Machine Learning For Intelligent Maintenance And Quality Control: A Review Of Existing Datasets And Corresponding Use Cases

Conference Paper · April 2022

DOI: 10.15488/11280

CITATIONS
4

READS
490

4 authors:

Nicolas Jourdan
Technische Universität Darmstadt
12 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Lukas Longard
Technische Universität Darmstadt
8 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Tobias Biegel
Technische Universität Darmstadt
6 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Joachim Metternich
Technische Universität Darmstadt
215 PUBLICATIONS **2,507** CITATIONS
SEE PROFILE

2nd Conference on Production Systems and Logistics

Machine Learning For Intelligent Maintenance And Quality Control: A Review Of Existing Datasets And Corresponding Use Cases

Nicolas Jourdan[1,*], Lukas Longard[1,*], Tobias Biegel[1,*], Joachim Metternich[1]
[1]TU Darmstadt, Institute of Production Management, Technology and Machine Tools (PTW), Darmstadt, Germany

Abstract

The advent of artificial intelligence and machine learning is influencing the manufacturing industry profoundly, enabling unprecedented opportunities to improve manufacturing processes within the three dimensions time, quality and cost. With the introduction of digitization and industry 4.0, increasing amounts of data become available for processing and use in smart manufacturing systems. However, the various use cases for machine learning in manufacturing often require problem-specific datasets for training and evaluation of algorithms which are difficult to acquire, hindering both practitioners and academic researchers in this area. As the respective data frequently contains sensitive information, manufacturing companies rarely release datasets to the public. Further, the relevant attributes and features of available datasets are usually not evident, requiring time-consuming analysis to evaluate if a dataset fits a given problem. As a result, it can be challenging to develop and evaluate machine learning methods for manufacturing systems due to the lack of an overview of available datasets. This paper presents a comprehensive overview of 47 existing, publicly available datasets, mapped to various use cases in manufacturing with the goal of simplifying and stimulating research. The characteristics of the datasets are compared using a set of descriptive attributes to provide an outline and guidance for further research and application of machine learning in manufacturing. In addition, suitable performance metrics for the evaluation of classification use cases in manufacturing are presented.

Keywords

Machine Learning; Artificial Intelligence; Manufacturing; Dataset; Benchmark; Metric; Evaluation;

1. Introduction and methodology

Machine Learning (ML) techniques increasingly transcend from research to practical applications in various industries. One of the industry areas that received significant attention in this context is the manufacturing industry [1]. The growing interest in manufacturing-related ML applications is fueled by the digitization of manufacturing processes in the context of industry 4.0 and the Internet of Things (IoT) [2]. However, the introduction of ML in manufacturing faces several challenges, with one of the most important being the acquisition of datasets for the development, training and evaluation of ML algorithms in high quantity. A sufficient data basis is crucial for the development of ML algorithms and strongly influences the achievable performance of the system [1]. Not only the quantity, but also the quality of the available data is of importance. Issues such as missing values, class imbalance, varying sampling frequencies and data types as well as high dimensionality have to be handled by the developers through pre-processing the data before

[*] Authors contributed equally

training the algorithms [3]. Compared to other ML application fields such as autonomous driving, only few publicly available datasets exist for manufacturing. Companies often see process data as sensitive information that cannot be shared due to privacy concerns [4]. As a result, the majority of studies that show successful applications of ML in manufacturing use cases do not share their training and testing datasets publicly, preventing an effective comparison between approaches [1]. In conclusion, an overview of publicly available datasets for ML applications in manufacturing is required to assist researchers and practitioners in the development and evaluation of algorithms and to enable the comparison of ML approaches in research studies. Few of such reviews (e.g. [4]) exist and to the best of the authors knowledge, none exist that account for modalities such as images, which are increasingly used in manufacturing-related ML applications [5–7]. In this study, the search for datasets was conducted on open platforms that provide dataset and code hosting for research and public competition purposes. The platforms include in alphabetical order: *GitHub [8], Kaggle [9], Mendeley Data [10], NASA Prognostics Center of Excellence (PCoE) [11], OpenML [12], University of California Irvine (UCI) Machine Learning Repository [13].* In addition to the resulting file storage ressources of the datasets, a search regarding accompanying publications that first release, describe and/or use the datasets for research, was conducted. Using the resulting publications, snowballing was applied to identify additional datasets that are hosted on platforms such as the universities of the corresponding authors. In total, 47 datasets have been identified and analysed regarding the comparison parameters. In addition to the selection or creation of an appropriate dataset for model training, the performance evaluation is an integral part of the model development process. Especially for classification tasks, the selection of an appropriate evaluation metric requires a deep understanding of the pursued task and relevant requirements [14]. Thus, a search for studies that utilize the identified datasets was conducted using Scopus, yielding 127 publications. These publications were consequently analysed regarding the applied performance evaluation metrics. The remainder of the paper is structured as follows: Chapter 2 introduces the identified datasets, sorted by the respective use cases. In Chapter 3, the most widespread classification metrics used for the evaluation of ML applications in manufacturing-related research are explained and critically analysed. Lastly, the conclusion is presented in Chapter 4.

2. Datasets and use cases in manufacturing

With the ongoing digitization there is an increasing number of research efforts emerging that focus on manufacturing-related problems. The datasets listed in this paper are suitable to address a subset of these problems and may be used to train and test ML methods specifically designed for manufacturing.

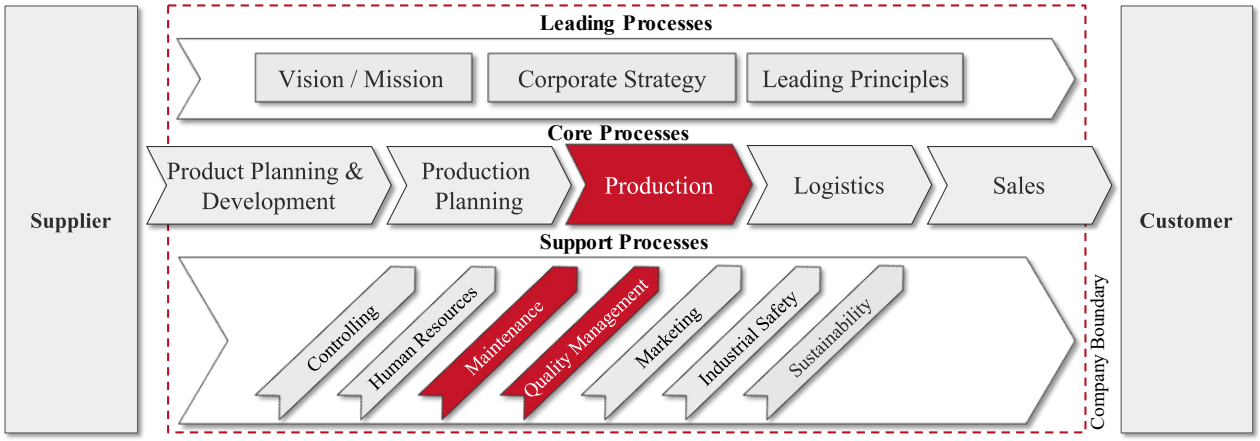

Figure 1: Process map highlighting the supporting areas that contain the use cases for all the presented datasets, adapted from [15].

