Metadata Modeling and Use of Domain Knowledge to Support Industrial Data Analytics

Peter Reimann[1]

[1]Graduate School of Excellence advanced Manufacturing Engineering, University of Stuttgart, Nobelstr. 12, Stuttgart, Germany

Abstract

Industrial Data Analytics refers to data analyses across different phases of the industrial product life cycle. The specific characteristics of available industrial data often pose challenges for common data management and data analysis methods. This paper gives an overview on the projects of the research group ICT Platform for Manufacturing at the Graduate School of Excellence advanced Manufacturing Engineering (GSaME) of the University of Stuttgart. These projects are related to Industrial Data Analytics and offer approaches to addressing the domain-specific data characteristics. Relevant research areas are metadata management, use of domain knowledge to improve data preparation, the management of machine learning (ML) models, and approaches to meta-learning and automated machine learning (AutoML). In addition, this paper details on two specific research contributions. Firstly, it discusses a metadata model that facilitates a democratized access to data in virtual product development projects. The second contribution is an approach to exploit domain knowledge during data preparation in order to address two of the most important challenging data characteristics in industrial data: a multi-class imbalance and a data bias that is due the high variety of underlying products.

Keywords

Industrial Data Analytics, Domain-specific Data Characteristics, Metadata Modeling, Domain Knowledge

1. Introduction

Industrial Data Analytics refers to problems and solution approaches to data management, data provision, and data analytics across different phases of the industrial product life cycle [1]. This paper gives an overview on the research topics of the research group ICT Platform for Manufacturing at the Graduate School of Excellence advanced Manufacturing Engineering (GSaME) of the University of Stuttgart. This research group deals with both application-oriented and fundamental research in the area of Industrial Data Analytics. It examines data and their potential for data analysis in various phases of a product life cycle, e.g., for analyzing simulation data in the product development phase [2], sensor data from test benches in the production phase [3, 4], or data from the product usage phase describing the configurations of sold products [5].

The specific characteristics of available industrial data pose challenges for common data management and data analysis methods [6, 7, 8, 9]. For instance, data may come in diverse and heterogeneous formats and be contained in isolated data silos across different organizational units of a company. This makes it difficult or even impossible to acquire relevant data for a particular analysis [2]. Moreover, the high product diversity increases the number and complexity of patterns and correlations contained in data [10, 4]. Here, machine learning algorithms often fail to correctly identify all these patterns and correlations.

The specific characteristics of industrial data and the resulting challenges for data management and data analysis methods often lead to fundamental research questions that have not yet been considered in scientific literature. The research group contributes to answering these questions by developing novel approaches that are tailored to the specific characteristics of industrial data. Here, relevant research topics include how to address different kinds of data bias and data set shifts in real-world industrial data [11, 5] or how to improve data quality in unstructured text data and text analysis pipelines [12]. Other areas of research are related to metadata management [13], use of domain knowledge to improve data preparation and data analysis [14, 15], the management of machine learning (ML) models [16], as well as approaches to meta-learning [17] and automated machine learning (AutoML) [18, 19].

After giving an overview on related research projects of the group ICT Platform for Manufacturing in Section 2, this paper details on two specific contributions in Sections 3 and 4. The first contribution is a metadata model that connects data from heterogeneous and previously isolated data sources in virtual product development [13]. The second major contribution is an approach to exploit domain knowledge from a taxonomy during data preparation in order to address two of the most important challenging data characteristics and kinds of bias in industrial data: a multi-class imbalance and a data bias that is due the high variety of underlying products [10, 14].

