A Review of Publicly Available Datasets from Manufacturing Systems

Valentina Golendukhina
Computer Science Department
University of Innsbruck
Innsbruck, Austria
valentina.golendukhina@uibk.ac.at

Bianca Wiesmayr
LIT CPS Lab
Johannes Kepler University Linz
Linz, Austria
bianca.wiesmayr@jku.at

Michael Felderer
German Aerospace Center (DLR)
University of Cologne
Cologne, Germany
michael.felderer@dlr.de

***Abstract*—Smart manufacturing systems adapt to changes in the environment, which are detected via sensors. Decisions are then made by the control software of such manufacturing systems, e.g., by integrated AI-based algorithms. The design and evaluation of such algorithms require the availability of high-quality datasets. This paper provides an overview of the existing publicly available manufacturing datasets, offering a detailed exploration of the current landscape of shared data resources in the manufacturing sector and highlighting the utility of these datasets. The review identifies nine notable datasets with extensive documentation and comprehensive data coverage. Each of these datasets is described in detail and can be used for developing intelligent manufacturing systems, assessing their quality, and reporting on open gaps.**

***Index Terms*—Dataset, Control software, Manufacturing systems, Industry 4.0**

I. INTRODUCTION

Manufacturing enterprises use high-end automated production systems to meet the requirements posed by market demands, such as changes in customer needs, customization of products, or cost reduction [1]. These systems realize a set of processes that are required for the production of items, e.g., handling materials, processing the materials in manufacturing steps (such as molding, casting, or compacting materials), testing and inspecting the results, and controlling and planning the production [2]. Sensors can provide a variety of data points with information about these processes. They measure physical quantities, such as the temperature, current, or voltage of a system. Additionally, more abstract information about a production system is available, for instance, in the form of production orders. Modern manufacturing systems may additionally include complex perception methods, such as cameras for tracking their environment [1]. Based on the collected information, the software can control actuators, adjust production workflows [1], predict faults in the production system [3], and analyze the cause of a failure [4]. Data-driven methods, such as machine learning (ML), help realize and effectively automate these processes.

The heterogeneous data sources that are typical for manufacturing systems are a specific challenge for the development of data-driven algorithms [5]. The public availability of high-quality datasets can support the development of intelligent algorithms for production systems [6].

The primary motivation behind this study is to provide a base for understanding the diverse sensor outputs and data heterogeneity present in publicly available manufacturing datasets. Such an understanding is crucial for developing methodologies that can effectively interpret and leverage these data in real-world manufacturing settings. Moreover, it can be beneficial for knowledge transfer of manufacturing flow and initial learning of manufacturing processes and the typical data types for these processes. Furthermore, exploring datasets from the manufacturing domain offers significant benefits for ML development. ML models that are exposed to a wide array of data types and structures shall be used to generate knowledge of manufacturing processes [5] and can potentially improve the adaptability to changing environments, a common challenge in the dynamic field of manufacturing. Also, extensive documentation and background knowledge of the manufacturing processes can contribute to the explainability and understandability of the results generated by ML models. Lastly, the structured collection of high-quality manufacturing datasets offers a baseline for further examination of manufacturing processes and generalization of faults in the production flow and sensor data issues.

In this paper, we systematically review the existing white and grey literature on datasets from manufacturing systems to identify qualitative datasets that reflect the processes in manufacturing. We thoroughly examined 52 datasets, found nine publicly available and sufficiently documented datasets, and provided their overview, data types included, and purpose. Moreover, this review aims to identify challenges and issues in the currently published datasets and to outline the potential for integrating data-based analyses into manufacturing systems.

The remainder of this paper is structured as follows. Section II provides the background information on existing studies listing datasets collection and manufacturing datasets in particular. Section III describes the methodology behind the systematic dataset review process. Subsequently, Section IV presents the results, and Section V provides a discussion of the findings, identified challenges, and limitations. Finally, Section VI concludes the paper.

This work was supported by the Austrian Research Promotion Agency (FFG) in the frame of the project ConTest [888127]

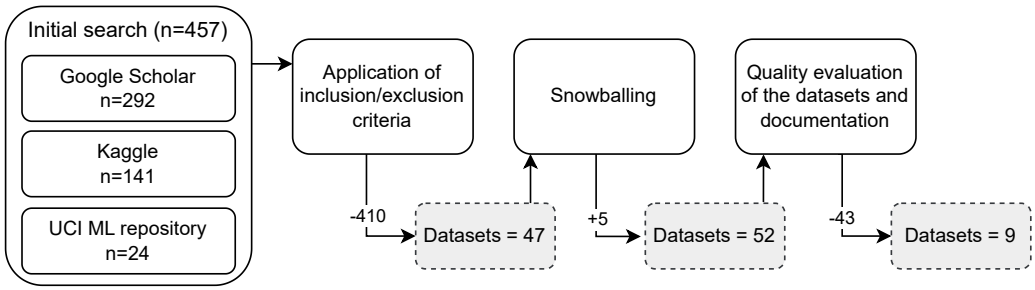

Fig. 1. Datasets review process

II. RELATED WORK

We consider related work that reviews datasets for manufacturing systems and categorizes applications of data-driven methods for manufacturing.

