Computers in Industry 149 (2023) 103930

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

A survey on machine learning based analysis of heterogeneous data in industrial automation

Simon Kamm *, Sushma Sri Veekati, Timo Müller, Nasser Jazdi, Michael Weyrich

University of Stuttgart, Institute of Industrial Automation and Software Engineering, Pfaffenwaldring 47, 70550 Stuttgart, Germany

ARTICLE INFO

Keywords:
Machine learning
Multi-modal machine learning
Adaptive machine learning
(Physics-) informed machine learning
Heterogeneous data integration
Heterogeneous data management

ABSTRACT

In many application domains data from different sources are increasingly available to thoroughly monitor and describe a system or device. Especially within the industrial automation domain, heterogeneous data and its analysis gain a lot of attention from research and industry, since it has the potential to improve or enable tasks like diagnostics, predictive maintenance, and condition monitoring. For data analysis, machine learning based approaches are mostly used in recent literature, as these algorithms allow us to learn complex correlations within the data. To analyze even heterogeneous data and gain benefits from it in an application, data from different sources need to be integrated, stored, and managed to apply machine learning algorithms. In a setting with heterogeneous data sources, the analysis algorithms should also be able to handle data source failures or newly added data sources. In addition, existing knowledge should be used to improve the machine learning based analysis or its training process. To find existing approaches for the machine learning based analysis of heterogeneous data in the industrial automation domain, this paper presents the result of a systematic literature review. The publications were reviewed, evaluated, and discussed concerning five requirements that are derived in this paper. We identified promising solutions and approaches and outlined open research challenges, which are not yet covered sufficiently in the literature.

1. Introduction

Within the industrial automation domain including manufacturing, more and more data is acquired from diverse data sources (e.g. different sensors or text input). Between these data, there is a system-dependent correlation. This leads to increasingly heterogeneous data, also known as part of the big data dimension *variety*. Nevertheless, this increasingly heterogeneous data has to be analyzed to gain insight and generate knowledge or even wisdom based on the available data and information to gain insight and generate knowledge or even wisdom based on the available data and information (following the DIKW, i.e. Data, Information, Knowledge, Wisdom, pyramid (Rowley, 2007)). With the recent progress in the field of machine learning, new approaches analyzing heterogeneous data for industrial automation emerged. Additionally, novel approaches to integrate and manage heterogeneous data are developed, which are mandatory to enable the recently evolved analysis algorithms in an industrial environment. These fields (data integration and management as well as machine learning) need to be linked and combined to analyze and gain benefits from the heterogeneous data, which is defined in more detail in the following.

Data and information coming from different data sources incorporate differences in data formats, data representations, and data modalities, which can be used as a non-complete colloquial definition of heterogeneous data: *Heterogeneous data occurs, when the data samples have* differences, or *the acquired data shows differences in a data-related property, such as their format, their representation, their meaning* etc. Within (Jirkovsky and Obitko, 2014; L'heureux et al., 2017; Wang, 2017; Jirkovsky et al., 2016), similar classifications of data heterogeneity are given. Following them, it can be differentiated between five classes of heterogeneity: Syntactical, semantical, statistical, terminological, and semiotic heterogeneity. In L'heureux et al. (2017), syntactic, semantic, and statistical heterogeneity are highlighted. Syntactic heterogeneity is described as heterogeneity in data types or file formats, While semantic heterogeneity covers different meanings and interpretations. It is further described, that heterogeneity in statistics (statistical heterogeneity) means the difference in statistical properties among the different samples within an overall dataset. The four same types of heterogeneity are defined in (Jirkovsky and Obitko, 2014; Wang, 2017; Jirkovsky et al.,

* Corresponding author.
E-mail address: simon.kamm@ias.uni-stuttgart.de (S. Kamm).

https://doi.org/10.1016/j.compind.2023.103930
Received 16 January 2023; Received in revised form 16 March 2023; Accepted 15 April 2023
Available online 23 April 2023

2016), syntactic, terminological, semantic, and semiotic heterogeneity. Syntactic and semantic heterogeneity is described as synonymous with L'heureux et al., (2017). Terminological heterogeneity is the difference in terms, e.g. a different natural language expression for the name of the same sensor. Semiotic heterogeneity denotes different interpretations of the data from the data sources by different people.

