Having no connection to the controller hinders the automated segmentation and thus the process-wise anomaly detection. A non-intrusive solution to monitor and prevent process failures consist of windowing the data stream with a fixed-sized window length and processing the windows steam independently from the process ID.

3.2. *Real-world Challenges*

Generalization is still one of the primary challenges for industrial ML due the continuous disturbances. Driven by market demand and technical progress, CNC machining production processes are constantly changing with R&D advancement, which goes along with modifications in the tool process operations. Another type of disturbance is caused by the noisy environment in the shop floor and the high imbalance of the normal/abnormal classes. This section presents the different industrial challenges based on the Bosch CNC Machining Dataset described in the previous section.

3.2.1. *Environmental challenges*

During machining, the different process operations are conducted in high-speed, requiring a frequent mounting and unmounting of tools on the spindle chuck. These factors lead occasionally to process failures mainly caused by tool misalignment, chip clamping, chip in chuck, tool breakage, etc. To reach the optimal product quality, after each batch an expert on the shop floor controls the resulting workpiece in a gauging station and annotate the process health. Nevertheless, labeling during production is still very challenging. Due to the manual drudgery gauging, some processes are wrongly labelled and precise annotations are missing. The published dataset focuses on the quality process failures, i.e., the OK class refers to a healthy process and NOK refers to a faulty process.

A common challenge in industrial datasets is the strong OK/NOK unbalance, especially in process monitoring tasks. Figure 5 shows an unbalance rate of 816:35 between the OK/NOK in our dataset. In our real production, the amount of OK samples are significantly higher. To provide an exemplary dataset, a reasonable number of OK processes were selected from the different time periods, which reduces the class imbalance.

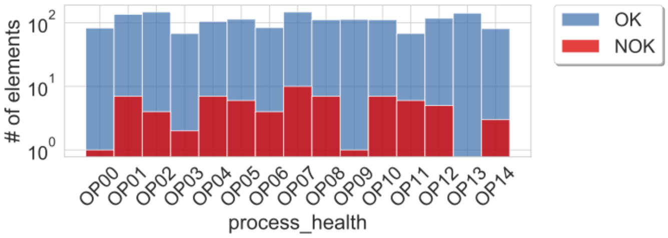

Fig. 5: Class distribution per process operation.

Besides the process failures, some condition anomalies occur and are detected only after machine maintenance. These anomalies are caused mainly by components wear, hydraulic issues, incorrect settings, etc. However, before reaching a critical phase, a slight deterioration/change over time is seen, causing additional noise in the vibration data. This causes a drift in the OK class between different time frames. In addition to ageing drift, a discrepancy between the conditions of the machines and machine components increases the challenge in real-world applications. The within-class discrepancy over time and between machines is studied in Figure 6.

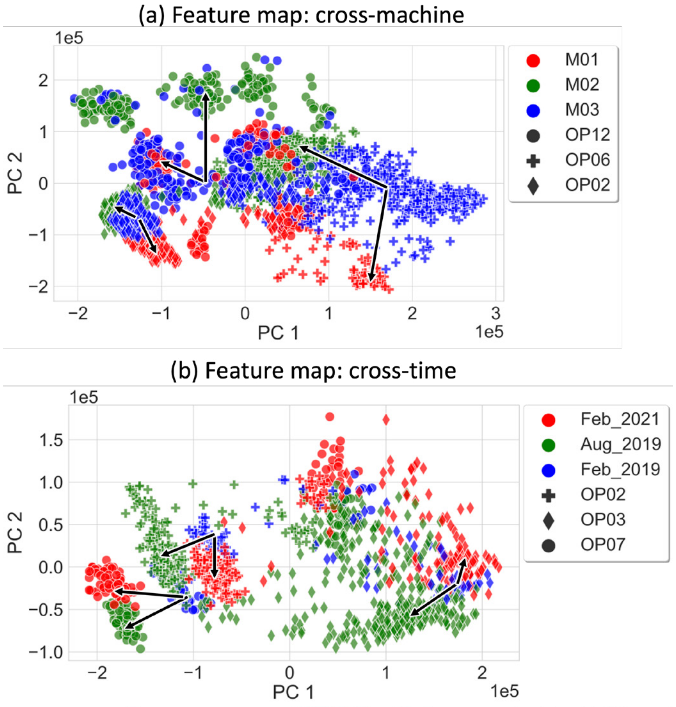

Fig. 6: Example of feature maps of 3 different OPs reduced into 2D using principal component analysis [17]. **(a)** plots the drift between the 3 machines (in Aug_2019). **(b)** plots the drift between the 3 largest time frames (in M01).