Overall, datasets can be categorized according to their scope [3]. Some datasets only cover individual elements, such as a single component of a system (e.g., a bearing [4]), while others cover individual machines or a whole production line [3]. Furthermore, datasets that are available in the literature can be categorized based on their target activity.

Hagmeyer et al. [4] reviewed 70 datasets for prognostics and diagnostics for industrial applications including datasets from the manufacturing domain. Although some of the datasets intersect with the current paper, the main focus of the authors was on datasets for predictive maintenance.

In another review by Jovicic et al. [7], the authors collected and analyzed 15 time series datasets for predictive maintenance from the energy industry. They exclusively searched in data repositories and evaluated these datasets based on the source (real or simulated), the size, the number of faults, and the accessibility of the datasets. Due to the focus of the review, only datasets that explicitly contained faults were included. Hence, also simulated datasets were considered useful because real datasets contain only a small number of faults. Regarding documentation, license information and a description of the variables were considered essential information.

Dogan and Birant [8] provided a comprehensive overview of the ML techniques applied in the manufacturing domain. The authors identified three manufacturing datasets that are most commonly used in ML studies: SECOM [9], Steel Plates Faults [10], and Bosch Production Line Performance [11]. The SECOM dataset [9] originates from a semiconductor manufacturing process and aims at predicting the outcome of the production. For 1567 examples, the timestamp and various feature values are provided together with the outcome. However, a description of these features is not part of the dataset, and only a few values per day were recorded. Hence, it does not allow any insight into the production process. The Steel Plates Faults dataset [10] has different variables including various lengths, the type of steel, luminosity, and several kinds of associated faults. The main purpose of the dataset is pattern recognition with computer vision algorithms, data about the manufacturing process itself is not included. Also, the Bosch Production Line Performance dataset [11] aims at predicting which parts fail quality control. A large number of features are included, but they are anonymized and represent measurement values along the production process. The authors of the review [8] provided the names of the other datasets used in each reviewed study, although a large number of such datasets was only discussed in the reviewed studies, but was not publicly available. The main focus of this paper was on reviewing the applied ML models, while our review focuses on publicly available datasets.

Kaupp et al. [3] identified several production lines related datasets out of which two datasets did not reveal the sensor types [9], [11], and the other two datasets focused on energy consumption of a factory topic. Also, the authors provide an extensive list of datasets collected from individual machines. The main purpose of the paper was to provide a dataset that fills the gap of data acquired by real sensors with clearly labeled features including the relevant measurement units. Hence, the quality of the identified dataset was not investigated.

All discussed reviews focus on time series or image data. In addition, image datasets are available in the literature, e.g., for identifying anomalies [7] or as a basis for human-robot interaction, e.g., [12].

III. METHODOLOGY

In this study, we systematically identify and collect publicly available datasets from manufacturing systems with the primary goal of uncovering datasets that demonstrate manufacturing processes recorded by two or more distinct sensors and include documentation of the underlying processes. The following section details our methodology for the identification and aggregation of these datasets, emphasizing our criteria for dataset selection and the search strategies employed.

A. Search Process

After performing initial searches in databases, the final search string was formulated including the terms retrieving the most true positive results. It is provided in Table I.

The search process was conducted using a combination of academic search engines and data repositories, specifically targeting sources known for their collections of industrial and scientific datasets.

TABLE I
SEARCH STRING

”dataset” AND (”manufacturing” OR ”production” OR ”automation” OR ”factory”)

The search for relevant datasets was conducted in three stages to ensure a thorough selection process that is depicted in Figure 1. Initially, a targeted search string was used to identify datasets from academic paper titles in selected search engines and databases. Inclusion and exclusion criteria were then applied to titles and abstracts. To ensure no significant sources were overlooked, a snowballing process was employed on the significantly relevant papers to uncover additional datasets. This led to the final selection and detailed examination of 52 datasets.

We utilized Google Scholar, ACM Digital Library, and IEEE Xplore to access scholarly articles and technical reports that detail experiments and studies involving multi-sensor datasets in manufacturing environments. Simultaneously, we explored data-centric platforms such as the UCI Machine Learning Repository[1], NASA Open Data Portal[2], and Kaggle[3], which are repositories that often contain real-world datasets, including those relevant to industrial applications. After applying the predefined search strings and deleting duplicates, i.e., any datasets that appear across multiple databases or research papers, we found three sources that provided the most relevant results: Google Scholar, Kaggle, and the UCI Machine Learning Repository.