In order to facilitate and further stimulate future research in this area, we identify common use cases for ML applications in manufacturing and map them to relevant datasets with the aim of accelerating the search for a suitable dataset based on the respective use case. The use cases defined in this paper are assigned to the two supporting processes: maintenance and quality management, as illustrated in the process map in Figure 1. The process map depicts the processes within the company boundaries that are necessary to meet customer demands. Each value adding activity requires support from indirectly value adding processes [15]. In terms of the core process production, the highlighted supporting processes embody two of the primary application areas of ML research in manufacturing nowadays and account for all the datasets listed in this paper [16].

The tables containing the identified datasets for the respective use cases are structured as follows: The *Name* column provides a short identifier to the datasets as well as a short description of the individual setting. The identifier is adopted from the original dataset source if available, otherwise it is newly created. Since a subset of the presented datasets is contained in [4], we adopted the respective names if applicable. Further, the *Type* and *Count* of the available non-target variable features is shown. In case of images, the resolution of the images is displayed as the feature count. The *Target Variable* column describes whether there are labels available for supervised learning: "C (*N*)" indicates that classification labels for *N* classes are given, while "R" indicates a supervised regression task. In some cases, labels for both, classification and regression, are available. The *Instances* column indicates the number of samples, e.g., rows, a given dataset contains. Further, *Official Train/Test Split* specifies, whether the dataset publishers provide a designated train / test data split for evaluation. This is especially important for comparability in research studies, as the same test split must be used to be able to compare the performance of different approaches. A consequent † indicates, that the target labels for the official test split are not publicly available, but rather hidden behind an evaluation server, which guarantees a fair benchmarking of approaches. The column *Data Source* highlights whether the dataset was collected from a real process, or rather generated synthetically using a simulation. Lastly, *Format* shows the file formatting of the raw data. In some cases, multiple formats are given, e.g., images in PNG-format and corresponding labels in XML-format. In the references section, the corresponding URL to the dataset is given together with the publication where the dataset was first introduced, if available. The use cases for each supporting process are presented in the following.

2.1 Maintenance (predictive maintenance and condition monitoring)

Predictive maintenance and condition monitoring are two terminologies that are used interchangeably by some researchers, while others view condition monitoring as part of the broader concept of predictive maintenance [17–20]. For the purpose of this paper, we follow the latter approach and thus consider condition monitoring as being a part of predictive maintenance. The use case predictive maintenance involves the data-driven assessment of the health status of machine components and sees its major objective in predicting the Remaining Useful Lifetime (RUL) of these components in order to reduce maintenance cost while simultaneously preventing unplanned downtimes. In machining applications such as milling, turning, or drilling, this, for instance involves the monitoring of the cutting tool to assess the current wear state, followed by the prediction of the RUL. Another typical scenario can be seen in the monitoring and RUL prediction of bearings. In this paper, we consider the data-driven health assessment as being integral to condition monitoring while the prediction of the RUL constitutes the broader case of predictive maintenance. The corresponding datasets for predictive maintenance and condition monitoring are exhibited in Table 1 on the following page.

Table 1: Datasets for predictive maintenance and condition monitoring

Name	Year	Features		Target Variable	Instances	Official Train/Test Split	Data Source	Format
		Type	Count					
Diesel Engine Faults Features [21] *Fault detection based on pressure curves and vibration.*	2020	Signal	84	C (4)	3.500	×	Syn.	MAT
Degradation of a Cutting Blade [22] *Wrapping machine process data over 12 months with a degrading cutting tool.*	2019	Signal	9	-	1.062.912	×	Real	CSV
CNC Mill Tool Wear [23] *CNC process data of wax milling with worn/unworn tools.*	2018	Signal	48	C (3*2)	25.286	×	Real	CSV
Condition Monitoring of Hydraulic Systems [24] *Test rig process data of multiple load cycles with various fault types and severity levels.*	2018	Signal	17	C (5*(2-4))	2.205	×	Real	Other
Production Plant Data for Condition Monitoring [22] *Anonymized process data of component run-to-failure experiments.*	2018	Signal	26	-	228.414	×	Real	CSV
Versatile Production System [25] *Popcorn production process data with multiple process steps.*	2018	Signal	5-85	-	80.000	×	Real	CSV
Degradation Measurement of Robot Arm Position Accuracy [26] *Target- and actual values of robotic arm tool position, velocity and current for health assessment.*	2017	Signal	73	-	155.000	×	Real	CSV
APS Failure at Scania Trucks [27] *Anonymized counters and histograms for air pressure system fault detection.*	2016	Signal	170	C (2)	76.000	✓	Real	CSV
Maintenance of Naval Propulsion Plants [28] *Gas turbine process data for component decay state prediction.*	2016	Signal	16	R	11.934	×	Syn.	Other
Plant Fault Detection [29] *Anonymized process data for plant fault detection.*	2015	Signal	10	C (6)	8.938.370	×	Real	CSV
Asset Failure and Replacement [30] *Anonymized data for asset fault detection.*	2014	Signal	1	C (2)	447.341	✓†	Real	CSV
Maintenance Action Recommendation [31] *Anonymized process and maintenance data of an industrial asset for maintenance action recommendation.*	2013	Signal	32	C (14)	2.097.152	✓†	Real	CSV
Anemometer Fault Detection [32] *Anemometer measurements for fault detection*	2011	Signal	16 16-20	-	345.700 208.800	✓†	Real	Other
Gearbox Fault Detection [33] *Test rig accelerometer data for fault detection.*	2009	Signal	3	-	> 10 Mio.	×	Real	CSV
Li-Ion Battery Aging [34] *Battery test rig data during charge and discharge cycles for degradation detection.*	2008	Signal	12	-	2.167	×	Real	MAT
Turbofan Engine Degradation Simulation [35] *C-MAPSS simulation sensor data of various conditions and fault modes.*	2008	Signal	26	-	262.256	✓	Syn.	Other
Bearing [36] *Bearing test rig accelerometer data of run-to-failure experiments.*	2007	Signal	4-8	-	61.440	×	Real	CSV
Milling [37] *Milling process- and external sensor data for tool wear detection.*	2007	Signal	13	R	1.503.000	×	Real	MAT
CWRU Bearing Data [38] *Bearing test rig accelerometer data for fault detection.*	n.A.	Signal	5	C (2)	> 10 Mio.	×	Real	MAT

2.2 Quality management

The subprocess quality management embodies the use cases process monitoring, predictive quality, quality inspection and process parameter optimization. The respective use cases and corresponding datasets will be introduced in the following subsections.

