Domain Knowledge in CRISP-DM: An Application Case in Manufacturing

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 56-2 (2023) 7603–7608

Domain Knowledge in CRISP-DM: An Application Case in Manufacturing

Nils Lundén [*,**], **Ebru Turanoglu Bekar** [*], **Anders Skoogh** [*], **Jon Bokrantz** [*]

[*] *Department of Industrial and Materials Science, Chalmers University of Technology, Gothenburg, Sweden (e-mail: nils.lunden@volvo.com, ebrut@chalmers.se, anders.skoogh@chalmers.se, jon.bokrantz@chalmers.se)*

[**] *ESW Stategy & Solutions, Volvo GTO, Gothenburg*

Abstract: To keep up with shifting technology trends and remain competitive, more manufacturing companies are investigating how to utilize data analytics to improve their processes. An issue these companies often face today is the need for more competence to perform advanced analytics projects within their departments. By using a human-in-the-loop approach and efficiently utilizing current domain knowledge in combination with data analytics, the higher success of implementation can be achieved. A common approach today to perform data analytics projects is to use the general Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. This methodology does not consider the challenges specific to manufacturing and how to include domain expertise. This paper, therefore, suggests how the CRISP-DM methodology can be adapted to compensate for these issues. The adapted methodology is demonstrated in a case study for improving quality in the machining process by using interpretable machine learning models that can be used to assist experts when performing root cause analysis. This contributes to showing how to use domain experts' knowledge better and how data analytics can be used in conjunction with domain-specific methods.

Keywords: Human-in-the-loop, Domain-knowledge, CRISP-DM, Machine learning, Manufacturing, Root cause analysis, Data analytics

1. INTRODUCTION

Digitalization and data analytics have attracted significant interest within the manufacturing industry during the last few years. More companies are looking into how to gain valuable insights from the large quantities of data at their disposal. However, this is a challenging transition for most companies, where one main reason is the lack of data analytics competencies within manufacturing departments (Dogan and Birant, 2021). Furthermore, manufacturing problems are usually complex and require domain-specific knowledge and tools, which are difficult to replace by only relying on data analytics. Quality problems, a common use case for data analytics in manufacturing, are an example of a complex area requiring skilled domain experts to deduce causes of process instability. Advanced analytical methods such as machine learning (ML) are often not transparent, making it difficult for humans to understand the model's behaviors (Soldatos and Kyriazis, 2021). This makes it challenging to identify the root causes of quality problems and improve manufacturing processes based on analytical results, as well as create acceptance of the result among domain experts.

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a common methodology for performing data analytics in manufacturing due to its ability to integrate business goals with the analytics process (Huber et al., 2019). CRISP-DM is a general methodology and not specific to manufacturing and the unique challenges in this area. It needs to describe how to use available domain knowledge best and how it can be implemented with outdated legacy systems. Therefore, this makes it difficult for manufacturing companies to use the methodology efficiently (Lundén, 2022).

This paper proposes an adapted CRISP-DM methodology to enable the fusion of domain-specific knowledge with the data analytics process. An application case is performed where the methodology is implemented to improve quality in a machining operation. The result demonstrates the needs and benefits of human-in-the-loop when performing data analytics projects in manufacturing. It is also shown how interpretable ML can augment process experts' ability to perform root cause analysis (RCA). This is done by developing models with high transparency and explainable results.

The paper is structured as follows: Section 2 presents related work to these topics. Section 3 describes the adapted CRISP-DM methodology. Section 4 shows the results from the case study, and section 5 provides a discussion of the results and concluding remarks.

2. RELATED WORK

Industry 4.0 has been an important research subject focused on how new technology can be an enabler within

10.1016/j.ifacol.2023.10.1156

manufacturing. In recent state-of-the-art research, the follow-up, Industry 5.0, has also gained increasing attention. Cotta et al. (2021) explains that the main difference with Industry 5.0 is the human-centric approach where human needs are prioritized in production. They elaborate that with this mindset, technology should guide and assist humans in their work rather than adapting the workforce to rapidly changing technology. In this way, advanced data analytics approaches such as ML can be used complementary to traditional methods already known to the domain experts for dealing with quality problems for example. This section discusses which previous suggestions have been made to apply CRISP-DM with a human-centric approach and also how the analytics result can better assist domain experts by using interpretable ML.