To ensure the relevance and quality of the datasets included in our study, we established specific inclusion and exclusion criteria. **Inclusion criteria** were as follows:

Publicly accessible: Only publicly available datasets or datasets accessible after signing in as a research institution were selected for this review. Thus, the datasets requiring a membership were excluded, e.g., Kamp-ai[4] – a large Korean database providing real-world manufacturing datasets from the industry partners.

Data from two or more different sensors: A key focus of our study was on datasets that include data collected from multiple sensors covering different processes undergoing the manufacturing process.

Sufficiently documented: Each dataset's description and documentation were assessed to ensure that the process of data collection and information about sensors are included. The identified quality criteria for documentation are described in Table II. Only datasets with a score of three were added to the final table. The documentation must clearly describe the type of sensors used, the data collection and preprocessing methods, and the manufacturing processes.

Year of publishing 2004-2024: We focused on datasets published or updated within this time frame to ensure that the sensor technology reflected modern capabilities in terms of frequency, precision, and ease of data acquisition.

[1]https://archive.ics.uci.edu
[2]https://data.nasa.gov/browse?limitTo=datasets
[3]https://www.kaggle.com
[4]https://www.kamp-ai.kr

TABLE II
CRITERIA FOR ASSESSING THE QUALITY OF DATASET DOCUMENTATION

Score	Requirements
0	no documentation
1	variables names are not anonymized
2	variables provided, collection/preparation methods are described
3	variables, collection/preparation method, processes described

Tabular data: We specifically focused on datasets in tabular format, as sensor data is typically organized in this form.

Language: To ensure that the datasets are accessible to the widest possible research community, only those datasets documented in English or easily translated to English were considered.

We also established a set of **exclusion criteria** to further refine our dataset selection process:

Images or videos: We excluded datasets that primarily and exclusively consist of image or video data.

Process manufacturing related: Although process manufacturing systems are significant, our study specifically excludes these as they often involve chemical production and measuring approaches different from those in discrete manufacturing.

To further refine the quality of datasets sourced from platforms like Kaggle, where data is uploaded by a wide range of contributors and varies significantly in quality and relevance, we introduced additional criteria specifically tailored for such repositories:

- medium and large datasets (Kaggle filtering option),
- description provided, including sufficient information about the data origin and the types of sensors used,
- 10 or more upvotes: inspired by the GitHub stargazers approach [13], we applied a popularity filter requiring datasets to have received 10 or more upvotes on the platform. Popularity metrics can be indicative of community validation and can be an effective way to filter out less reliable and irrelevant datasets.

B. Data Extraction

After the search process was completed and the final set of datasets was identified, two researchers extracted detailed information from each dataset, provided that such information was available. The extracted details included:

- Dataset title.
- The list of authors.
- White paper: Any accompanying academic or technical white paper that describes the dataset in detail.
- Venue: The platform or repository where the dataset was published or made available.
- Year of publication.
- Data source: This includes whether the data originated from simulations, experiments, or actual real-world operations.
- Data acquisition: The method of data collection – continuous, periodic, or condition-dependent acquisition.

- Data preparation steps: Details of any preprocessing or cleaning steps that the data underwent before publication.
- Sensor types and their physical quantities: information about the types of sensors used and the physical measurements they capture.
- The purpose of the dataset, e.g., monitoring or anomaly detection.
- Link: The URL from where the dataset can be accessed.

To ensure objectivity and minimize researcher bias, the data inspection and extraction process was conducted iteratively by two independent researchers. Each dataset was initially reviewed by one researcher to extract relevant data points, which were subsequently examined and validated by a second researcher. This review process ensured the accuracy and reliability of the information collected but also provided a check against potential biases that might arise from individual interpretations.

IV. RESULTS

This section provides comprehensive information about the datasets discovered during the systematic review process and underlines the main categories of publicly available manufacturing datasets including examples that were excluded from the final list. Subsequently, nine chosen datasets are described.

A. Categorization of Results

Initially, over 50 datasets were analyzed, which were subsequently categorized into six groups to facilitate a clearer understanding of their applications and characteristics. The categories are provided and described below including dataset examples that were not included in the final selection. It is important to recognize that these categories are not mutually exclusive and should be seen as characteristics that emphasize the salient features of the datasets.

1) Human and Machine Interaction: This category of datasets often contains video and image data of human-machine interaction, e.g., during the assembly process [14] or for human location and activity detection in the facility [15]. However, the datasets including human-focused activities are not limited to visual data. Individual recordings are also used in tabular datasets on process management and event logging, e.g. in the extensively documented textile weaving dataset for machine learning [16]. Such datasets provide insights into workflow efficiency, and ergonomic and safety practices.