2.2.1 Process monitoring

The analysis of sensor-based process data can yield valuable information for the purpose of process control and quality monitoring [39]. The idea of process monitoring is to understand the variation in a process and to assess its current state [40]. A widely used technique in this field is control charting which involves two distinct monitoring phases, i.e. phase I and phase II [41]. In phase I, control charts are used to retrospectively test whether the process was in control after the data have been sampled from the process. The result of this phase is a Normal Operating Condition (NOC) dataset in which the underlying process is assumed to be in-control. With the help of the NOC dataset, control limits are established based on which a new observation of process data will be evaluated. This is the objective of phase II. Process monitoring has been an active research field throughout the last decades [42]. Especially within the process industry, the application of Multivariate Statistical Process Monitoring (MSPM) methods gained popularity [43,44]. In terms of discrete manufacturing, recent research focusses on the initiation of a paradigm shift from the conventional post-process Statistical Process Control (SPC), i.e. inferring the process condition based on measurements taken from the manufactured product to the so called in-process SPC that aims at inferring the process condition based on actual process data [45]. In both fields of industry, the application of ML, especially Deep Learning (DL) is receiving more and more attention and provides promising results for future research in this field [46,47]. The corresponding datasets for process monitoring are exhibited in Table 2.

Table 2: Datasets for process monitoring

Name	Year	Features		Target Variable	Instances	Official Train/Test Split	Data Source	Format
		Type	Count					
High Storage System Anomaly Detection [48] *Storage test rig process data for anomaly detection.*	2018	Signal	20	C (2)	91.000	×	Syn.	CSV
Genesis Pick-and-Place Demonstrator [49] *Material sorting test rig process data for anomaly detection.*	2018	Signal	23	C (3)	32.440	×	Real	CSV
Tennessee Eastman Process Simulation Dataset [50] *Simulated chemical process data for anomaly detection with different fault types.*	2017	Signal	51	C (21) / R	> 10 Mio.	✓	Syn.	RData
Robot Execution Failures [51] *Force and torque measurements of an industrial robot with different erroneous operating conditions.*	1999	Signal	89	C (13)	463	×	Real	Other
Mechanical Analysis [52] *Vibration measurements of electromechanical devices with different erroneous operating conditions.*	1990	Signal	7	C (6)	209	✓	Real	MAT
CWRU Bearing Data [38] *Bearing test rig accelerometer data for anomaly detection.*	n/a	Signal	5	C (2)	> 10 Mio.	×	Real	MAT

2.2.2 Predictive quality and quality inspection

The use case predictive quality incorporates the scenario where the prediction of the product quality is of primary concern. The accurate prediction of the product quality can be used to better control the manufacturing process [53]. The costs of delayed discovery of nonconformities in the product lifecycle increase exponentially the further the product moves down the value-chain [54]. Therefore, it becomes useful to predict if a product will fail specification tests in later stages of the process if the cycle times of a process chain are very long [55].

Table 3: Datasets for predictive quality and quality inspection.

Name	Year	Features		Target Variable	Instances	Official Train/Test Split	Data Source	Format
		Type	Count					
Casting Product Quality Inspection [6] *Grayscale images of pump impeller castings with and without defects.*	2020	Image	300×300 512×512	C (2)	7.348	✓	Real	JPG
GC10-DET [56] *Grayscale images of metal surfaces with various defect types and corresponding bounding box annotations.*	2020	Image	Varying	C (10)	3.570	×	Real	JPG, XML
Mechanic Component Images [7] *Grayscale images of air conditioner pistons with various defect types.*	2020	Image	86×90	C (3)	285	×	Real	PNG
Multi-Stage Continuous Flow Process [57] *Anonymized process data of a production line with quality measurements of part dimensions.*	2020	Signal	116	-	14.088	×	Real	CSV
Plastic Extrusion Defects [58] *Process data of a plastic extrusion process.*	2020	Signal	470	-	226.536	×	Real	CSV
AITEX [59] *Grayscale images of textile fabrics with various defect types and corresponding segmentation masks.*	2019	Image	4096×256	C (13)	245	×	Real	PNG, Mask
Deep PCB [60] *Grayscale images of circuit boards with various defect types and corresponding bounding box annotations.*	2019	Image	640×640	C (7)	1.500	✓	Real	JPG, Mask
Severstal Steel Defect Detection [61] *Grayscale images of steel surfaces with various defect types and corresponding segmentation polygons.*	2019	Image	1600×256	C (5)	18.074	✓†	Real	JPG, CSV
Turning Dataset for Chatter Diagnosis [62] *Sensory data of a turning test rig and varying strengths of chatter.*	2019	Signal	8	C (4)	>10 Mio.	×	Real	MAT
Magnetic Tile Defect [63] *Grayscale images of magnetic tile surfaces with various defect types and corresponding segmentation masks.*	2018	Image	248×373	C (6)	1.344	×	Real	JPG, PNG
TIG Welding [5] *Grayscale images of a welding process with various defect types.*	2018	Image	800×974	C (6)	33.254	✓	Real	PNG, JSON
Mining Process [64] *Process data of a mining process for impurity prediction in ore concentrate.*	2017	Signal	24	R	737.454	×	Real	CSV
Bosch Production Line Performance [65] *Anonymized process data of production lines with and without defects.*	2016	Signal	4264	C (2)	2.368.435	✓†	Real	CSV
WM811K Wafer Maps [66] *Defect matrices of semiconductor wafers with various defect types.*	2014	2D Defect Matrix	Varying	C (9)	811.457	×	Real	MAT
NEU Surface Defect Database [67] *Grayscale images of metal surfaces with various defect types and corresponding bounding box annotations.*	2013	Image	200×200	C (6)	1.800	×	Real	BMP, XML
Steel Plate Faults [68] *Geometric measurements of steel plates with various defect types.*	2010	Signal	27	C (7)	1.941	×	Real	CSV
HCI Industrial Optical Inspection [69] *Synthetic grayscale images of textured surfaces with corresponding defect ellipses.*	2007	Image	512×512	C (2)	16.100	✓	Syn.	PNG, Other