34th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), June 7-9, 2023, Hirsau, Germany

peter.reimann@gsame.uni-stuttgart.de (P. Reimann)
0000-0002-6355-4591 (P. Reimann)

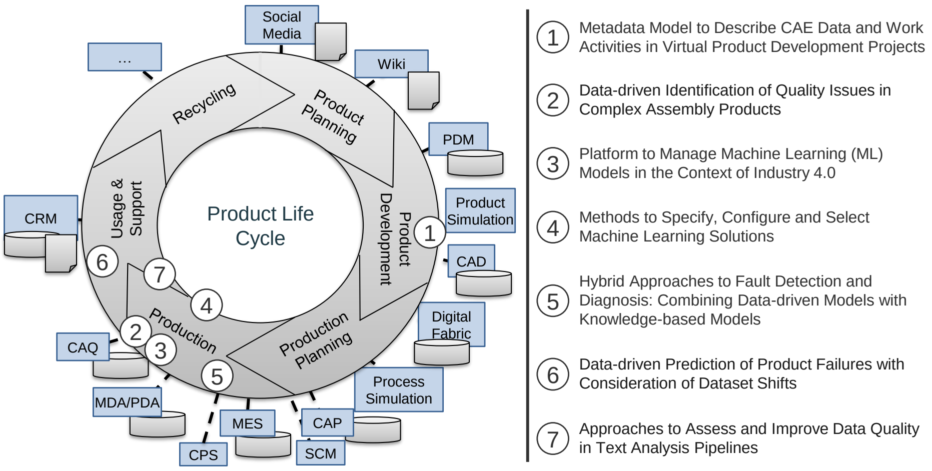
Product Life Cycle
Recycling
Product Planning
Product Development
Production Planning
Production
Usage & Support
Social Media
Wiki
PDM
Product Simulation
CAD
Digital Fabric
Process Simulation
CAP
SCM
MES
CPS
MDA/PDA
CAQ
CRM
...
1
2
3
4
5
6
7
1 Metadata Model to Describe CAE Data and Work Activities in Virtual Product Development Projects
2 Data-driven Identification of Quality Issues in Complex Assembly Products
3 Platform to Manage Machine Learning (ML) Models in the Context of Industry 4.0
4 Methods to Specify, Configure and Select Machine Learning Solutions
5 Hybrid Approaches to Fault Detection and Diagnosis: Combining Data-driven Models with Knowledge-based Models
6 Data-driven Prediction of Product Failures with Consideration of Dataset Shifts
7 Approaches to Assess and Improve Data Quality in Text Analysis Pipelines

Figure 1: Overview on research projects of the research group ICT Platform for Manufacturing and their association to individual phases of a typical product life cycle.

2. Research Group ICT Platform for Manufacturing

Figure 1 gives an overview on the projects and research topics related to Industrial Data Analytics in the research group ICT Platform for Manufacturing. Furthermore, the figure assigns the projects to the phases of a typical product life cycle, from which the projects mainly acquire and analyze data. The first project deals with metadata modeling in the area of virtual product development [2]. The second project focuses on the production phase and how to identify quality issues in complex assembly products, e.g., truck engines [3]. The major contributions of these two projects are discussed in Sections 3 and 4.

The major outcome of the third project is a platform to manage machine learning (ML) models, e.g., classification or regression models [16]. Data scientists of a company may develop ML models for specific use cases and upload them in the platform. The ML models are then associated with appropriate metadata to further describe them. This metadata covers, amongst others, life cycle information of the ML model, e.g., whether it is in the training or application phase or whether it has already been retired and is thus not used anymore [20]. Moreover, the metadata covers semantic information [16]. This for instance includes information about the domain-specific use case, e.g., fault detection or fault diagnosis, or about the machine in a production line for which the ML model has been developed. It helps other stakeholders find appropriate models for their specific use cases and thus effectively facilitates reusability of ML models.

The contributions of the fourth project offer structured methods to specify, configure and select whole ML solutions. These ML solutions constitute combinations and configurations of ML tools and software to perform, e.g., data collection, data preprocessing, model training, and model deployment [21]. One contribution of this project is a method to structure the collabaration among different stakeholdes, e.g., data scientists, IT experts, business analysts, engineers or other domain experts during development projects for ML solutions [22]. In addition, AssistML is a novel concept that enhances the ML model management platform of the previous project by approaches to meta-learning [17]. This way, AssistML automates the discovery and recommendation of ML solutions for a given use case, and it makes this task feasible for non-experts such as citizen data scientists.