2) Individual Machines: The most common type among the found datasets. These datasets concentrate on specific machinery or movements within the manufacturing environment. Several datasets provide insights into bearing functioning and faults [17], [18]. Also, Kaggle mostly provides datasets on single types of machinery, e.g., analysis on a rotating shaft based on vibration data [19], or machinery fault dataset [20].

3) ML-Related Datasets Without Sensor Description: Commonly found on platforms like Kaggle, these datasets are used for developing ML models and do not typically include detailed sensor data or process descriptions. Such datasets provide a large number of features, but the origins of the data and information on the collection process are often either anonymized or not provided. The two largest datasets which are also considered benchmark datasets in the community are SECOM [9] - data from a semi-conductor manufacturing process providing information on 591 anonymized features and Bosch production line performance [11] describing chocolate soufflé production with the data covering over 1100 features.

4) Image and Video Datasets for Quality Assessment: This type of dataset includes visual recordings used to inspect the final products for defects. Quality inspection datasets can provide synthetically generated, real-world data, and real data enhanced with synthetically generated data. Such datasets always include annotations for defective and regular elements. An example of such a dataset is the steel plate faults dataset [10] providing a classification of seven different types of steel plate faults. An overview of industrial datasets for object detection in manufacturing is also found in [21].

5) Higher Abstraction Level Datasets: This category includes datasets not centered around sensor data but rather on the management aspects of manufacturing, such as order flow and production throughput or quality. An example is a publicly available textile weaving dataset for machine learning [16], which includes information on the weaving waste. Hence, it can be used to predict the waste and production outputs. Data entries include information such as the order number, yarn lengths, finished fabrics, and total production. It was recorded manually over nine months and the available dataset includes both, the unprocessed information and a pre-processed version that is suitable for ML training.

6) Datasets Covering a Production Line: These datasets aim to provide a holistic view of a production line, including all or key associated machinery and processes, and show the interdependencies and interaction between different sensors measuring the process flow. Such datasets are the focus of our review and are described in detail in the following subsection as well as in Table III.

B. Selected Datasets

In the following, we further describe the nine datasets that we consider useful for developing applications that cover a manufacturing system. The datasets are listed in Table III.

1) Bosch CNC Machining Dataset [22]: The benchmark dataset includes real-world industrial vibration data gathered from brownfield CNC milling machines over two years at a production plant using a smart data collection system. To provide data comparable to an industrial scenario, the researchers collected data from three different CNC machines. The authors provide detailed information about the sensor locations, recording procedures, frequencies, and the manufacturing processes involved. The dataset is designed primarily for the ML scientific community to address various challenges including feature drifts across different machines and over time, the wide variety of tool operations during production, and the imbalance in the dataset concerning the number of samples per class. The data is made available in the GitHub repository in the hierarchical data format (.h5) and can be

easily accessed with the code in Python 3.8 provided for data loading.

2) CONTEXT: An Industry 4.0 Dataset of Contextual Faults in a Smart Factory [3]: This dataset was recorded from a demonstrator for a smart factory with several stations. Hence, it does not provide data from a real production scenario but offers a realistic view. The dataset specifically aims to provide data that is clearly labeled and assigned to individual stations and covers different kinds of physical quantities. Additionally, the sensors and their value ranges including units are clearly stated. The recorded data is enhanced through so-called sensing units which collect additional information about the production process, but also compensate for lost information. In addition to the normal operation, this dataset covers fault scenarios, i.e., a leakage in the pressure system, a reduced throughput, or a missing part. The data are available in .CSV format.

3) A modular Ice Cream Factory Dataset on Anomalies in Sensors (MIDAS) [23]: The MIDAS dataset forms the basis for training ML for anomaly detection. This is a simulated dataset containing data from six production stages: mixer, pasteurizer, homogenizer, aging and cooling, dynamic freezer, and hardening. The authors simulated both the expected operation of the system and abnormal behavior with three types of injected anomalies. The paper extensively describes each step of the production process including the measurement units and anomalies ingestion process. Although the authors did not provide the simulation environment, they made the data available on the GitHub platform in .CSV format including Jupyter Notebook files for data reading, transformation to data frames, and model training.

4) Robotic Arm Dataset (RoAD) [24]: This dataset offers a detailed characterization of a collaborative robotic arm's motion and energy consumption within a production line, compiled and annotated by researchers. Although the dataset originates from a single robotic arm, we considered it relevant, e.g., as an assembly station in manufacturing. It includes data from the robotic arm's seven joints, capturing parameters like current, frequency, phase angle, power, power factor, reactive power, and voltage, alongside annotations for both standard and anomalous scenarios. Additionally, an API that is hosted on GitHub and is accessible via Python facilitates data manipulation and analysis.