2.1 CRISP-DM in manufacturing

It is important to include business understanding and expertise in industrial contexts when conducting data analytics projects. Without proper domain understanding, there is a risk of only detecting trivial or already known patterns which are not relevant to the business goal (Ladj et al., 2021). For this reason, the cross-industry standard process for data mining (CRISP-DM) has become widely spread for data mining in manufacturing. CRISP-DM is a general framework for translating business problems into data analytics goals (Huber et al., 2019). It constitutes six phases that are performed sequentially: Business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The initial *business understanding* phase is where an understanding of the project objectives from a business perspective is formed and then converted into a data mining problem (Wirth and Hipp, 2000). The following *data understanding* phase concerns creating a better understanding of the data and potential data quality problems. The *data preparation* phase is thereafter used to create a proper structure and transformation of the data that can be used for modeling. In the *modeling* phase, the aim is to select appropriate ML algorithms and train models, which are later evaluated in the *evaluation* phase according to setting business goals. The final *deployment* phase aims to use gained insights from the project to implement improvements in the organization. A more in-depth description of the CRISP-DM methodology is described by Wirth and Hipp (2000).

The CRISP-DM model can be viewed as a general guideline for planning and documentation of the data mining process. It needs to be adapted to the domain where it is executed to consider the domain-specific requirements. Commonly discussed topics when applying CRISP-DM in manufacturing involve experts within different fields in the manufacturing environment in multiple steps of the process and that higher emphasis needs to be put on data acquisition (Martínez-Plumed et al., 2019).

Huber et al. (2019) propose that a technical understanding phase is added to create a better understanding of which parameters are needed to gather data from by consulting technical experts. They argue that a technical realization phase should be performed to investigate how to acquire this data with sufficient quality. (Ungermann et al., 2019) discusses that domain knowledge should be used to prioritize what parameters should be collected initially if the realization of data collection is time-consuming or expensive.

Kristoffersen et al. (2019) points out that it is essential to re-use gained insights when applying CRISP-DM in manufacturing to facilitate scaling of data analytics. They propose that *analytic profiles* should be used to document steps taken in projects. These profiles can be used to implement the following use cases of similar nature. The authors also argue that manufacturing domain experts should be involved later in the data preparation process to validate if the transformed data still represents the original business problem well.

It is commonly mentioned that successful implementation of data analytics in manufacturing requires integrating domain-specific tools for improved business and data understanding. Ishikawa diagrams, Failure Mode and Effects Analysis (FMEA), and Five Whys (5W) are mentioned as examples (Huber et al., 2019; Ungermann et al., 2019; Kampker et al., 2018; Pradhan et al., 2007).

When applying CRISP-DM in manufacturing, there are several considerations to adapt it to the domain. So far, a comprehensive data analytics process covering all these areas has yet to be developed specifically for usage in manufacturing. This might make it difficult for new or aspiring practitioners of data analytics within the industry to make valuable use of this advanced data analytics. This inspired the development of a holistic version of CRISP-DM better suited for the manufacturing area that considers several of the mentioned issues with the original methodology.

2.2 Interpretable ML for root cause analysis

A common approach for dealing with quality issues is to identify the root cause of the problem so the issue can be corrected. Traditionally, RCA has been performed by experienced personnel at a site to describe the causal relations between process steps and the quality outcome with the help of manual methods as described by Lokrantz et al. (2018). However, they explain that the problem with this approach is that this valuable knowledge might sometimes be difficult to transfer between individuals or sites, and the experts might be biased in their assessments. Data analytics on a diagnostic level offers possibilities to perform objective RCA where models learn the relations in a machining process from large volumes of sensor data which traditional methods can not make good use of.