The fifth project deals with hybrid approaches and their application to fault detection and fault diagnosis in a production line [15]. Hybrid approaches combine data-driven models with knowledge-based models, such as physics-based simulation models. These different kinds of data-driven and knolwedge-driven models enhance and complement each other. This is particularly useful in cases when pure data-driven solutions are not adequate due to a lack of data. A major contribution of this project is *PUSION*, a generic and automated framework for decision fusion in classification ensembles that may be composed of diverse data-driven or knowledge-driven models [18]. By combining the decisions of these diverse classification models via decision fusion algorithms, the overall prediction accuracy may be enhanced.

The focus of the next project is related to methods for data-driven prediction of product failures during the product usage phase [5]. Here, different kinds of data set shifts may occur, i.e., changes of the statistical data distribution over time [23]. This means that both the decision boundaries of classification patterns and the

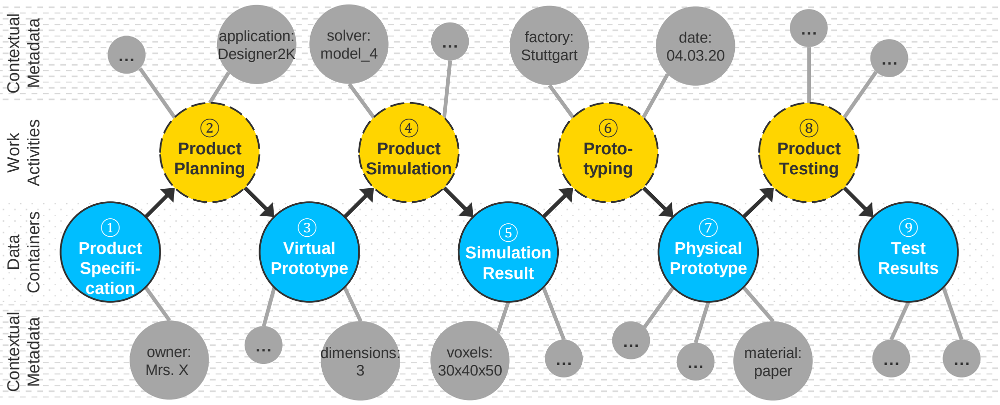

Figure 2: Example of an instance of the metadata model that describes both data, metadata, and work activities of virtual product development projects, cf. Ziegler at al. [13].

statistical distribution of these patterns may change. This has to be reflected via adequate approaches to detect data set shifts and to adapt classification models to the new statistical distribution if necessary.

Finally, one project concerns data quality of text data. It introduces the QUALM concept for continuous data quality measurement and improvement at different steps of text analysis pipelines [12]. QUALM data quality indicators quantify text characteristics, e.g., the number of abbreviations or spelling mistakes, and give hints how these may affect the quality of analysis results. Corresponding QUALM modifiers use these text characteristics to enhance the text quality. An example of such a modifier is an approach to select the best-fitting training data for an analysis task based on the similarity between this training data and the input data [24]. In addition, QUALM offers a hybrid method for information extraction, which exploits both structured and unstructured data sources to yield more relevant information from these sources [25].

3. Metadata Model for Virtual Product Development Projects

Product development projects in companies are mainly virtual and digitized thanks to several computer-aided systems, e.g., for Computer-Aided Design (CAD), Computer-Aided Engineering (CAE), and Computer-Aided Testing (CAT) [26] These CAx systems produce a huge amount of data that offer additional opportunities for data analysis, e.g., to gain insights how to improve a product design. However, several challenges hinder exploiting the full potential for data analysis [13]. In particular, different CAx systems store their data in various heterogeneous formats, e.g., proprietary 2D or 3D geometry files, plain text, images, videos, XML documents, or CSV-formated files. In addition, different data are contained in isolated data silos across individual organizational units that are often not willing to share their data with other stakeholders, even from the same company. Altogether, this makes it nearly impossible for domain experts, i.e., development engineers, to acquire and explore the data they need for a particular data analysis.