In practice, the application of predictive quality requires the existence of sufficient quality data to find the dependencies between the generally more accessible process data on the basis of which the quality of the product shall be predicted in the future. This can be difficult especially in terms of low volume discrete production systems [70]. Typical applications of predictive quality can be seen in the prediction of the surface quality, surface roughness as well as deformations or chatter marks [71]. In this paper, quality inspection entails the assessment of the quality of a manufactured product at certain stages of the manufacturing process. In a recent review paper [72] the authors conducted a thorough investigation based on the last three decades of the state of the art in so called zero defect manufacturing. The authors subdivide quality inspection based on the respective manufacturing stage into three different phases, i.e. prior to, during or after the manufacturing of the product. In terms of this study, we summarize all three aforementioned phases under the term quality inspection. ML Methods such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are used in this field for signal- and image processing with the goal of assessing the quality of the manufactured parts [5]. The corresponding datasets are exhibited in Table 3 on the previous page.

2.2.3 Process parameter optimization

Process parameters are generally chosen based on human judgement and experience in combination with the use of handbooks that provide recommendations, which may lead to a loss of productivity and quality [73]. Consequently, the selection of the optimal process parameters such as cutting speed, depth of cut, etc. plays an important role in today's highly competitive manufacturing industries and provides the opportunity to achieve high quality products with less cost and time constraints [74]. The field of application for process parameter optimization with the help of ML has received a lot of interest in recent research. In [75] the authors provide an extensive review for the application of ML for the optimization of process parameters. The main areas mentioned include milling, turning, gear hobbing and boring, finishing, welding and plastic injection molding. Next to supervised ML methods such as ANN or SVM, evolutionary optimization techniques such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO) or Simulated Annealing (SA) have been used for process parameter optimization [76]. The corresponding datasets are exhibited in Table 4.

Table 4: Datasets for process parameter optimization.

Name	Year	Features		Instances	Official Train/Test Split	Data Source	Format
		Type	Count				
Laser Welding [77] *Process parameter recordings for correlation with weld quality indicators such as weld depth and geometrical dimensions.*	2020	Signal	13	361	×	Real	XLS
3D Printer [78] *Process parameters of a 3D printer for correlation with print quality indicators such as roughness, tension and elongation.*	2018	Signal	12	50	×	Real	CSV
Tool Path Generation [79] *Shape deviation measurements and corresponding simulated cutting conditions.*	2018	Signal	9	4.968	×	Real, Syn.	CSV
Mercedes-Benz Greener Manufacturing [80] *Car feature configurations to be correlated with the required test time of the configurations.*	2017	Signal	378	8.420	✓†	Real	CSV
SECOM [81] *Semiconductor process measurements and corresponding yields for determination of key factors to yield.*	2008	Signal	591	1.567	×	Real	Other

3. Evaluation metrics

Besides data selection, another integral part of the model development process in ML is the performance evaluation. The aim of the performance evaluation is to find a model that best represents the underlying data and also performs well on new data [1]. For this purpose, a model is evaluated on a separate held-out test set using appropriate performance metrics after the training process. The selection of an appropriate evaluation metric requires a deep understanding of the pursued task with all its characteristics [82]. Despite the many discussions in the field of performance metrics in science, misleading or inadequate metrics are often used [83]. The majority of the identified datasets in this paper are suitable for classification as well as regression tasks. Both types require task-specific metrics to evaluate the respective performance of the models. Due to the large number of available metrics and their susceptibility to changing framework conditions (e.g. imbalanced classes), the selection of classification metrics often turns out to be difficult. Haixiang et al. evaluated 517 papers concerned with imbalanced classification across multiple domains and found out that 201 out of those (38%) were using accuracy as an evaluation metric [84]. In contrast, for regression, the relation and appropriateness of several evaluation metrics have been analysed thoroughly [85,86] and the difference between existing metrics is sufficiently clear. Moreover through the continuous character of the output (and measures), the selection of metrics is facilitated [14]. As a result, only classification metrics are further elaborated and critically discussed in the context of manufacturing in the following. Subsequently, the distribution of these is evaluated by analysing papers related to the datasets found.

3.1 Classification metrics

A common method for evaluating the performance of classifiers is the confusion matrix. It is applicable for problems where the output includes two or more classes. In the confusion matrix for binary classification problems, the classes are called positive and negative while the labels true and false indicate whether a prediction matches the true value or not. Most of the classification performance metrics can be derived directly (e.g. sensitivity, precision) or indirectly (e.g. Receiver Operating Characteristic (ROC), Precision-Recall Curve (PRC)) from the confusion matrix [83]. Accuracy describes the portion of correctly predicted data points out of all data points. While it is often used as a single metric to evaluate classification problems, the pure focus on maximizing accuracy is viewed critically by some researchers [83,87]. The reason for this is that classification accuracy considers the same misclassification costs for false positive and false negative errors. For most real-world problems one type of classification error (i.e. type I, type II) is more expensive than another. This issue is especially important when dealing with imbalanced datasets which frequently appear in manufacturing use cases. Suppose a model predicts NOK parts (positive class) at a quality gate which represents a problem with two classes: class A (OK parts) is 95% of the dataset and class B (NOK parts) is the remaining 5%. By simply predicting class A for every sample, the model can reach an accuracy of 95%, which seems to be a good score, but it is not. To overcome this, Seliya et al. point out that a classifier should be evaluated not only by one, but a set of performance metrics. Through this approach, several performance aspects can be considered and differentiated conclusions can be drawn [88,89]. In the given example of quality control, the correct prediction of the minority class may be of higher importance since a faulty delivery to the customer is to be avoided at all costs which promotes the use of recall as the primary evalution metric. Though, precision cannot be ignored as a low precision may lead to high quality control costs due to a high number of tests. This highlights, that the selection of a relevant metric is highly dependend on the actual use case. A metric that takes both recall and precision into account is the F-score. It uses the harmonic mean in place of the arithmetic mean, thus punishing the extreme values more [82]. A special case of this metric is the F_β-score, that allows the user to emphasize on either recall or precision [90]. Neither of the above mentioned metrics take into account the number of true negatives [91]. Specificity is used to determine the proportion of actual negative cases which got predicted correctly.