Different approaches can be taken to achieve data-driven RCA depending on the previous knowledge of the causal effects in the machining process. Miguéis et al. (2022) describes how RCA is performed when there is a low understanding between parameters and the problem. The idea builds on creating associations between the parameters and the problems, which can be derived by feature importance from ML classification models. They explain that parameters strongly associated with the problem are likely to be root causes and can be further examined by experts. The requirement for applying this approach is to have a labeled data set so that supervised learning models can be used to learn the relations between input parameters and the labeled quality outcome. It is further explained by Miguéis et al. (2022) that classification models with an interpretive structure should be used to fully understand which features have the highest importance for the problem. The interpretability aspect when applying ML for RCA is elaborated on by Mueller et al. (2018). They distinguish between white-box models and black-box models, where humans can more easily understand white-box models' structures. They mention Decision Tree models as a suitable example for RCA application as the feature importance can easily be extracted from such a model. Black-box models such as Support Vector Machines or Neural Networks should be avoided in these applications as they are highly non-linear and complex to interpret (Mueller et al., 2018).

It is important to differentiate between correlation and causation when performing RCA, as only finding correlated values might not reveal the root cause of the problem. The methods explained above are mainly helpful in finding the correlations between input and output data and can act as decision support for domain experts to find causation among the parameters. Therefore, the model developed in the case study was created with the focus on being explainable and clear to enable further investigation by process experts.

3. ADAPTATION OF CRISP-DM FOR MANUFACTURING

The adapted CRISP-DM methodology aims to address the shortcomings for implementation in manufacturing described in Section 2.1. The adapted version emphasizes how to better use domain knowledge, describes a data collection strategy, and how to reuse knowledge. A methodology flowchart is shown in Figure 1 where the added key elements to the standard CRISP-DM methodology are highlighted. Lundén (2022) presents a detailed review of the methodology. In this section, the main ideas are presented for how to adapt CRISP-DM for manufacturing but with maintained generality for data analytics. In Section 4, it is shown how this methodology can be used with ML for addressing quality issues.

During the initial business understanding phase, the needed domain experts should be identified to get a holistic view of the problem. It is important that multidisciplinary knowledge can be shared efficiently between data analysts

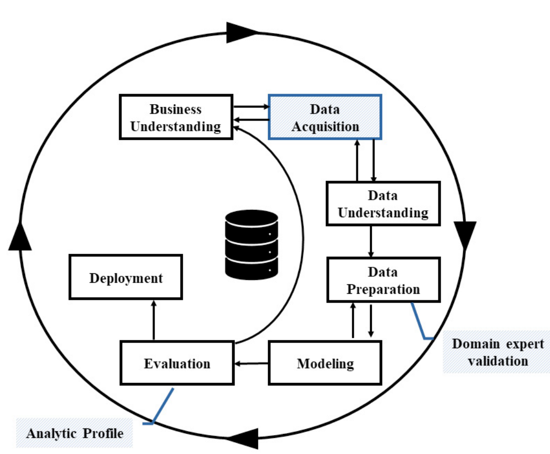

Fig. 1. Adapted workflow of the CRISP-DM methodology

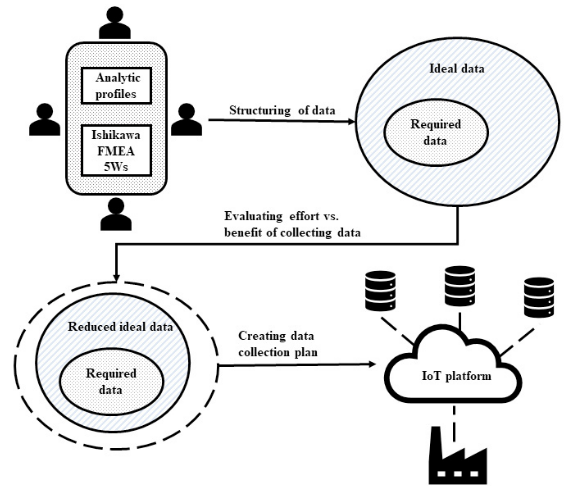

Fig. 2. Data acquisition flow from identification of parameters to collection

and domain experts. Therefore, basic knowledge within both fields is ideal for efficient collaboration. Quantifiable business goals relating to production KPIs should be set and then translated into data analytics goals.