Considering the data stream challenges, the raw data from 3 different processes are first snipped using a sliding window with window length equal to 4096. This value has been defined empirically, due to the nature of the collected data and the known process steps. From each window, the most com-

This is a resupply of March 2023 as the template used in the publication of the original article contained errors. The content of the article has remained unaffected.

mon features for industrial time-series data are extracted using a generic feature extractor Tsfresh [18]. This includes summary statistics, characteristics of samples distribution and observed dynamics. To visualize the discrepancies, the high-dimensional features have been reduced to 2 dimensions using principal component analysis [17]. Figure 6.a visualizes the discrepancy cross-machines in a single time interval (Aug_2019) and highlights the challenge of scaling data-driven algorithms to solve industrial tasks. In a similar manner, in Figure 6.b, the drift of the data over time is depicted for the 3 largest time intervals from a single machine (M01).

To encounter the mentioned challenge, generalization of the ML models must be the main evaluation criteria. By building the training dataset, some processes should be kept aside to evaluate the performance of the models. The dataset published with this paper provides suitable content and structure to enable ML researchers to develop more robust models for such unavoidable environmental challenges from real life.

3.2.2. Industrial challenges

To enhance the ML generalization, our dataset presents an example of 15 different tool operations. As mentioned previously, each OP is characterized by a unique parametrization that results in different patterns in the time series signal, making it difficult to predict health status. Using the same pipeline as in Section 3.2.1, the features are extracted from the different OPs and the high-dimensional extracted features are mapped in a two-dimensional space using principle component analysis [17]. An overview of the reduced features is presented in Figure 7.

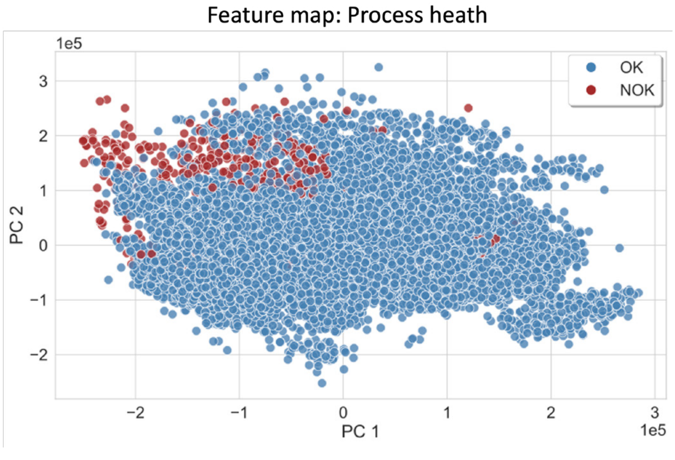

Fig. 7: Feature maps of the complete dataset reduced into 2D using principle component analysis [17].

In Figure 7, the distancing between the OKs and NOKs of the different OPs is illustrated. Some processes of the NOK class are easily distinguishable from the OK class. In others, it is difficult to distinguish between the OK class and the NOK class due to the difference in severity of the anomaly's impact. An example is shown in Figure 8, where a comparison between OP07 and OP08 in time and frequency-domain is conducted. It shows that the impact is more severe in OP07 than in OP08 and a clear divergence between the two processes in both time and frequency domains. However, a common observation is that the anomaly can be detected in frequencies which are integer multiples of the spindle speed. For this example of OP07, the frequency characteristics in the 200 Hz and 400 Hz regions therefore have visibly higher amplitude compared to the healthy process.