Ziegler et al. [13] thus propose a metadata model to describe all related CAx data and to address the above-mentioned challenges. This metadata model not only describes data, but it offers a connected view on data, metadata, and work activities in virtual product development projects. Figure 2 shows an idealized example of an instance of the metadata model. It covers several data containers [27, 28] (blue in the figure) that abstract from heterogeneous data formats and point to the underlying data sources or files via URIs. Furthermore, the metadata explicitly describes the work activities (yellow) that are carried out by development engineers in virtual product development projects. These work activities are connected to the data containers the activities consume and produce. So, the whole metadata describes full and connected views of project workflows including the activities and the data. The grey elements shown in Figure 2 represent metadata that further describe either data containers or work activities via contextual information.

A major benefit of this metadata model is that product development engineers easily understand it, because it is based on the project workflows and work activities these engineers carry out in their daily work life. This facilitates a democratized data access, so that product development engineers may easily find the data associated to the work activities in development projects they are familiar with. It facilities an expert-led data exploration and a subsequent data analysis to gain sophisticated insights from the underlying data. The metadata model

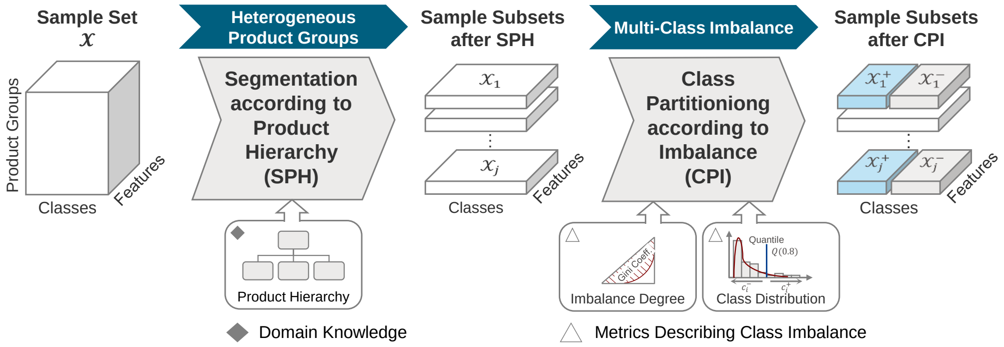

Figure 3: Major steps of the approach that uses domain knowledge and metrics about class distributions to address the challenges of heterogeneous product groups and multi-class imbalance, cf. Hirsch et al. [10].

is implemented in a data lake and as a property graph structure in a graph database system [13]. This system offers APIs to navigate through the graph structure and to issue queries, e.g., in Cypher, to search for specific data sets. It may be used to realize various kinds of data analysis, ranging from reports over data mining to even process mining [29].

4. Approach to Exploit Domain Knowledge for Data Preparation

Hirsch et al. [11] introduce a challenging use case of data-driven identification of quality issues in complex assembly products, e.g., truck engines. The use cases constitutes a multi-class classification problem, where each class corresponds to one of 84 engine components, e.g., cylinders, fuel injectors, or turbochargers. The problem is to train a classification model that is able to identify one of these multiple classes and engine components that are the cause of a particular quality issue.

With 1050 data instances, the data set of this use cases is of rather low size and contains several kinds of noise [11]. Nevertheless, literature comprises techniques for classification ensembles [30], e.g., Random Forest [31], which are able to deal with these two data characteristics. However, such data-driven methods usually still show poor prediction performance, as they are not able to address two additional kinds of domain-specific data characteristics [11]. The first one is a *multi-class imbalance*, i.e., the class labels occur in an imbalanced way in the data [32]. Here, many learning algorithms tend to ignore the patterns of class labels that are underrepresented. The second challenging data characteristic results from the fact that companies offer *heterogeneous product groups* with a high product variety. Here, the class patterns, i.e., decision boundaries in the feature space of a particular class usually differ across individual product groups. So, the number of class patterns is even higher than the number of classes, and the class patterns may be represented by different and overlapping ranges of feature values. This makes it hard for learning algorithms to identify and clearly distinguish all class patterns. Furthermore, these learning algorithms again tend to ignore the class patterns of seldom product groups that are underrepresented in data.