In Liang et al. (2022), data is seen from a multimodal view, where different modalities (e.g. text and video) exist. Six dimensions of heterogeneity are named: Element representation, distribution, structure, information, noise, and relevance. Element representation is the basic element of how a modality is represented, e.g. a set of characters for text. This dimension "*measures heterogeneity in the sample space or representation space of modality elements*" (Liang et al., 2022). Distribution is the frequency and likelihood of elements in modalities. Structure means the difference in the modality-internal structure. For instance, the text modality follows a language-dependant structure. The information dimension shows the difference in the information contained within the different modalities, since not all modalities may contain the same level of information. Noise means, that in each modality, different noise levels and distributions can occur. The final dimension relevance measures the difference in the relevance of each modality for a specific task. These six dimensions fit and overlap with the previously introduced and established definitions of data heterogeneity, seeing data heterogeneity more from the modality perspective, while the previously introduced definitions are more general. To have an overview of heterogeneous data and also group scenarios, we introduce the HEterogeneous DAta CUbe (HEDACUBE) as shown in Fig. 1.

The introduced definitions are adjusted to give a view of heterogeneous data and its terms and definition, for instance, "distribution" is removed since it is covered in the class "statistical heterogeneity". Within this cube, the data source dimensions can be matched to the heterogeneity classes for a given scenario. Heterogeneity is no binary variable, therefore the third dimension shows the smooth transition from homogeneous to heterogeneous data. The class statistical heterogeneity refers to a dataset, whereas all other dimensions and classes refer to individual samples. Non-meaningful combinations were highlighted, where the heterogeneity class does not match the data source dimension. In the following, some examples will be given. Statistical heterogeneity can for instance occur within the noise, the information, or the relevance of data sources. All these dimensions can vary over a whole dataset from a statistical point of view. Semiotic heterogeneity, the different interpretations of the data is related to noise, information, and relevance, which all can be differently interpreted by people. Semantic heterogeneity can happen for all dimensions, with a possible difference in coverage or perspective affecting the information and relevance. Syntactic heterogeneity (heterogeneity in data types or file formats) can take place in the information, structure, noise, or element representation of a data source. The contained information can, for example, differ between two file formats when a file format has a higher compression rate in comparison to another file format (such as jpg-format with a higher compression rate compared to tiff-format). Terminological heterogeneity is linked with the dimensions of element representation and structure, where different terms may be used to represent elements or within the structure of the data source.

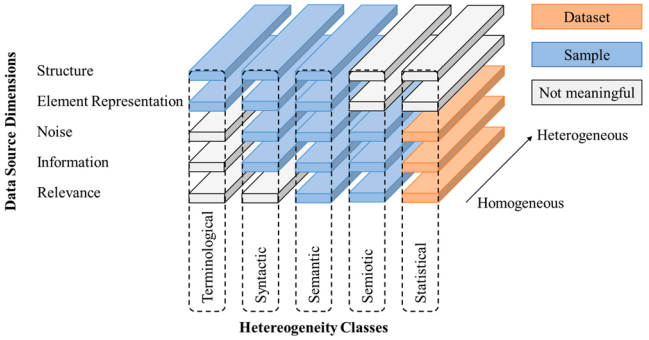

Fig. 1. Introduction of HEDACUBE (heterogeneous data cube).

As discussed in Kamm et al. (2021), syntactical and semantical heterogeneity are the two most relevant data heterogeneity classes concerning industrial automation, since the other classes are easier to resolve or can not be resolved at all. Terminological heterogeneity can be resolved by renaming or giving naming rules. Semiotic heterogeneity is nearly impossible to be resolved since different interpretations by different people can always occur. By a reduced semantic heterogeneity and thus clearer semantic modeling of the data, semiotic heterogeneity can also be reduced. In addition. statistical heterogeneity needs to be considered, since differences in statistics may occur within data from an industrial automation system. Following this, heterogeneous data within this survey covers semantical, syntactical, or statistical heterogeneous data with a focus on semantical and syntactical heterogeneous data over the introduced six data source dimensions.