All metrics mentioned so far are single-threshold metrics, which means that they are defined for an individual score threshold (cut-off) of a classifier and cannot give an overview of the different performance levels at varying thresholds [90]. Through performance curves, the changing metrics at varying thresholds can be captured [90]. The most widespread curves are the ROC and the PRC. The ROC plot shows the trade-off between recall and specificity at varying thresholds [92] and an operating point, i.e. threshold, needs to be chosen according to the use case requirements. A single performance metric that can be derived from the ROC curve is the Area Under the ROC Curve (AUROC) score. An AUROC of 0.5 results from random choice while an AUROC of 1.0 shows a perfect classifier [93]. As with the ROC curve, the Area Under the PRC (AUPRC) is also used as a single metric. Differently to AUROC though, the baseline of AUPRC changes with class imbalance [90].

3.2 Use of classification metrics in publications

After presenting and discussing the state of the art in terms of classification metrics and the associated difficulties, the analysis of 49 different publications dealing with classification algorithms on the identified datasets is explained below. The selection is based on a backward search starting from the datasets found. Similar to the findings of Haixiang et al., accuracy is the most widely used metric in classification tasks. Almost 72% (35) of all analysed publications use accuracy as a performance metric, while for 39% of the publications, accuracy was the only metric used (see Figure 2). Furthermore, 45% of the publications only use one metric to evaluate their results. Although several authors highlight the widespread use of performance curves as an evaluation metric, this could not be fully confirmed in the analysis conducted. Altogether only eleven publications either used ROC, AUROC, PRC or AUPRC to evaluate their results. It should be noted that none of these used both ROC (AUROC) or PRC (AUPRC) and thus could not encompass all performance aspects. Only about 25% (12) of the publications studied used three or more metrics for evaluation.

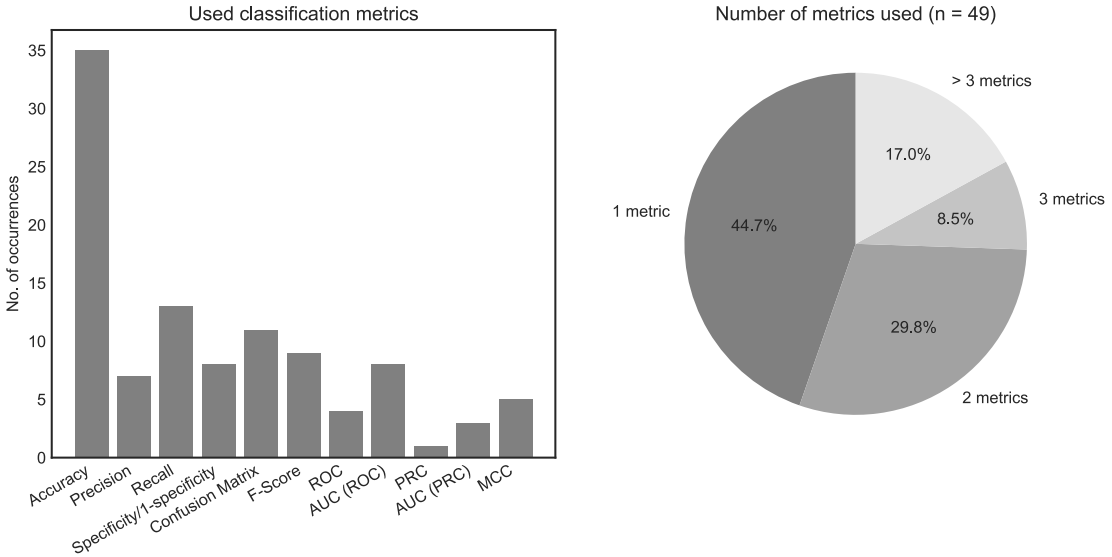

Figure 2: Analysis of publications regarding evaluation metrics (n = 49)

4. Conclusion

In this paper we provide a comprehensive overview and comparison of datasets suitable for the development of ML applications in the manufacturing sector as well as corresponding metrics for effective performance evaluation of classification problems. The identified use cases include predictive maintenance, condition monitoring, process monitoring, predictive quality, quality inspection and process parameter optimization. The analysis has the aim of stimulating research in this field as well as to promote the use of public datasets

for evaluation. This is required to compare the performance of different approaches objectively in research, which is often not possible due to the use of proprietary datasets that are not shared because of data privacy concerns. Further, manufacturing companies can employ the public datasets in the development of algorithms for their specific facilities and gain practical knowledge and experience in the process [4]. Additionally, we showed that a large part of the analysed studies solely use accuracy for performance evaluation of classification problems, which may not be expressive enough in all use cases. As an opportunity for future work, it would be of interest to identify or create datasets, corresponding tasks and metrics that can serve as a standard benchmark for certain use cases in manufacturing, comparable to other industry areas such as autonomous driving.

Acknowledgements

The research leading to these results has received funding from the European Institute of Technology (EIT Manufacturing) under grant agreement number 21026 (IVE), as well as from the German Federal Minisitry of Education and Research under grant agreement number 02L19C150 (KompAKI).

References

[1] Wuest, T., Weimer, D., Irgens, C., Thoben, K.-D., 2016. Machine learning in manufacturing: advantages, challenges, and applications. Production & Manufacturing Research 4 (1), 23–45.

[2] Diez-Olivan, A., Del Ser, J., Galar, D., Sierra, B., 2019. Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. Information Fusion 50, 92–111. https://www.sciencedirect.com/science/article/pii/S1566253518304706.

[3] Pham, D.T., Afify, A.A., 2005. Machine-learning techniques and their applications in manufacturing. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture 219 (5), 395–412.

[4] Krauß, J., Dorißen, J., Mende, H., Frye, M., Schmitt, R.H., 2019. Machine learning and artificial intelligence in production: application areas and publicly available data sets. Production at the leading edge of technology, 493–501.

[5] Bacioiu, D., Melton, G., Papaelias, M., Shaw, R., 2019. Automated defect classification of Aluminium 5083 TIG welding using HDR camera and neural networks. Journal of Manufacturing Processes 45, 603–613. https://www.kaggle.com/danielbacioiu/tig-aluminium-5083. Accessed 31 March 2021.

[6] Kantesaria, N., Vaghasia, P., Hirpara, J., Bhoraniya, R., 2020. Casting product image data for quality inspection. https://www.kaggle.com/ravirajsinh45/real-life-industrial-dataset-of-casting-product. Accessed 19 March 2021.