A data acquisition phase is added to accommodate the complexity of data collection in brownfield factories. Like the ideas presented by Huber et al. (2019), the phase is divided into steps involving the identification of relevant data sources and, after that, acquisition. An overview of this process can be seen in Figure 2. The first step is identifying relevant parameters using domain-specific tools like Ishikawa diagrams. This data is, after that, structured into an ideal data set (what ideally should be collected if possible) and a required data set containing parameters that must be collected to succeed with the project. Based on cost versus benefit assessments for collecting the parameter, it might be necessary to reduce the ideal data set. When the CRISP-DM process is iterated, more parameters in the ideal data set can be collected if necessary. The collection is realized by setting up a connection between IoT sources from the shopfloor as well as other data sources to the company's analytics platform.

As discussed by Kristoffersen et al. (2019), a validation performed by domain experts who will use the model should be performed during the data preparation phase. This confirms that the data is not altered beyond its intended purpose. For conducting RCA, the extracted features from the data must be understandable.

Modeling and evaluation are performed as usual. Whether the project is successful or not, documentation should be completed where an analytics profile is created after the evaluation. This will support the following projects when deciding what domain experts to include, what data to use, and what methods. The next section presents an implementation of the adapted CRISP-DM methodology for the case study.

4. CASE STUDY RESULT

The case study was performed at a truck manufacturing company within a department where engine components are machined. A multi-purpose machine in which machines intake and exhaust holes in cylinder blocks has long been an issue for the department due to dimensional errors causing high scrap rates. Several manufacturing experts have been involved in a project trying to find the root causes of dimensional errors using different methods and have yet to be successful results. The case study aimed to use data analytics to assist with RCA and provide new insights to domain experts.

As presented in the previous section, the adapted CRISP-DM methodology incorporated the experts' knowledge in the data analytics process. This section shows the implementation of the steps in the methodology.

4.1 Business understanding

To achieve a holistic view of the problem, people with different competencies were involved from the first phase. The team consisted of production engineers, machine operators, IoT technicians, and data analysts. The business goal for the case was to provide data-driven insights that can assist with RCA, which ultimately will reduce the number of scraped workpieces produced by the machine. The data analytics goal was to create a model that can determine which parameters have the highest impact on quality.

4.2 Data acquisition

A workshop was performed with involved experts to identify all parameters which might have an impact. An Ishikawa diagram was used to structure relevant data related to the workpiece, machining process, machine tool, machining result, machining time, and human factors. The required data from this set was those related to temperatures and tools, as the domain experts hypothesized that these parameters were critical. Some of the identified parameters were excluded due to high acquisition effort. Vibrations were an example of such parameters, as they resided in a separate system that was difficult to connect to.

The machine is old and operates in a brownfield factory environment. No significant investments were made for the case study. Collecting data from the factory's current legacy system was essential. ThingWorx [1] was used as the data analytics platform to which the parameters were connected using standard interfaces. After a historical data set had been collected, the data was exported from the platform in csv-format. The analytics was performed with Python using the open-source scikit-learn [2] library.

4.3 Data understanding

Data from 5 739 machine cycles were collected for analysis. Each sample was labeled with an OK or NOK quality result. The labeling was performed with an automatic measuring machine which asserts that machined holes are within tolerances. In Figure 3, the output from a subset of the samples is seen, where a few samples have been outside the upper or lower tolerance limit. Quality output was also gathered from a log written by machine operators, where more details of the defects were provided.

[1] https://www.ptc.com/en/products/thingworx

[2] https://scikit-learn.org/

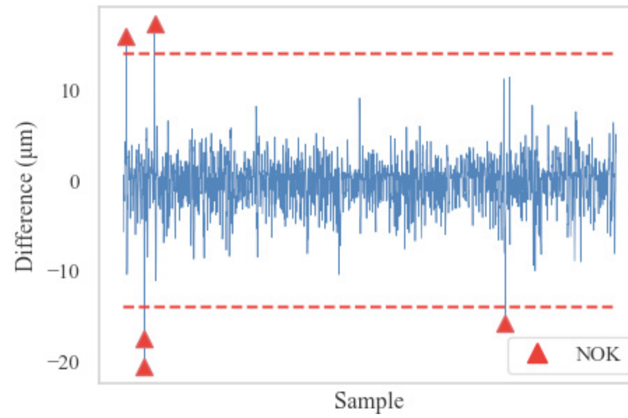

Fig. 3. Measured deviation from the nominal diameter of one of the exhaust holes in the cylinder block over a subset of the samples