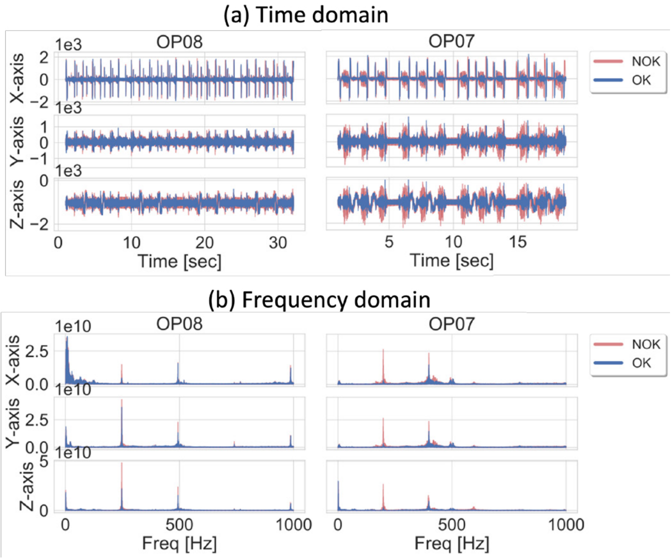

Fig. 8: Comparison of 2 different tool operations: OP07 vs. OP08.

To achieve rapid processing and non-intrusive solutions, time series signals are usually windowed at fixed length (WS). This technique is generally used as a data augmentation technique, especially for NOK data. The drawback of segmenting NOK data is that the label of small segments may not correspond to the complete process. This effect is mainly observed in the first and last extracts, where anomalies are not present yet. When labelling the published data, we truncated the start and end of the OP from the NOK samples. However, this issue can appear in the middle of the process due to fast position change. This can be seen in Figure 9, where a small snippet from the middle of OP08 of the OK and NOK classes matches exactly. To encounter this issue, a reasonable choice of WS needs to be defined.

The CNC Machining dataset provides the needed variety of samples and classes with different levels of discrimination that allow the research community to work on solutions in a systematic way and investigate the robustness of the data-driven methods to industrial challenges.

3.2.3. Dataset partitioning

By publishing this dataset, we encourage the research of ML models and learning techniques for noisy time-series data. To realistically measure performance in the real-world challenges, we propose three strategies for partitioning the CNC Machining dataset. With a machine-wise partitioning, as in Figure 10.a, the ability to perform on a new machine outside the training set is addressed. Using time-wise partitioning, as in Figure 10.b,

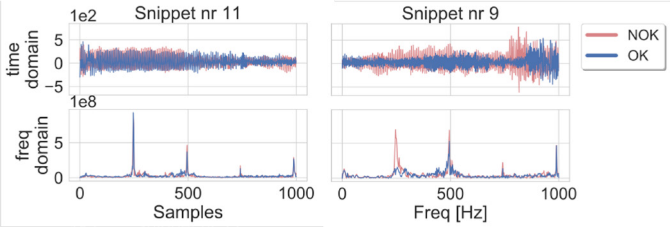

Fig. 9: Data segmentation causing faulty labels. Data taken from ”OP08_Feb2019_000” and windowed with $ws = 1000$. For sake of perfect overlap, the OK sample is cropped to the range [4650, 64231].

we address a data drift over time, by withholding some time intervals exclusively for validation and testing.

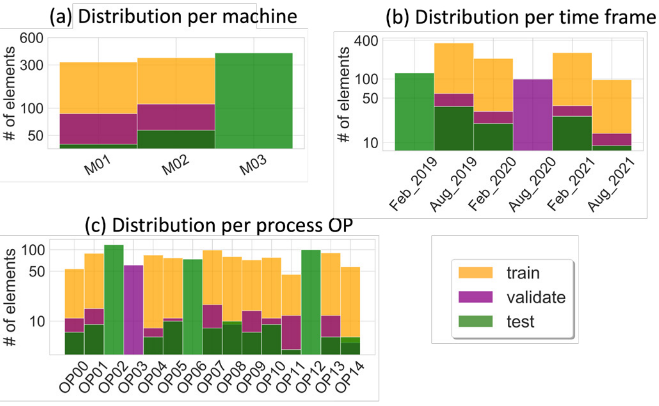

Fig. 10: Three strategies for dataset partitioning.

We intentionally suggest doing this already for the validation set and not only for the test set to be able to check overfitting during training. A third option for partitioning is the application of the same strategy on each process as in Figure 10.c.