[7] Paladi, S., 2020. Mechanic Component Images. https://www.kaggle.com/satishpaladi11/mechanic-component-images-normal-defected. Accessed 19 March 2021.

[8] GitHub. https://github.com/. Accessed 31 March 2021.

[9] Kaggle. https://www.kaggle.com/. Accessed 31 March 2021.

[10] Mendeley Data. https://data.mendeley.com/. Accessed 31 March 2021.

[11] NASA Ames Prognostics Center of Excellence (PCoE);. https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/. Accessed 27 November 2020.

[12] OpenML. https://www.openml.org/home. Accessed 31 March 2021.

[13] University of California Irvine. UCI Machine Learning Repository. University of California Irvine. https://archive.ics.uci.edu. Accessed 3 November 2020.

[14] Ferri, C., Hernández-Orallo, J., Modroiu, R., 2009. An experimental comparison of performance measures for classification. Pattern Recognition Letters 30 (1), 27–38.

[15] Bogaschewsky, R., Rollberg, R., 1998. Prozeßorientiertes Management. Springer, Berlin, Heidelberg.

[16] Kang, Z., Catal, C., Tekinerdogan, B., 2020. Machine learning applications in production lines: A systematic literature review. Computers & Industrial Engineering 149, 106773.

[17] Goyal, D., Saini, A., Dhami, S.S., Pabla, B.S., 2016. Intelligent predictive maintenance of dynamic systems using condition monitoring and signal processing techniques — A review. 2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Spring), 1–6.

[18] Wang, H., Ye, X., Yin, M., 2016. Study on predictive maintenance strategy. International Journal of u-and e-Service, Science and Technology (9), 295–300.

[19] Hu, J., Chen, P., 2020. Predictive maintenance of systems subject to hard failure based on proportional hazards model. Reliability Engineering & System Safety 196.

[20] Mobley, R.K., 2002. An introduction to predictive maintenance, 2nd ed. Butterworth-Heinemann, Amsterdam, New York, 438 pp.

[21] Pestana, D., 2020. Diesel engine faults features dataset: (3500-DEFault). https://data.mendeley.com/datasets/k22zxz29kr/1. Accessed 20 March 2021.

[22] Birgelen, A.v., Buratti, D., Mager, J., Niggemann, O., 2018. Self-organizing maps for anomaly localization and predictive maintenance in cyber-physical production systems. Procedia CIRP 72, 480–485. https://www.kaggle.com/inIT-OWL/one-year-industrial-component-degradation; https://www.kaggle.com/inIT-OWL/production-plant-data-for-condition-monitoring. Accessed 31 March 2021.

[23] Kovalenko, I., Saez, M., Barton, K., Tilbury, D., 2017. SMART: A system-level manufacturing and automation research testbed. Smart Sustain. Manuf. Syst. 1 (1), 20170006. https://www.kaggle.com/shasun/tool-wear-detection-in-cnc-mill. Accessed 31 March 2021.

[24] Helwig, N., Pignanelli, E., Schutze, A., 2015. Condition monitoring of a complex hydraulic system using multivariate statistics. IEEE International Instrumentation and Measurement Technology Conference, 210–215. https://archive.ics.uci.edu/ml/datasets/Condition+monitoring+of+hydraulic+systems. Accessed 31 March 2021.

[25] Institut für industrielle Informationstechnik - inIT, Technische Hochschule Ostwestfalen-Lippe, 2018. Versatile Production System. https://www.kaggle.com/inIT-OWL/versatileproductionsystem. Accessed 20 March 2021.

[26] Qiao, H., 2017. Degradation measurement of robot arm position accuracy. National Institute of Technology (NIST). https://www.nist.gov/el/intelligent-systems-division-73500/degradation-measurement-robot-arm-position-accuracy. Accessed 19 March 2021.

[27] Lindgren, T., Biteus, J., 2016. APS failure at Scania trucks data set. Scania CV AB. https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks. Accessed 20 March 2021.

[28] Coraddu, A., Oneto, L., Ghio, A., Savio, S., Anguita, D., Figari, M., 2016. Machine learning approaches for improving condition-based maintenance of naval propulsion plants. Proceedings of the IMechE 230 (1), 136–153. http://archive.ics.uci.edu/ml/datasets/Condition+Based+Maintenance+of+Naval+Propulsion+Plants. Accessed 31 March 2021.

[29] PHM Society. 2015 PHM Society Conference Data Challenge. PHM Society. https://github.com/robot007/PHM15. Accessed 31 March 2021.

[30] PHM Society, 2014. 2014 PHM Society Conference Data Challenge. PHM Society. https://phmsociety.org/conference/annual-conference-of-the-phm-society/annual-conference-of-the-prognostics-and-health-management-society-2014/phm-data-challenge-2/. Accessed 18 March 2021.

[31] PHM Society, 2013. 2013 PHM Society Conference Data Challenge. PHM Society. https://phmsociety.org/conference/annual-conference-of-the-phm-society/annual-conference-of-the-prognostics-and-health-management-society-2013/phm-data-challenge/. Accessed 31 March 2021.

[32] PHM Society, 2011. 2011 PHM Society Conference Data Challenge. PHM Society. https://phmsociety.org/phm_competition/2011-phm-society-conference-data-challenge/. Accessed 18 March 2021.

[33] Goebel, K., 2009. Gearbox Fault Detection Dataset, PHM Data Challenge 2009. https://c3.nasa.gov/dashlink/resources/997/. Accessed 31 March 2021.

[34] Mcintosh, D., 2010. Li-ion battery aging datasets. NASA Ames Prognostics Center of Excellence (PCoE). https://c3.nasa.gov/dashlink/resources/133/. Accessed 20 March 2021.

[35] Saxena, A., Goebel, K., 2008. Turbofan engine degradation simulation data set. NASA Ames Prognostics Center of Excellence (PCoE). https://c3.nasa.gov/dashlink/resources/139/. Accessed 20 March 2021.

[36] Lee, J and Qiu, H and Yu, G and Lin, Ja and others, 2007. Bearing data set. IMS, University of Cincinnati, NASA Ames Prognostics Center of Excellence (PCoE), Rexnord Technical Services. https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/. Accessed 20 March 2021.

[37] Agogino, A., Goebel, K., 2007. Milling data set. NASA Ames Prognostics Data Repository; BEST lab, UC Berkeley. https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/. Accessed 20 March 2021.