The data was collected in multiple formats. Some parameters had to be excluded due to many missing values or irregular collection frequencies. There was also an issue with unbalance since only 0.35% of the samples had a NOK result. This is a common issue for quality-related data analytics cases in manufacturing and is usually dealt with by using a balancing technique (Taisch et al., 2020).

4.4 Data preparation

The data preparation was performed in four steps. In the first step, *data cleaning*, samples with many missing values or outliers were removed. After this, *data integration* was performed to combine data from different sources into a single tabular format.

Data tranformation was carried out to extract features from the data. Christ et al. (2016) proposed a method to extract features from time series collected from industrial sensors. This technique was applied by fragmenting the data into equal and synchronized time frames. Afterward, as suggested by the authors, the time series were split into time windows where the data was aggregated into statistical features within each time window. In this case, equal and non-overlapping time windows were used for simplicity and easier interpretation. The time series features and all other parameters used as features were normalized to a standard zero-to-one scale. This was done to reduce any negative impact the variation in measurement units might have on the modeling result.

In the last data preparation step, *dimension reduction*, Analysis of Variance (ANOVA) was used to reduce the number of features used for modeling. The algorithm ranked the features by their ability to predict the quality outcome. A test design was set up that tried different amounts of features during modeling.

As proposed in Section 3, a validation by domain experts should be performed after the data preparation step. It was ensured that features were not altered to an extent where it would be difficult to interpret their meaning. The experts saw partitioning into time windows for sensor data as useful for RCA as this makes it possible to determine when the cycle problems occur.

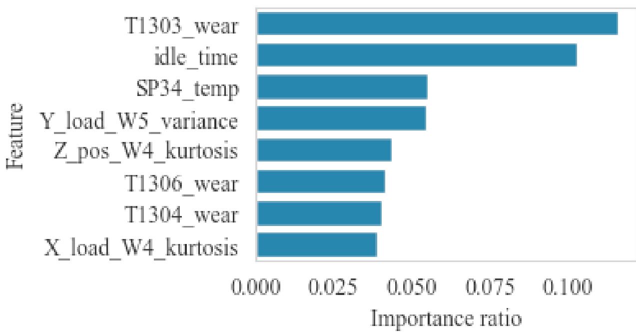

Fig. 4. Top eight important features for class prediction with the Random Forest model

4.5 Modeling

As the data set is labeled, a supervised ML approach was used. As suggested by Mueller et al. (2018), tree-based models offer high interpretability and can derive feature importance. Therefore, Decision Tree, Random Forest, and XGBoost were chosen as potential training algorithms. The detailed theoretical background information regarding these algorithms can be found in an edited book Han et al. (2022). Logistic Regression was also evaluated as this model offers high interpretability (Brownlee, 2020).

To reduce the negative impact of an unbalanced data set, Synthetic Minority Oversampling Technique (SMOTE) was used to create new synthetic samples with NOK labels. To avoid introducing too much bias in the data set, SMOTE was combined with a random undersampling technique as suggested by Chawla et al. (2002).

The data set was split into training and test data with a 75 to 25 ratio. The number of features, time windows (for time series parameters), and the balancing ratio was changed iteratively to find the best-performing models. The best performing configurations and the performance for each model can be seen in Table1.

Four common evaluation metrics were used for model evaluation: Accuracy, precision, recall, and F1-score. Accuracy measures the ratio of correct predictions, whereas the other metrics evaluate the percentage of correct true positive and true negative predictions. Handelman et al. (2019) provides detailed explanations of how the metrics are calculated. Due to the unbalance between the classification label, accuracy is not a good indicator of model performance. The precision, recall, and F1-score better explain the model's abilities to distinguish between OK and NOK samples.