4. Conclusion

Generalization is still a major challenge for industrial ML applications. To overcome this limitation, we proposed in this paper a challenging dataset from a real production plant. We depicted our smart data collection system based on an edge-to-cloud IoT architecture. The main benefit of this approach is, firstly, to retrofit brownfield CNC machinery where a direct measurement is extremely complicated, and secondly, enable the data science and domain expertise collaboration. With the presented system, vibration data has been collected from 3 different machines over a long time-interval. The data analysis showed that, with a low-cost accelerometer mounted in the rear side of the machine, process anomalies are detectable. The advantage of this approach is to avoid the extreme conditions from the front side, i.e. the machining area. Finally, to enhance ML and machine monitoring researches, we highlighted the environmental and industrial challenges embedded in the presented dataset. Some dataset scenarios have been proposed to enable the researchers to work on solutions in a systematic way. Future research will focus on development of robust ML architectures. Labeling and segmenting time-series data remain important topics and will be further investigated.

References

[1] Tim Stock and Günther Seliger. Opportunities of sustainable manufacturing in industry 4.0. *Procedia Cirp*, 40:536–541, 2016.

[2] A Quatrano, Simone De, ZB Rivera, and D Guida. Development and implementation of a control system for a retrofitted cnc machine by using arduino. *FME Transactions*, 45(4):565–571, 2017.

[3] Romulo G Lins, Bruno Guerreiro, Robert Schmitt, Jianing Sun, Marcio Corazzim, and Francis R Silva. A novel methodology for retrofitting cnc machines based on the context of industry 4.0. In *2017 IEEE International Systems Engineering Symposium (ISSE)*, 1–6, 2017.

[4] Chandra Nath. Integrated tool condition monitoring systems and their applications: a comprehensive review. *Procedia Manufacturing*, 48:852–863, 2020.

[5] Daniel Frank Hesser and Bernd Markert. Tool wear monitoring of a retrofitted cnc milling machine using artificial neural networks. *Manufacturing letters*, 19:1–4, 2019.

[6] T Mohanraj, S Shankar, R Rajasekar, NR Sakthivel, and Alokesh Pramanik. Tool condition monitoring techniques in milling processa review. *Journal of Materials Research and Technology*, 9(1):1032–1042, 2020.

[7] Zhiyuan Lu, Meiqing Wang, and Wei Dai. Machined surface quality monitoring using a wireless sensory tool holder in the machining process. *Sensors*, 19(8):1847, 2019.

[8] Vinh Nguyen and Shreyes N Melkote. Manufacturing process monitoring and control in industry 4.0. In *Proceedings of 5th International Conference on the Industry 4.0 Model for Advanced Manufacturing*. Springer, 144–155, 2020.

[9] System level Manufacturing and Automation Research Testbed (SMART) at the University of Michigan. Cnc milling dataset. https://www.kaggle.com/shasun/tool-wear-detection-in-cnc-mill, 2018. (Date accessed: 14.12.2021).

[10] A. Agogino and K. Goebel. Best lab, uc berkeley. milling data set, nasa ames prognostics data repository. https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/, 2007. (Date accessed: 14.12.2021).

[11] Bosch Connected Devices and Solutions GmbH. Connected industrial sensor solution. https://www.bosch-connectivity.com/media/downloads/ciss/ciss_datasheet.pdf, 2020. (Date accessed: 14.12.2021).

[12] Yang Hui, Xuesong Mei, Gedong Jiang, Tao Tao, Changyu Pei, and Ziwei Ma. Milling tool wear state recognition by vibration signal using a stacked generalization ensemble model. *Shock and Vibration*, 2019.

[13] Grzegorz Wszołek, Piotr Czop, Jakub Słoniewski, and Halit Dogrusoz. Vibration monitoring of cnc machinery using mems sensors. *Journal of Vibroengineering*, 22(3):735–750, 2020.

[14] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, jan 1949.

[15] Wen Sun, Jiajia Liu, and Yanlin Yue. Ai-enhanced offloading in edge computing: When machine learning meets industrial iot. *IEEE Network*, 33(5):68–74, 2019.

[16] Emre Yigitoglu, Mohamed Mohamed, Ling Liu, and Heiko Ludwig. Foggy: A framework for continuous automated iot application deployment in fog computing. In *2017 IEEE International Conference on AI & Mobile Services (AIMS)*, 38–45, 2017.

[17] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[18] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77, 2018.

This is a resupply of March 2023 as the template used in the publication of the original article contained errors. The content of the article has remained unaffected.