[38] Case Western Reserve University. CWRU Bearing Data Center. Case Western Reserve University. https://csegroups.case.edu/bearingdatacenter. Accessed 19 March 2021.

[39] D. E. Lee, Inkil Hwang, C. M. O. Valente, J. F. G. Oliveira, David A. Dornfeld, 2006. Precision manufacturing process monitoring with acoustic emission. Condition monitoring and control for intelligent manufacturing, 33–54.

[40] William H. Woodall, Douglas C. Montgomery, 2014. Some current directions in the theory and application of statistical process monitoring. Journal of Quality Technology 46 (1), 78–94.

[41] Bersimis, S., Panaretos, J., Psarakis, S., 2005. Multivariate statistical process control charts and the problem of interpretation: a short overview and some applications in industry, in: Proceedings of the 7th Hellenic European Conference on Computer Mathematics and its Applications, Athens Greece.

[42] Tang, P., Peng, K., Dong, J., Zhang, K., Zhao, S., 2020. Monitoring of nonlinear processes with multiple operating modes through a novel gaussian mixture variational autoencoder model. IEEE Access 8, 114487–114500.

[43] Qin, S.J., 2003. Statistical process monitoring: basics and beyond. J. Chemometrics 17 (8-9), 480–502.

[44] Yu, J., Liu, X., Ye, L., 2021. Convolutional long short-term memory autoencoder-based feature learning for fault detection in industrial processes. IEEE Trans. Instrum. Meas. 70, 1–15.

[45] Maggioni, M., Marzorati, E., Grasso, M., Colosimo, B.M., Parenti, P., 2014. In-process quality characterization of grinding processes: A sensor-fusion based approach. ASME 2014 12th Biennial Conference.

[46] Lv, F., Fan, X., Wen, C., Bao, Z., 2018. Stacked sparse auto encoder network based multimode process monitoring. 2018 International Conference on Control, Automation and Information Sciences (Iccais), 227–232.

[47] Yan, S., Yan, X., 2020. Quality-driven autoencoder for nonlinear quality-related and process-related fault detection based on least-squares regularization and enhanced statistics. Ind. Eng. Chem. Res. 59 (26), 12136–12143.

[48] Hranisavljevic, N., Niggemann, O., Maier, A. A novel anomaly detection algorithm for hybrid production systems based on deep learning and timed automata. arXiv preprint arXiv:2010.15415 2020. https://www.kaggle.com/inIT-OWL/high-storage-system-data-for-energy-optimization. Accessed 31 March 2021.

[49] Birgelen, A.v., Niggemann, O., 2018. Anomaly detection and localization for cyber-physical production systems with self-organizing maps. IMPROVE-Innovative Modelling Approaches for Production Systems to Raise

Validatable Efficiency, 55–71. https://www.kaggle.com/inIT-OWL/genesis-demonstrator-data-for-machine-learning/home. Accessed 31 March 2021.

[50] Rieth, C.A., Amsel, B.D., Tran, R., Cook, M.B., 2017. Additional tennessee eastman process simulation data for anomaly detection evaluation. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6C3JR1. Accessed 20 March 2021.

[51] Camarinha-Matos, L.M., Lopes, L.S., Barata, J., 1996. Integration and learning in supervision of flexible assembly systems. IEEE Trans. Robot. Automat. 12 (2), 202–219. http://archive.ics.uci.edu/ml/datasets/Robot+Execution+Failures. Accessed 31 March 2021.

[52] Bergadano, F., Giordana, A., Saitta, L., Marchi, D. de, Brancadori, F., 1990. Integrated learning in a real domain. Machine Learning Proceedings, 322–329. http://archive.ics.uci.edu/ml/datasets/Mechanical+Analysis. Accessed 31 March 2021.

[53] Xu, P., Ren, Z., Shen, Y., Yu, W., He, L., 2021. Quality prediction of discrete manufacturing process based on CGAN&Catboost hybrid model. J. Phys.: Conf. Ser. 1757 (1), 12072.

[54] Jimmy Chhor, Stefan Gerdhenrichs, Felix Mohrschladt, Robert H. Schmitt, 2019. Development of a machine learning model for a multi-correlative sample-based prediction of product quality for complex machining processes, in: , Production at the leading edge of technology. Springer Vieweg, Berlin, Heidelberg, pp. 523–532.

[55] Krauß, J., Pacheco, B.M., Zang, H.M., Schmitt, R.H., 2020. Automated machine learning for predictive quality in production. Procedia CIRP 93, 443–448. https://www.sciencedirect.com/science/article/pii/s2212827120306016.

[56] Lv, X., Duan, F., Fu, X., Gan, L., 2020. Deep metallic surface defect detection: The new benchmark and detection network. Sensors. https://www.kaggle.com/zhangyunsheng/defects-class-and-location.

[57] Liveline Technologies, 2020. Multi-stage continuous-flow manufacturing process. Liveline Technologies. https://www.kaggle.com/supergus/multistage-continuousflow-manufacturing-process/metadata. Accessed 19 March 2021.

[58] Kaggle, 2020. Find a defect in the production extrusion line. https://www.kaggle.com/podsyp/find-a-defect-in-the-production-extrusion-line/metadata. Accessed 20 March 2021.

[59] Silvestre-Blanes, J., Albero-Albero, T., Miralles, I., Pérez-Llorens, R., Moreno, J., 2019. A public fabric database for defect detection methods and results. Autex Research Journal (4), 363–374. https://www.aitex.es/afid/. Accessed 31 March 2021.

[60] Tang, S., He, F., Huang, X., Yang, J. Online PCB defect detector on a new PCB defect dataset. https://github.com/Charmve/Surface-Defect-Detection/tree/master/DeepPCB. Accessed 31 March 2021.

[61] Severstal, 2019. Steel defect detection. Severstal. https://www.kaggle.com/c/severstal-steel-defect-detection/overview. Accessed 19 March 2021.

[62] Yesilli, M., 2019. Turning dataset for chatter diagnosis using machine learning. Mendeley Data. http://dx.doi.org/10.17632/hvm4wh3jzx.1.

[63] Huang, Y., Qiu, C., Guo, Y., Wang, X., Yuan, K., 2020. Surface defect saliency of magnetic tile. The Visual Computer (36), 85--96. https://github.com/abin24/Magnetic-tile-defect-datasets. Accessed 31 March 2021.

[64] Magalhães Oliveira, E., 2017. Quality prediction in a mining process. https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process. Accessed 19 March 2021.