Table 1. Configuration and performance metrics for the evaluated models

Configuration/ Metric	Random Forest	Decision Tree	Logistic Reg.	XG-Boost
No. win.	5	5	5	8
Balance ratio	3:10	1:5	1:5	1:5
No. Features	50	50	40	50
Accuracy	0.99	0.99	0.98	0.99
Precision	1.00	0.25	0.13	0.60
Recall	0.75	1.00	0.75	0.75
F1-score	0.86	0.40	0.22	0.67

The Random Forest model had the highest F1-score among investigated models. The feature importance was extracted from this model with the scikit-learn library and can be seen in Figure 4. Among the most important features were tool wear, machine idle time before the cycle, spindle temperature, axis load, and axis positions.

4.6 Evaluation

Due to the high imbalance in the data with few samples with NOK quality results, it was deemed that more data should be collected over a more extended period. This will ensure a more robust training set for the models with reduced impact of bias.

Making features interpretable with unambiguous names can give domain experts clear insights about important parameters. The top-ranked features can be further examined through visualization. Figure 5 shows an example of how one of the top-ranked features, the y-axis load variance in time window 5 (Y_load_W5_variance), can be illustrated. This allows domain experts to further identify anomalous patterns during specific times in the machining cycle.

5. DISCUSSION AND CONCLUSIONS

In this paper, suggestions have been made for adapting the CRISP-DM methodology into a more holistic approach for implementing data analytics in manufacturing with a fusion of domain knowledge. It has been discussed in literature (Huber et al., 2019; Ungermann et al., 2019) that data acquisition is a common problem for realizing the true potential of data analytics in manufacturing. The implementation of the suggested methodology successfully clarified to the company how to use domain experts' knowledge to identify critical data and prioritize what data to collect. Other suggestions were also made for how to best utilize manufacturing domain knowledge in different phases of CRISP-DM, such as validation, previously suggested by Kristoffersen et al. (2019). This proved advantageous in the case study as it allowed for early adoption of the analytics process to achieve a modeling result suitable for the process experts who will use the result.

The case study where the adapted methodology was implemented focused on how to assist with RCA in quality use

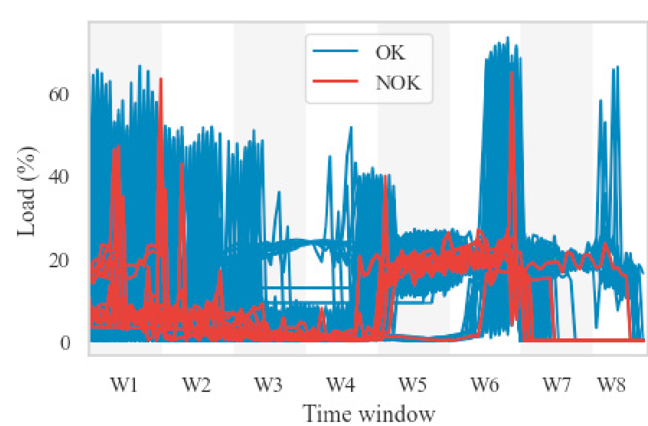

Fig. 5. Visualization of load as a percentage of max load for the Y_load_W5_variance parameter

cases. It was demonstrated how this might be achieved by creating easily understood features for modeling and using ML models with high interpretability. This allowed for finding parameters with a high correlation to the quality result and provided new insights to the process experts. Finding correlating parameters is not enough to automate the RCA process as causality in the data then needs to be found, which is a difficult task with ML (Pradhan et al., 2007). The model results should therefore be deployed with an interactive tool that domain experts can use in combination with other domain-specific tools for performing RCA. The result showed examples of data visualizations that can guide the experts in finding deviating patterns in the machining cycles.

The findings in this study provide more clarity as to how data analytics can be implemented in manufacturing in a human-centric manner. It shows how to implement a human-in-the-loop approach, which can help companies where it is uncertain how to transition to data analytics can be made in the best way with their current competencies. The case study also demonstrates how humans are involved in defining goals, selecting data, labeling data, validating the process, and interpreting the results. By using efficient cross-functional collaboration between data analysts and domain experts in this way, the knowledge barriers to implementing data analytics can be reduced.