5) Football Manufacturing Production Line Dataset [25]: This is a benchmark dataset for knowledge graph generation containing realistic data from a football production line. The dataset was created in collaboration with production line managers, supervisors, and engineers from the manufacturing industry and provides information about nine different types of machinery including laser cutting, oval printing, high-frequency cutting, glue sprayer, heating panel, forming could, ball shaping, ball seam glue, and heat drying conveyor. The production process is thoroughly described and provides information about the types of sensors embedded in each machinery and the respective measurements and their ranges. Researchers provided the data generation environment on GitHub that can be run in Python as well as distinct data files on Zenodo in .OWL format.

6) An Industrial Manufacturing Dataset [6]: The dataset is collected from a real-world CNC production line during the manufacturing of eight different types of products for around seven months. The authors provide a description of the three types of sensors and the process of data recording. The dataset was collected to train ML models for predictive maintenance and made available for results replication and further research. To support the classification tasks, the authors provided the ground truth data including labels for the normal and defective functioning of the machines ranging from minor to major and critical defect types and labeled by production workers. The dataset is provided in the hierarchical data format (.h5) on a stand-alone platform to support sustainable replication of the research results.

7) AI4I 2020 Predictive Maintenance Dataset [26]: This is a synthetic realistic dataset that models the data on six features: product, air temperature, process temperature, rotational speed, torque, and tool wear. The data is provided for the normal operation and machine failure states including five failure modes. For each sensor, the measurement units and interval values are given. The dataset aims to increase the understandability of the models trained on the provided data. The data can be easily accessed on the UCI Machine Learning Repository website as a single .CSV file.

8) Intel Lab Data [27]: The dataset consists of approximately 2.3 million sensor readings gathered from 54 sensors positioned in the Intel Berkeley Research lab. It provides detailed sensor activity over roughly a month and includes the sensors' locations on a lab map. Although it does not specifically focus on manufacturing, the dataset offers extensive insights into sensor operations during the observed period. While the dataset does not provide extensive details about the sensors themselves, it does include information on the measurement units used for measuring humidity, temperature, light, and voltage. All data (in .TXT format) and documentation can be accessed on a specially dedicated web page.

9) PRONTO Heterogeneous Benchmark Dataset [28]: The dataset was collected in the Process System Engineering Laboratory of Cranfield University utilizing a fully automated industrial-scale flow facility designed for water, air, and oil flow experiments. The facility contains various sensors including pressure, flow rate, temperature, and density sensors. The final dataset represents data from heterogeneous sources such as process measurements, alarm records, high-frequency pressure and ultrasonic sensors, operation logs, and video recordings. The authors extensively document the data collection process for each sensor and provide comprehensive information about the measurement units, frequency, and recording process and conditions. The data is collected during the normal and manually introduced faulty operating conditions. As the goal of the project was to deliver heterogeneous data from a manufacturing process, the dataset contains various data types including .CSV, .MP4, .JPG, and .mat for continuous data. The dataset is available on Zenodo.