[65] Robert Bosch GmbH, 2016. Bosch production line performance. Robert Bosch GmbH. https://www.kaggle.com/c/bosch-production-line-performance/overview. Accessed 19 March 2021.

[66] National Taiwan University, CS Dept. WM811K: Wafer Map. National Taiwan University, CS Dept. http://mirlab.org/dataSet/public/. Accessed 19 March 2021.

[67] Song, K., Yan, Y., 2013. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Applied Surface Science 285, 858–864. http://faculty.neu.edu.cn/yunhyan/NEU_surface_defect_database.html. Accessed 31 March 2021.

[68] Semeion, Research Center of Sciences of Communication, Via Sersale 117, 00128, Rome, Italy. Steel Plate Faults Dataset. https://archive.ics.uci.edu/ml/datasets/steel+plates+faults. Accessed 11 May 2021.

[69] Wieler, M., Hahn, T., 2007. Weakly supervised learning for industrial optical inspection. DAGM. https://hci.iwr.uni-heidelberg.de/content/weakly-supervised-learning-industrial-optical-inspection. Accessed 19 March 2021.

[70] Gittler, T., Relea, E., Corti, D., Corani, G., Weiss, L., Cannizzaro, D., Wegener, K., 2019. Towards predictive quality management in assembly systems with low quality low quantity data – a methodological approach. Procedia CIRP 79, 125–130.

[71] Kim, D.-H., Kim, T.J.Y., Wang, X., Kim, M., Quan, Y.-J., Oh, J.W., Min, S.-H., Kim, H., Bhandari, B., Yang, I., Ahn, S.-H., 2018. Smart machining process using machine learning: A review and perspective on machining industry. Int. J. of Precis. Eng. and Manuf.-Green Tech. 5 (4), 555–568.

[72] Psarommatis, F., May, G., Dreyfus, P.-A., Kiritsis, D., 2020. Zero defect manufacturing: state-of-the-art review, shortcomings and future directions in research. International journal of production research 58 (1), 1–17.

[73] Kant, G., Sangwan, K.S., 2015. Predictive modelling and optimization of machining parameters to minimize surface roughness using artificial neural network coupled with genetic algorithm. Procedia CIRP 31, 453–458.

[74] Bouacha, K., Terrab, A., 2016. Hard turning behavior improvement using NSGA-II and PSO-NN hybrid model. Int J Adv Manuf Technol 86 (9-12), 3527–3546.

[75] Weichert, D., Link, P., Stoll, A., Rüping, S., Ihlenfeldt, S., Wrobel, S., 2019. A review of machine learning for the optimization of production processes. Int J Adv Manuf Technol 104 (5), 1889–1902.

[76] Yusup, N., Zain, A.M., Hashim, S.Z.M., 2012. Evolutionary techniques in optimizing machining parameters: Review and recent applications (2007–2011). Expert Systems with Applications 39 (10), 9909–9927.

[77] Rinne, J., 2020. Screening datasets for laser welded steel-copper lap joints. Mendeley Data. http://dx.doi.org/10.17632/2s5m3crbkd.2.

[78] Okudan, A., 2018. 3D printer dataset for mechanical engineers. TR/Selcuk University. https://www.kaggle.com/afumetto/3dprinter. Accessed 19 March 2021.

[79] Dittrich, M.-A., Uhlich, F., Denkena, B., 2019. Self-optimizing tool path generation for 5-axis machining processes. CIRP journal of manufacturing science and technology 24, 49–54. https://data.mendeley.com/datasets/smyg6cfwpk/1. Accessed 31 March 2021.

[80] Daimler AG, 2017. Mercedes-Benz greener manufacturing. https://www.kaggle.com/c/mercedes-benz-greener-manufacturing/data. Accessed 19 March 2021.

[81] McCann, M., Li, Y., Maguire, L., Johnston, A., 2008. Causality challenge: Benchmarking relevant signal components for effective monitoring and process control. JMLR: Workshop and Conference Proceedings 6c, 277–288. http://archive.ics.uci.edu/ml/datasets/SECOM. Accessed 19 March 2021.

[82] Brabec, J., Machlica, L., 2018. Bad practices in evaluation methodology relevant to class-imbalanced problems, 4 pp. http://arxiv.org/pdf/1812.01388v1.

[83] Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. Beyond Accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. Advances in Artificial Intelligence, 1015–1021.

[84] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 73, 220–239.

[85] Harrell, F.E., 2015. Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis, Second edition ed. Springer, Cham, Heidelberg, New York, 582 pp.

[86] Bishop, C.M., 1995. Neural networks for pattern recognition. Oxford University Press; Clarendon Press, Oxford, 482 pp.

[87] Provost, F., Fawcett, T., Kohavi, R., 1998. The case against Accuracy estimation for comparing induction algorithms. Proceedings of the 15th international conference on machine learning ICML.

[88] Seliya, N., Khoshgoftaar, T.M., van Hulse, J., 2009. A study on the relationships of classifier performance metrics, in: , 2009 21st IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2009). IEEE, pp. 59–66.

[89] Vishwakarma, G., Sonpal, A., Hachmann, J., 2021. Metrics for benchmarking and uncertainty quantification: Quality, applicability, and best practices for machine learning in chemistry. Trends in Chemistry 3 (2), 146–156.

[90] Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one 10 (3).

[91] Powers, D.M.W., 2007. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001.

[92] Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861–874.

[93] Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.

Biography

Nicolas Jourdan, M. Sc. (*1992) is a research assistant and PhD student at the Institute of Production Management, Technology and Machine Tools (PTW) at the Technical University of Darmstadt, Germany since 2020. His research interests include the robustness analysis of machine learning models for manufacturing applications.

Lukas Longard, M. Sc. (*1992) is a research assistant and PhD student at the Institute of Production Management, Technology and Machine Tools (PTW) at at the Technical University of Darmstadt, Germany since 2019. His research interests include the further development and introduction of machine learning to digital shop floor management - a management tool in the production area.

Tobias Biegel, M. Sc. (*1993) is a research assistant and PhD student at the Institute of Production Management, Technology and Machine Tools (PTW) at the Technical University of Darmstadt, Germany since 2019. His research interests include the application of Deep Learning for in-process multivariate statistical process control (MSPC) in manufacturing processes.

Prof. Dr.-Ing. Joachim Metternich (*1968) has been the head of the Institute for Production Management, Technology and Machine Tools (PTW) at the Technical University of Darmstadt, Germany since 2012. In addition, Prof. Metternich is the spokesman for the Center of Excellence for Medium-Sized Businesses in Darmstadt and president of the International Association of Learning Factories.