Hirsch et al. propose an approach to data preparation that effectively addresses both challenges arising from a multi-class imbalance and from heterogeneous product groups [10, 14]. Figure 3 shows the major steps of this approach. It divides the whole data set $\mathcal{X}$ into several subsets $\mathcal{X}_j \subseteq \mathcal{X}$. After this data preparation, a classification model is trained for each of the data subsets $\mathcal{X}_j$.

The first step, *Segementation according to Product Hierarchy* (SPH), uses domain knowledge from a product hierarchy or product taxonomy to divide the data set into one subset $\mathcal{X}_j$ for each product group. As only data of one particular product group is contained in each subset $\mathcal{X}_j$, this subset contains a significantly less number of class patterns, i.e., usually only one pattern for each remaining class in $\mathcal{X}_j$. In addition, these class patterns are more evenly distributed in each subset. So, learning algorithms have much less problems to identify and distinguish all class patterns, even those that have previously been underrepresented in the whole data set $\mathcal{X}$

The second step, *Class Partitioning according to Imbalance* (CPI), further divides some of the subsets $\mathcal{X}_j$ to address multi-class imbalance. Therefore, CPI first determines the degree of class imbalance in each subset $\mathcal{X}_j$ resulting from SPH using the Gini coefficient as an imbalance metric. If the value of the Gini coefficient as higher than a threshold, e.g., 30 %, the subset $\mathcal{X}_j$ is divided into a subset $\mathcal{X}_j^+$ containing only data instance of majority classes and a subset $\mathcal{X}_j^-$ for instances of minority classes. Here, CPI uses a quantile approach to determine the point of intersection between majority and minority classes.

Hirsch et al. [10] apply their approach to the data of the above-mentioned uses case for a data-driven identification of quality issues in assembled truck engines and discuss the evaluation results. In addition, the authors prove the generality of their approach by applying it to several synthetic data sets that show varying data and class distributions [14]. In both evaluations, they compare their approach exploiting domain knowledge with a data-driven baseline that applies Random Forest and a feature selection technique to the whole data set $\mathcal{X}$. They show that their approach leads to an average increase of classification accuracy between 4 and 13 %-points. Furthermore, it leads to a reduction of the number of rework steps needed to repair faulty truck engines.

5. Summary

This paper gives an overview on the projects of the research group ICT Platform for Manufacturing at the Graduate School of Excellence advanced Manufacturing Engineering (GSaME) of the University of Stuttgart. These projects are related to Industrial Data Analytics covering data analyses across the whole industrial product life cycle. The research topics of this group include methods to address different kinds of data bias, data set shifts, or text data quality, as well as areas such as metadata management, use of domain knowledge to improve data preparation, management of machine learning (ML) models, and approaches to meta-learning and automated machine learning (AutoML). In addition, this paper details on two specific research contributions. Firstly, it discusses a metadata model that facilitates a democratized access to data of different CAx systems in virtual product development projects. The second contribution is an approach to exploit domain knowledge during data preparation in order to address two of the most important challenging data characteristics in industrial data: a multi-class imbalance and a data bias that is due the high variety of underlying products.

Acknowledgments

This work was supported by the German Research Foundation (DFG) within the Excellence Initiative II and by the Ministry of Science, Research and Arts of the State of Baden-Wurttemberg within the sustainability support of the projects of the Excellence Initiative II. Further thanks go to the colleagues that have been part of the research group ICT Platform for Manufacturing from 2017 to 2022: Vitali Hirsch, Cornelia Kiefer, Marco Spieß, Alejandro Gabriel Villanueva Zacarias, Christian Weber, Yannick Wilhelm, and Julian Ziegler.

References

[1] C. Gröger, Industrial Analytics –- An Overview, it – Information Technology 64 (2022) 55–65. doi:10.1515/itit-2021-0066.