TABLE III
AN OVERVIEW OF THE DATASETS COVERING A PRODUCTION LINE

	Dataset	White Paper	Venue	Year	Data Source	Data Acquisition	Data Preparation	Sensor Types	Goal	Link
1	Bosch CNC Machining Dataset	Tnani, M.A., Feil, M., and Diepold, K., 2022. Supervised Learning for CNC Machining Anomaly Detection	UCI ML / GitHub	2022	Real	Periodic	Data is manually segmented and structured	Acceleration, vibration	Process monitoring, fault prediction	https://github.com/boschresearch/CNC_Machining
2	CONTEXT: An Industry 4.0 Dataset of Contextual Sensors From Various Processes	Kaupp, L., Webert, H., Nazemi, K., Humm, B., and Burzlaff, F., 2021. Context-Aware Systems in Industry 4.0	Zenodo / Conf. on IMS	2021	Experiment	Continuous, condition-dependent	The data from different sources is manually synchronized	Temperature, light intensity, pressure, magnetic	Contextual faults prediction	https://zenodo.org/records/4034867
3	A Modular Ice Cream Factory Dataset	Markovic, T., Leon, M., Leander, B., and Punnek, S., 2023. Sensor Data Enhancement and Anomaly Detection in a Modular Ice Cream Factory	IEEE Access / GitHub	2023	Simulation	Continuous	Normalized	Temperature, flow rate	Anomaly detection	https://github.com/vujicictjana/MIDAS
4	Robotic Arm Dataset (RoAD)	Mascolini, A., Gaiardelli, S., Ponzio, F., Dall'Alba, D., 2023. Data-Driven Approaches for Robotic Arm Fault Prediction	IECON / GitLab	2023	Experiment	Continuous	Filtered by a Kalman filter to reduce measurement noise	Acceleration, velocity, orientation, temperature	Anomaly detection	https://gitlab.com/AlessioMascolini/roaddataset
5	Football Manufacturing Production Line Dataset	Yahya, M., Ali, A., Mehmood, Q., Yang, L., Brett, P., 2021. Enhancing Football Manufacturing with Semantic Web Technologies	GitHub / Semantic Web Journal	2021	Simulation	Continuous	Raw	Power consumption, temperature, pressure, location	Simulation	https://github.com/MuhammadYahta/ManufacturingDataset
6	An Industrial Manufacturing Dataset	Hoelzl, G., Zausinger, J., Kranz, M., Fleischmann, B. and Soller, S., An Industrial Manufacturing Dataset together with Anomaly Detection Results integrated in an Open & Stand Alone Sharing Platform for Sustainable Replication.	AllData23	2023	Real	Periodic	Split by product	Rotation speed, electric current	Anomaly detection	https://www.hasisaurus.at/DataSet.html
7	AI4I 2020 Predictive Maintenance Dataset	Matzka, S., 2020, September. Explainable artificial intelligence for predictive maintenance applications. In 2020 third international conference on artificial intelligence for industries (ai4i) (pp. 69-74). IEEE.	UCI ML	2020	Synthetic	Continuous	Raw	Rotation speed, temperature, torque	Explainability studies	https://archive.ics.uci.edu/dataset/601/ai4i+2020+predictive+maintenance+dataset
8	Intel Lab Data	Bodik, P., Hong, W., Guestrin, C., Madden, S., Paskin, M., Thibaux, R. Intel Lab Data, 2004	-	2004	Real	Continuous	Raw	Light, temperature, humidity, voltage	Fault prediction	https://db.csail.mit.edu/labdata/labdata.html
9	PRONTO Heterogeneous Benchmark Dataset	Stief, A., Tan, R., Cao, Y., Ottewill, J.R., Thornhill, N.F. and Baranowski, J., 2019. A heterogeneous benchmark dataset for data analytics: Multiphase flow facility case study. Journal of Process Control, 79, pp.41-55.	Zenodo	2019	Real	Continuous, condition-dependent	Raw	Temperature, pressure, flow rate, density	Monitoring, fault prediction	https://zenodo.org/records/1341583#.X8Lsx17Q8Wp

V. DISCUSSION

This section discusses the results of the paper and highlights the issues in the publicly available datasets in the manufacturing domain. Then, the threats to validity and their mitigation strategies are described.

A. Datasets Issues

A challenge in this study was the small number of publicly available datasets with clearly labeled features. In the final list of datasets, four datasets originate from a real process, two are collected from an experiment, and three are simulated or synthetic datasets. With the progression of Industry 4.0 and the broad implementation of the industrial Internet of Things (IoT), there has indeed been a notable increase in data availability. However, when analyzing the available open datasets, the data often comes with inadequate documentation, presenting significant challenges for researchers and practitioners aiming to leverage this data effectively. As highlighted by Matzka [26], the scarcity of well-documented predictive maintenance datasets is an issue in the ML community that was one of the motivations for the creation of a synthetic dataset. The same motivation was provided by the authors of the CONTEXT dataset [3]. Their dataset originates from a demonstrator machine because datasets from real-world machines are frequently anonymized and obfuscated on purpose to hide sensitive information about a production process.

Crucial details such as sensors' names, types, positions, and the specific manufacturing processes they monitor are frequently not disclosed. The anonymization, often intended to preserve confidentiality, is common for production processes. Often the data is anonymized in a way that no sensor or process is identifiable to achieve confidentiality purposes, e.g. Bosch production line performance dataset [11] and numerous other datasets hosted on platforms like Kaggle. This also explains the small number of datasets in the manufacturing and production domain that is available in the literature. Moreover, no datasets from the Kaggle platform were added to the final list due to their poor documentation.

The accurate identification of sensor types is crucial for the precise interpretation of data and for ensuring that the data is applicable to specific industrial scenarios. This level of detail is essential for researchers and practitioners to understand and utilize the data effectively in context-specific applications. However, a significant issue in ML research is the frequent omission of citations or the non-disclosure of the training datasets. This trend is shown in [8]. The authors provide a review of over 50 studies relevant to data mining in manufacturing, but only three links to training datasets were provided in the paper.

The lack of a unified standard for citing datasets in scholarly articles poses challenges for researchers attempting to find references within white papers, leading to issues in verifying data sources and replicating studies. Furthermore, while some articles may include accompanying files, these are often limited to scripts for processing and analyzing data rather than the datasets themselves. This limitation further complicates efforts to access and utilize the actual data.

Moreover, the absence of units of measurement – such as pounds or kilograms – in datasets further complicates the interpretation of data and typical ranges of values that occur in production systems. Without standard units, the data's utility is reduced, as it becomes challenging to apply the findings across different systems or to compare them with other datasets effectively.