Following Jirkovsky and Obitko (2014), semantical heterogeneity can be further split into a difference in coverage, a difference in granularity, and a difference in perspective. The difference in coverage means two data sources describe different regions at the same level of detail. Two data sources describing the same region from the same perspective in different levels of detail have a difference in granularity. And the difference in perspective (or difference in scope) happens, when two data sources "*describe the same region* […], *at the same level of detail, but from different perspective*" (Jirkovsky and Obitko, 2014).

Following our previous work (Kamm et al., 2021) and the introduced definition of heterogeneous data, different challenges arise on machine learning based analysis of heterogeneous data in industrial automation, which will be discussed in the following sub-chapters.

1.1. Challenge group 1: data integration, storage, and description

As already discussed, data in industrial automation arise in different data sources (e.g. system parameters, environmental sensors, text documents, etc.), which results in a large semantic heterogeneity (Hildebrandt et al., 2020). Since data analysis algorithms consume homogeneous data in general, the data must be organized before being analyzed (Desai, 2018). Due to this huge heterogeneity of data and thus also data formats, fixed data integration, and database interfaces are not sufficient (Faul et al., 2016). After integration, the heterogeneous data needs to be managed with its often complex relations in a data storage system (usually a database system), where classical relational database systems (SQL databases) reach their limits for large heterogeneity and can not map the relationships (Henkel et al., 2015; Yoon et al., 2017). Another challenge is the different meanings and interpretations of data values and labels among different data sources, which needs to be considered (L'heureux et al., 2017). A formal semantic data description is necessary to reduce or even resolve this semantic heterogeneity of the data and ensure a common understanding (Jirkovsky and Obitko, 2014; Sahlab et al., 2021), which is mandatory to enable data analysis in a later stage. These challenges which arise when trying to reduce mainly the semantic heterogeneity are clustered in challenge group 1 related to data integration, storag and description. Based on the integrated, stored and described data, analysis algorithms can be applied in a later stage.

1.2. Challenge group 2: complex data analysis for the heterogeneous data

Datasets in industrial automation are often imbalanced, which is one challenge naturally arises when trying to analyze the data (Dai et al., 2020a). For instance, a failure classification dataset usually comprises more samples from the "good" class than samples from the "failure" class, because, in an industrial setting, failures are seldom taking place. This is referred to as statistical heterogeneity. An algorithm trained with an imbalanced dataset may tend to predict the majority class mainly,

because it was seen more often during training and the overall accuracy is still good. This needs to be considered when developing and evaluating analysis algorithms for heterogeneous data. There are specialized and very successful machine learning models for specific data modalities and tasks, such as object detection in images (Maschler et al., 2020), anomaly detection (Lindemann et al., 2020), or failure classification based on time-series data (Kamm et al., 2022a). However, these classical machine learning algorithms don't use the often given variety (Wilcke et al., 2017; Damoulas and Girolami, 2009). Because of their structure, they are designed to analyze one modality of data. Different data sources can be covered, but just when they are coming from the same modality. However, heterogeneous data from different modalities (e.g. images and time-series data), which are syntactical and semantical heterogeneous typically have a difference in perspective. The different modalities from different data sources have different scopes and thus can complement the information from another source to give a better view of the overall situation. For instance, a ultrasonic sensor can detect an object with its time-series signal, while a camera allows to further classify the object and thus judge, if the object is critical or not for a mobile robot. Therefore, fusing different data sources can bring advantages, although it brings new challenges for the analysis algorithms. When building machine learning models (e.g. deep neural networks), which are capable of analyzing heterogeneous data, another challenge arises: During the lifetime of an industrial automation system, it is expected, that new data sources are added or an existing sensor is failing or defective (Kamm et al., 2022b). Thus, it is crucial to ensure an adaptive and robust model, which is capable to handle newly added data sources and faulty/error-prone or missing data sources. The mentioned challenges can be grouped within challenge group 2 – the complex data analysis for heterogeneous data.