ACKNOWLEDGEMENTS

This paper was produced from the MSc thesis of the first author at the Chalmers University of Technology with supervision from the other authors. Special thanks are given to the Machining department at the company for providing the case study and the support from several domain experts who helped conduct the study. Thanks also to the Production Area of Advance at the Chalmers University of Technology for the research support.

REFERENCES

Brownlee, J. (2020). How to calculate feature importance with python. *Machine Learning Mastery. https://machinelearningmastery. com/calculate-feature-importance-with-python.*

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357.

Christ, M., Kempa-Liehr, A.W., and Feindt, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. *arXiv preprint arXiv:1610.07717.*

Cotta, J., Breque, M., De Nul, L., and Petridis, A. (2021). Industry 5.0: towards a sustainable, human-centric and resilient european industry. *European Commission Research and Innovation (R&I) Series Policy Brief.*

Dogan, A. and Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, 166, 114060.

Han, J., Pei, J., and Tong, H. (2022). *Data mining: concepts and techniques.* Morgan kaufmann.

Handelman, G.S., Kok, H.K., Chandra, R.V., Razavi, A.H., Huang, S., Brooks, M., Lee, M.J., and Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology*, 212(1), 38–43.

Huber, S., Wiemer, H., Schneider, D., and Ihlenfeldt, S. (2019). Dmme: Data mining methodology for engineering applications–a holistic extension to the crisp-dm model. *Procedia Cirp*, 79, 403–408.

Kampker, A., Heimes, H., Bührer, U., Lienemann, C., and Krotil, S. (2018). Enabling data analytics in large scale manufacturing. *Procedia Manufacturing*, 24, 120–127.

Kristoffersen, E., Aremu, O.O., Blomsma, F., Mikalef, P., and Li, J. (2019). Exploring the relationship between data science and circular economy: an enhanced crisp-dm process model. In *Conference on e-Business, e-Services and e-Society*, 177–189. Springer.

Ladj, A., Wang, Z., Meski, O., Belkadi, F., Ritou, M., and Da Cunha, C. (2021). A knowledge-based digital shadow for machining industry in a digital twin perspective. *Journal of Manufacturing Systems*, 58, 168–179.

Lokrantz, A., Gustavsson, E., and Jirstrand, M. (2018). Root cause analysis of failures and quality deviations in manufacturing using machine learning. *Procedia Cirp*, 72, 1057–1062.

Lundén, N. (2022). *Implementing data analytics for improved quality in manufacturing: a case study.* Master's thesis, Chalmers University of Technology.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J.H., Kull, M., Lachiche, N., Quintana, M.J.R., and Flach, P.A. (2019). Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering.*

Miguéis, V.L., Borges, J.L., et al. (2022). Automatic root cause analysis in manufacturing: an overview & conceptualization. *Journal of Intelligent Manufacturing*, 1–18.

Mueller, T., Greipel, J., Weber, T., and Schmitt, R.H. (2018). Automated root cause analysis of non-conformities with machine learning algorithms. *Journal of Machine Engineering*, 18.

Pradhan, S., Singh, R., Kachru, K., and Narasimhamurthy, S. (2007). A bayesian network based approach for root-cause-analysis in manufacturing process. In *2007 International Conference on Computational Intelligence and Security (CIS 2007)*, 10–14. IEEE.

Soldatos, J. and Kyriazis, D. (2021). Trusted artificial intelligence in manufacturing: A review of the emerging wave of ethical and human centric ai technologies for smart production.

Taisch, M., Casidsid, M., May, G., Morin, T., Padelli, V., Pinzone, M., Wuest, T., et al. (2020). World manufacturing report 2020: manufacturing in the age of artificial intelligence.

Ungermann, F., Kuhnle, A., Stricker, N., and Lanza, G. (2019). Data analytics for manufacturing systems–a data-driven approach for process optimization. *Procedia CIRP*, 81, 369–374.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, 29–40. Manchester.