Despite an increasing trend in the publication of datasets that is also observed in our findings and likely driven by a growing recognition of the replication crisis within the scientific community, the quality of these dataset publications reflects the issues described above. Many datasets are either not published at all or are poorly described, which leaves substantial gaps in the availability of accessible and usable data.

Enhancing the generalization and robustness of ML algorithms is a significant challenge that could be mitigated by the availability of diverse datasets from various manufacturing facilities. The provision of more datasets, particularly those encompassing a broad spectrum of operational contexts, would facilitate the advancement of ML models. These improvements would support more generalized applications across different industrial contexts, leading to more robust and adaptable ML solutions. Moreover, proper documentation of manufacturing processes and sensors can contribute to the understandability and explainability of ML models.

Additionally, practitioners can gain significantly from the availability of diverse and well-documented datasets. With adequate documentation, these datasets become invaluable tools not only for operational enhancement but also for educational purposes. Detailed dataset descriptions, including sensor specifications, measurement units, and contextual information about the manufacturing processes, allow practitioners to better understand and simulate real-world scenarios. This knowledge transfer is crucial for training and developing skills in experts within the manufacturing sector.

B. Limitations

Although our search string was carefully designed in pilot searches which were conducted by two authors independently, it may not cover all relevant results, e.g., due to synonyms that were not considered. We mitigated this bias by additionally applying backward and forward snowballing and comparing our results to those of related reviews. We mainly focused on searching in the index Google Scholar but also used our search string in formal databases like IEEE Xplore, which did not yield any additional results. Nevertheless, our results may be incomplete and there might be datasets that we could not identify within the databases.

The inclusion and exclusion of each dataset were carefully discussed among two of the authors based on the pre-defined criteria to reduce bias in the filtering process. To identify the most relevant results, we included white and grey literature, but we did not include any closed-access and commercial

datasets. This approach allowed us to comprehensively cover both theoretical and practical perspectives, ensuring a thorough exploration of available resources that describe multi-sensor interactions in manufacturing systems.

We did not contact authors to receive access to raw data because we were interested in publicly available datasets. However, we cannot exclude the possibility that we missed datasets where the download location was not explicitly mentioned in the paper.

VI. CONCLUSION

As manufacturing processes become increasingly digitized, the ability to effectively work with and understand this data is crucial. This paper has provided a detailed overview of nine currently available manufacturing datasets that are well-documented and cover multiple aspects of the production line, offering a comprehensive resource for researchers and practitioners. Employing a systematic approach, our research has aimed to identify all open datasets within the manufacturing domain, highlighting the extensive opportunities for data-driven innovation.

However, a significant challenge that persists is the lack of well-documented datasets coupled with a deficiency in datasets that incorporate domain experts' knowledge. This gap challenges both understandability and explainability, crucial factors in leveraging data for meaningful insights. Addressing these challenges is essential for advancing our capabilities in manufacturing analytics and fully realizing the potential of Industry 4.0.

REFERENCES

[1] A. Bannat, T. Bautze, M. Beetz, J. Blume, K. Diepold, C. Ertelt, F. Geiger, T. Gmeiner, T. Gyger, A. Knoll, C. Lau, C. Lenz, M. Ostgathe, G. Reinhart, W. Roesel, T. Ruehr, A. Schuboe, K. Shea, I. Stork, S. Stork, W. Tekouo, F. Wallhoff, M. Wiesbeck, and M. F. Zaeh, "Artificial cognition in production systems," *IEEE Transactions on Automation Science and Engineering*, vol. 8, no. 1, pp. 148–174, 2011.

[2] D. G. Sorensen, T. D. Brunoe, and K. Nielsen, "A classification scheme for production system processes," *Procedia CIRP*, vol. 72, pp. 609–614, 2018. 51st CIRP Conference on Manufacturing Systems.

[3] L. Kaupp, H. Webert, K. Nazemi, B. Humm, and S. Simons, "Context: An industry 4.0 dataset of contextual faults in a smart factory," *Procedia Computer Science*, vol. 180, pp. 492–501, 2021. 2nd International Conference on Industry 4.0 and Smart Manufacturing (ISM 2020).

[4] S. Hagmeyer, F. Mauthe, and P. Zeiler, "Creation of publicly available data sets for prognostics and diagnostics addressing data scenarios relevant to industrial applications," *International Journal of Prognostics and Health Management*, vol. 12, no. 2, 2021.

[5] S. Kamm, N. Sahlab, N. Jazdi, and M. Weyrich, "A concept for dynamic and robust machine learning with context modeling for heterogeneous manufacturing data," *Procedia CIRP*, vol. 118, pp. 354–359, 2023.