1.3. Challenge group 3: integration of existing knowledge into data-driven models

Since the required amount of data is growing tremendously with bigger (regarding the number of parameters) and more complex deep learning models, another challenge is to reduce the amount of desired training data. Therefore, the selection of the correct training data is crucial (Karpatne et al., 2017), since the models simply learn the parameter space of the training data and usually perform poorly outside of it, thus the exploration capability of the models is poor (Karpatne et al., 2017; Raissi et al., 2017). Furthermore, there is huge expertize for (the diverse) industrial applications, e.g. utilizing analytical models designed by experts. However, the existing knowledge is often not incorporated in data-driven applications, such as deep learning models (Raissi et al., 2017). Consequently, the hybrid (or gray-box) modeling combining data-driven and knowledge-driven approaches tackles these challenges by trying to use the advantages of both approaches. Incorporating existing knowledge may also be relevant for the noise dimension, when different noise is in different data sources, which may can be modeled by experts. This can be used to remove the noise before processing the data in data-driven model. Structure of data and also relevance may also be dimensions, where existing knowledge can be helpful for data-driven algorithms. However, overall many challenges arise when the often existing knowledge shall be integrated in a structured way into data-driven models, which is grouped within challenge group 3.

There are already related surveys in the field of data analytics for industrial automation that focus on the general topic of big data, where a minor aspect is the variety, while the main focus is often on the volume and velocity of big data. In Dai et al. (2020a), big data analytics for the manufacturing internet of things is discussed, where the big data lifecycle for manufacturing is divided into three consecutive stages: data acquisition, data preprocessing and storage, and data analytics. The heterogeneity of data is discussed for the first two stages (data acquisition and data preprocessing and storage) but is not further discussed for data analytics. In Munappy et al. (2019), different data management challenges for deep learning are identified based on expert interviews and literature reviews, where one aspect is heterogeneity in the data. However, this point is not further discussed there. Data analytics for big data with deep learning is the focus of Zhang et al., (2018). The authors discuss special deep learning models for heterogeneous data, the so-called multi-modal deep learning models since they cover different modalities of data (e.g., audio and video). In L'heureux et al. (2017), challenges for machine learning with big data are discussed, where one aspect is data heterogeneity. They classify the heterogeneity into different groups and discuss possible machine learning approaches to address the challenges mentioned in the publication. According to them, Deep Learning, Transfer Learning, and Lifelong Learning are addressing the challenges for data analysis coming with heterogeneous data which are discussed within the publication.

Within the mentioned surveys, some of the relevant aspects are discussed and addressed at least partially. However, none of them fully covers all challenge groups since they have another scope. Therefore, this publication aims to address the three mentioned challenge groups comprehensively to give a holistic overview of the existing literature in this field. The remaining paper is structured as follows: Chapter 2 explains the approach of the conducted systematic literature review and chapter 3 shows the quantitative review results. Subsequently, requirements are derived based on the identified challenges above and the relevant publications are evaluated and discussed concerning the mentioned requirements in chapter 4. Finally, chapter 5 provides a conclusion and outlook including the upcoming research challenges.

2. Methodology of the systematic literature review

To obtain a structured overview of the research field, we performed a systematic literature review (SLR). These reviews follow a defined methodology to reduce potential bias in the results of the review. This SLR follows the basic methodology from Kitchenham and Charters, (2007) and (Xiao and Watson (2019) to perform SLRs in software engineering. A popular guideline for SLRs is the PRISMA statement (Page et al., 2021). It defines a checklist and a set of methods to first develop and validate the review protocol as the basis for the SLR. An SLR can be split into three parts: planning the review, conducting the review, and finally reporting the review. Within each part, different sub-steps are mandatory. In the following, the planning phase is described in detail, as it serves as the base for the SLR. Following Kitchenham and Charters (2007), first, the *identification of the need for a review* and the *commissioning of a review* needs to be done. The need for a review is introduced in chapter 1 and also the commissioning is covered there. Following that, the *research question* shall be specified, the *review protocol* has to be *developed* and *evaluated*. Next, the research question is derived and introduced.

2.1. Research question

This SLR aims to analyze the current status of machine learning based analysis of heterogeneous data in industrial automation (including manufacturing). Therefore, existing research in this field is surveyed to identify gaps and opportunities for future work, so that the main research question guiding this study reads as:

How can heterogeneous data be analyzed by means of machine learning in the industrial automation domain?

Following Goodfellow et al. (2018), a machine learning algorithm has four components: The model, the dataset, the optimizer, and a cost function. From our point of view, the model and dataset are the basis for a machine learning algorithm, and an optimizer and a cost function are selected accordingly. Therefore, the model and the dataset are in the focus of this review, yielding specific sub-research questions for these two aspects. Since the used data modalities and sources directly influence the selected model, they should also be investigated by this review.