[2] J. Ziegler, et al., A Graph-based Approach to Manage CAE Data in a Data Lake, Procedia CIRP 93 (2020) 496–501. doi:10.1016/j.procir.2020.04.155.

[3] V. Hirsch, et al., Analytical Approach to Support Fault Diagnosis and Quality Control in End-Of-Line Testing, Procedia CIRP 72 (2018) 1333–1338. doi:10.1016/j.procir.2018.03.024.

[4] Y. Wilhelm, et al., Data Science Approaches to Quality Control in Manufacturing: A Review of Problems, Challenges and Architecture, in: Proc. of the 14th Symposium on Service-Oriented Computing (SummerSOC), Springer-Verlag, 2020, pp. 45–65. doi:10.1007/978-3-030-64846-6_4.

[5] M. Spieß, et al., Analysis of Incremental Learning and Windowing to handle Combined Dataset Shifts on Binary Classification for Product Failure Prediction, in: Proc. of the 24th International Conference on on Enterprise Information Systems (ICEIS), 2022, pp. 394–405. doi:10.5220/0011093300003179.

[6] G. Köksal, I. Batmaz, M. C. Testik, A Review of Data Mining Applications for Quality Improvement in Manufacturing Industry, Expert Systems with Applications 38 (2011) 13448–13467. doi:10.1016/j.eswa.2011.04.063.

[7] T. Wuest, et al., Machine Learning in Manufacturing: Advantages, Challenges, and Applications, Production & Manufacturing Research 4 (2016) 23–45. doi:10.1080/21693277.2016.1192517.

[8] C. Gröger, Building an Industry 4.0 Analytics Platform, Datenbank-Spektrum 18 (2018) 5–15. doi:10.1007/s13222-018-0273-1.

[9] C. Gröger, There Is No AI Without Data, Communications of the ACM 64 (2021) 98–108. doi:10.1145/3448247.

[10] V. Hirsch, P. Reimann, B. Mitschang, Exploiting Domain Knowledge to Address Multi-Class Imbalance and a Heterogeneous Feature Space in Classification Tasks for Manufacturing Data, PVLDB 13 (2020) 3258–3271. doi:10.14778/3415478.3415549.

[11] V. Hirsch, P. Reimann, B. Mitschang, Data-Driven Fault Diagnosis in End-of-Line Testing of Complex Products, in: Proc. of the 6th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, Washington, D.C., USA, 2019. doi:10.1109/DSAA.2019.00064.

[12] C. Kiefer, P. Reimann, B. Mitschang, QUALM: Ganzheitliche Messung und Verbesserung der Datenqualität in der Textanalyse,

Datenbank-Spektrum 19 (2019) 137–148. doi:10.1007/s13222-019-00318-7.

[13] J. Ziegler, et al., A Metadata Model to Connect Isolated Data Silos and Activities of the CAE Domain, in: Proc. of the 33rd International Conference on Advanced Information Systems Engineering (CAiSE), Springer International Publishing, 2021, pp. 213–228. doi:10.1007/978-3-030-79382-1_13.

[14] V. Hirsch, et al., Exploiting Domain Knowledge to Address Class Imbalance and a Heterogeneous Feature Space in Multi-Class Classification, International Journal on Very Large Data Bases (VLDB Journal) (2023). doi:10.1007/s00778-023-00780-6.

[15] Y. Wilhelm, et al., Overview on Hybrid Approaches to Fault Detection and Diagnosis: Combining Data-driven, Physics-based and Knowledge-based Models, Procedia CIRP 99 (2021) 278–283. doi:10.1016/j.procir.2021.03.041.

[16] C. Weber, P. Hirmer, P. Reimann, A Model Management Platform for Industry 4.0 - Enabling Management of Machine Learning Models in Manufacturing Environments, in: Proc. of the 23rd International Conference on Business Information Systems (BIS), 2020, pp. 403–417. doi:10.1007/978-3-030-53337-3_30.