[6] G. Hoelzl, J. Zausinger, M. Kranz, B. Fleischmann, and S. Soller, "An industrial manufacturing dataset together with anomaly detection results integrated in an open & stand alone sharing platform for sustainable replication," 2023. ALLDATA 2023: The Ninth International Conference on Big Data, Small Data, Linked Data and Open Data.

[7] E. Jovicic, D. Primorac, M. Cupic, and A. Jovic, "Publicly available datasets for predictive maintenance in the energy sector: A review," *IEEE Access*, vol. 11, pp. 73505–73520, 2023.

[8] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Systems with Applications*, vol. 166, p. 114060, 2021.

[9] M. McCann and A. Johnston, "SECOM Manufacturing Data." https://archive.ics.uci.edu/ml/datasets/secom, 2008. Accessed: 23.04.2024.

[10] M. Buscema, S. Terzi, and W. Tastle, "Steel Plates Faults Data Set." https://archive.ics.uci.edu/ml/datasets/Steel+Plates+Faults, 2010. Accessed: 23.04.2024.

[11] M. Risdal, Prasanth, R. Ghosh, Soundar, S. W., and W. Cukierski, "Bosch Production Line Performance." https://kaggle.com/competitions/bosch-production-line-performance, 2016.

[12] L. Duarte and P. Neto, "Event-based dataset for the detection and classification of manufacturing assembly tasks," *Data in Brief*, vol. 54, p. 110340, 2024.

[13] N. Munaiah, S. Kroh, C. Cabrey, and M. Nagappan, "Curating github for engineered software projects," *Empirical Software Engineering*, vol. 22, pp. 3219–3253, 2017.

[14] K. Moriwaki, G. Nakano, and T. Inoshita, "The brio-ta dataset: Understanding anomalous assembly process in manufacturing," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 1991–1995, IEEE, 2022.

[15] D. J. Rude, S. Adams, and P. A. Beling, "A benchmark dataset for depth sensor based activity recognition in a manufacturing process," *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 668–674, 2015.

[16] T. Ahmed and S. Uddin, "Textile weaving dataset for machine learning to predict rejection and production of a weaving factory," *Data in Brief*, vol. 47, p. 108995, 2023.

[17] C. S. of Engineering, "Bearing data set." Case Western Reserve University Bearing Data Center, 2023. Accessed: 24.04.2024.

[18] S. Kim, D. An, and J.-H. Choi, "Diagnostics 101: A tutorial for fault diagnostics of rolling element bearing using envelope analysis in matlab," *Applied Sciences*, vol. 10, no. 20, p. 7302, 2020.

[19] O. Mey, W. Neudeck, A. Schneider, and O. Enge-Rosenblatt, "Machine learning-based unbalance detection of a rotating shaft using vibration data," pp. 1610–1617, 09 2020.

[20] F. M. L. Ribeiro, "Machinery Fault Dataset." https://www.kaggle.com/datasets/uysalserkan/fault-induction-motor-dataset, 2021.

[21] H. M. Ahmad and A. Rahimi, "Deep learning methods for object detection in smart manufacturing: A survey," *Journal of Manufacturing Systems*, vol. 64, pp. 181–196, 2022.

[22] M.-A. Tnani, M. Feil, and K. Diepold, "Smart data collection system for brownfield cnc milling machines: A new benchmark dataset for data-driven machine monitoring," *Procedia CIRP*, vol. 107, pp. 131–136, 2022.

[23] T. Markovic, M. Leon, B. Leander, and S. Punnekkat, "A modular ice cream factory dataset on anomalies in sensors to support machine learning research in manufacturing systems," *IEEE Access*, vol. 11, pp. 29744–29758, 2023.

[24] A. Mascolini, S. Gaiardelli, F. Ponzio, N. Dall'Ora, E. Macii, S. Vinco, S. Di Cataldo, and F. Fummi, "Robotic arm dataset (road): A dataset to support the design and test of machine learning-driven anomaly detection in a production line," in *IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society*, pp. 1–7, IEEE, 2023.

[25] M. Yahya, A. Ali, Q. Mehmood, L. Yang, J. G. Breslin, and M. I. Ali, "A benchmark dataset for industry 4.0 production line and generation of knowledge graphs," *Semantic Web Journal*, 2021.

[26] S. Matzka, "Explainable artificial intelligence for predictive maintenance applications," in *2020 3rd Int. Conf. on Artificial Intelligence for Industries (AI4I)*, pp. 69–74, IEEE, 2020.

[27] P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin, and R. Thibaux, "Intel lab data." Online, 2004. Accessed: 25.04.2024.

[28] A. Stief, R. Tan, Y. Cao, J. R. Ottewill, N. F. Thornhill, and J. Baranowski, "A heterogeneous benchmark dataset for data analytics: Multiphase flow facility case study," *Journal of Process Control*, vol. 79, pp. 41–55, 2019.