Thus, separate sub-research questions for the data modalities and data sources are mandatory. Furthermore, the bibliometric key facts and the application field are taken into account, to investigate the temporal trend in this field and also see which applications are mostly covered by the approaches. Thus, the respective sub-research questions that should be answered by this SLR are derived and given in Table 1.

RQ1 provides an overview of the bibliometric key facts of the publications. Therefore, a possible trend over time can be investigated for this topic. RQ2 explores the data modalities and data sources, which are used for data analytics. RQ3 aims to provide an overview of existing datasets which are used to evaluate the approaches. The specific algorithms applied in the publications are presented in RQ4. This question shall result in an understanding, of which approaches are possibly suitable and how they were applied so far. Finally, RQ5 examines, which application scenarios in the industrial automation domain are addressed by the publications.

Based on this research question (including its sub-research questions), the final review protocol was developed and evaluated. The different relevant aspects of the review protocol are introduced and discussed in the following sections.

2.2. Data sources and search strategy

Web of Science, ScienceDirect, and IEEE Xplore are used as scientific databases for this SLR. To identify a meaningful keyword combination in terms of a search string, three groups of terms are defined: Algorithm group, data structures, and application domain. Overall 14 different combinations of keywords within the groups were tested until the final keywords were defined. We used VOSviewer (van Eck and Waltman, 2010) in this process to visualize potential clusters within the publications and adjust the search string correspondingly and also took care, that the number of results is feasible. Thereby, one combination retrieved around 16.000 results and thus had to be discarded. The final keyword combination is given in Table 2, where the different groups are combined with the *AND* operator. In addition, we excluded the medical domain by adding "*NOT (Medic* OR Patient)*", since a big medical cluster was identified with VosViewer, which is not relevant for this SLR.

As the SLR focuses on machine learning based data analysis, *machine learning* and *deep learning* were used as keywords within the algorithm group. In addition, *neural network* was added as a popular realization within deep learning. In the group data structures, *heterogeneous data* was used. Since heterogeneous data is often unstructured, too, *unstructured data* was added. Heterogeneous data stem from multiple data sources and often in multiple modalities, so *multisource data* and *multimodal data* were also included. The application domain is defined as industrial automation. Since related terms, such as industry or industry automation or automation are also often used, *Industr* Automat** was defined, where * can be any ending of the word. The search was conducted on the title, abstract, and keywords of published research articles (conference papers, journal papers, and books or chapters in a book). Table 3 summarizes the defined search settings for the SLR.

Table 4 shows the defined inclusion and exclusion criteria for conducting the SLR. We used four inclusion criteria and overall five exclusion criteria, which are split into two exclusion criteria for titles and three exclusion criteria for abstracts.

Table 1
Sub-research questions.

Nr.	Sub-Research Questions
RQ1	What are the bibliometric key facts of the identified publications?
RQ2	Which data modalities and sources are used for machine learning based data analysis?
RQ3	Which (public) dataset was used for the evaluation?
RQ4	What are the algorithms used within the presented approach?
RQ5	What are the application scenarios for analyzing heterogeneous data in industrial automation?

Table 2
Search string.

Algorithm Group	Data Structures	Application Domain
"Machine Learning" OR "Deep Learning" OR "Neural Network"	"Heterogeneous Data" OR "Unstructured Data" OR "Multisource Data" OR "Multimodal Data"	Industr* Automat*

Table 3
Overview of the search settings.

Search Settings	
Scientific Database	Web of Science, IEEE Xplore, ScienceDirect
Search Space	Title, Abstract, Keywords
Type of publication	Conference Papers, Journal Papers, Books, chapters in a book

Table 4
Inclusion and Exclusion Criteria during Conduction of the SLR.

Inclusion and Exclusion Criteria	
Inclusion Criteria	IC-1: Terms fulfill the search string IC-2: Conference Papers, Journal Papers, Books, or chapters in a book IC-3: Papers written in English IC-4: Publication date: 01.01.2015–31.12.2021
Exclusion Criteria for titles	EC-1.1: Application domain: hospitality, leisure, sport, tourism, agriculture, multidisciplinary, business, telecommunications EC-1.2: Papers that were present already in some other database (Duplicates)
Exclusion Criteria for abstracts	EC-2.1: Application domain is not industrial automation EC-2.2: No obvious usage of machine learning algorithms EC-2.3: Kind of heterogeneous data (e.g. only one data source)

After the review protocol was developed and evaluated, the SLR was conducted using the protocol in multiple steps. Fig. 2 visualizes the reduction process of the records retrieved through the defined search strings from the scientific databases as described above.