[17] A. G. Villanueva Zacarias, et al., AssistML: A Concept to Recommend ML Solutions for Predictive Use Cases, in: Proc. of the 8th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2021, pp. 148–155. doi:10.1109/DSAA53316.2021.9564168.

[18] Y. Wilhelm, et al., PUSION- A Generic and Automated Framework for Decision Fusion, in: Proc. of the 39th International Conference on Data Engineering (ICDE), IEEE, Anaheim, CA, USA, 2023.

[19] J. Voggesberger, P. Reimann, B. Mitschang, Towards the Automatic Creation of Optimized Classifier Ensembles, in: Proc. of the 25th International Conference on Enterprise Information Systems (ICEIS), Prague, Czech Republic, 2023, pp. 614–621.

[20] C. Weber, et al., A New Process Model for the Comprehensive Management of Machine Learning Models, in: Proc. of the 21st International Conference on Enterprise Information Systems (ICEIS), Heraklion, Greece, 2019, pp. 415–422. doi:10.5220/0007725304150422.

[21] A. G. Villanueva Zacarias, P. Reimann, B. Mitschang, A Framework to Guide the Selection and Configuration of Machine-Learning-based Data Analytics Solutions in Manufacturing, Procedia CIRP 72 (2018) 153–158. doi:10.1016/j.procir.2018.03.215.

[22] A. G. Villanueva Zacarias, R. Ghabri, P. Reimann, AD4ML: Axiomatic Design to Specify Machine Learning Solutions for Manufacturing, in: Proc. of the 21st International Conference on Information Reuse and Integration for Data Science (IRI), IEEE, 2020, pp. 148–155. doi:10.1109/IRI49571.2020.00029.

[23] J. G. Moreno-Torres, et al., A Unifying View on Dataset Shift in Classification, Pattern Recognition 45 (2012) 521–530. doi:10.1016/j.patcog.2011.06.019.

[24] C. Kiefer, P. Reimann, B. Mitschang, Prevent Low-Quality Analytics by Automatic Selection of the Best-Fitting Training Data, in: Proc. of the 53rd Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii, USA, 2020. doi:10.24251/HICSS.2020.129.

[25] C. Kiefer, P. Reimann, B. Mitschang, A Hybrid Information Extraction Approach Exploiting Structured Data Within a Text Mining Process, in: Proc. of the 18th Conference for Business, Technology and Web (BTW), Gesellschaft für Informatik (GI), Rostock, Germany, 2019, pp. 149–168. doi:10.18420/btw2019-10.

[26] S. Vinodh, D. Kuttalingam, Computer-aided Design and Engineering as Enablers of Agile Manufacturing, Journal of Manufacturing Technology Management 22 (2011) 405–418. doi:10.1108/17410381111112747.

[27] P. Reimann, et al., SIMPL – A Framework for Accessing External Data in Simulation Workflows, in: Proc. of the 14th Conference on Database Systems for Business, Technology and Web (BTW), Gesellschaft für Informatik (GI), Kaiserslautern, Germany, 2011, pp. 534–553.

[28] P. Reimann, H. Schwarz, B. Mitschang, A Pattern Approach to Conquer the Data Complexity in Simulation Workflow Design, in: R. M. et al. (Ed.), Proc. of the 22nd International Conference on Cooperative Information Systems (CoopIS), Springer-Verlag, Amantea, Italy, 2014, pp. 21–38.

[29] J. Ziegler, et al., A Graph Structure to Discover Patterns in Unstructured Processes of Product Development, in: Proc. of the 23rd International Conference on Information Reuse and Integration for Data Science (IRI), IEEE, 2022.

[30] M. Woźniak, M. Graña, E. Corchado, A Survey of Multiple Classifier Systems as Hybrid Systems, Information Fusion 16 (2014) 3–17. doi:10.1016/j.inffus.2013.04.006.

[31] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[32] M. Galar, et al., A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches, IEEE Transactions on Systems, Man, and Cybernetics 42 (2012) 463–484. doi:10.1109/TSMCC.2011.2161285.