The publications were included based on bibliometric properties, in our case the publication time (papers published from 01.01.2015 until 31.12.2021 are reviewed). The search was conducted, where 188 publications were found at Web of Science, 114 at IEEE Xplore, and 108 at ScienceDirect (410 in total). 21 publications were already excluded because of their given application domain. Here, we additionally filtered out Hospitality, Leisure, Sport, Tourism, Agriculture, Multidisciplinary, Business, and Telecommunications based on their tags in the scientific databases. Then 30 duplicates were removed and a title screening of the remaining 359 publications was performed. In this step, the publications are excluded, if the application domain or type of the publication is not fitting (here surveys are excluded). 230 were filtered out in this step. After this, abstract screening was done for the remaining 129

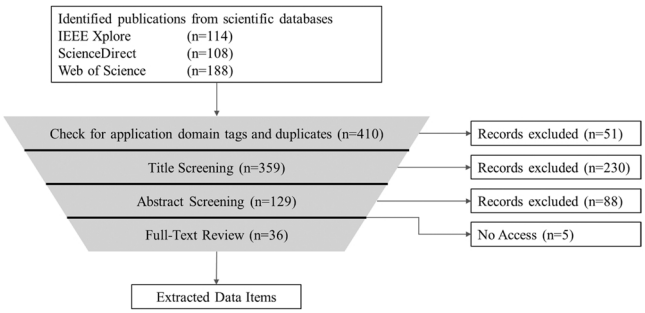

Fig. 2. Overview of the reduction process within the SLR.

publications and 88 were further filtered based on their application domain, algorithms usage (sort of algorithms), or usage of heterogeneous data. After that, 41 full papers remained and no access was given to five of them. Finally, 36 publications were reviewed and relevant data items were extracted for the detailed review. The titles and abstracts were screened by two reviewers independently. If there was no common classification (accepted or rejected) among them, a discussion was conducted to come to a decision. The full papers were read by one reviewer and the defined data fields are extracted and specific criteria were judged.

2.3. Data extraction

To answer the research questions RQ1 – RQ5 (see Table 1), different data items are extracted from the publications. The data items are summarized in Table 5.

With *D1*, the distribution of relevant publications over the years shall be traced to reveal current trends and answer RQ1. *D2* and *D3* are related to RQ2, to investigate which data modality (time-series, images, parameters, etc.) and data sources (e.g. temperature sensors or machine parameters) are analyzed. *D4* answers which dataset is used for evaluation, for instance, which public dataset is used, or if an own private dataset was created. *D5* and *D6* describe the methods and algorithms which are applied and the unique features described in the publication to answer RQ4. To investigate the application scenarios, *D7* and *D8* give information about the application field and the use case covered in the publication.

2.4. Limitations of the review

The review was conducted systematically following the previously mentioned schema. However, there are still possible limitations of the review. One threat of every review is the possible incompleteness of the conducted review, since relevant publications may not be found with the identified search string due to different wording or availability of the publications within the searched scientific databases. In addition, five of the identified 41 publications for the full-text review could neither be accessed, due to restricted access, nor were provided by the contacted authors. Furthermore, the survey was conducted for publications until 31.12.2021. It is expected that new relevant papers are published after this date, which are not covered in this survey. Thus, in the future, a delta search should be conducted in order to show the developments in this field.

3. Quantitative review results

In this section, the quantitative review results regarding the sub-research questions are shown and discussed. As mentioned, 410 overall publications are included in the SLR. For 359, the title screening was performed, where 230 were filtered out. For the remaining 129, abstract screening was conducted and another 93 publications were discarded. Finally, 36 publications constitute the basis for the following results to answer the sub-research questions separately and finally the overall research question. The facts are conducted with the help of the defined data items.

Table 5
Data Items extracted from each publication.

Number	Item Name	Description	Relevant RQ
D1	Publication Year	In which year was the article published?	RQ1
D2	Data Modality	Which data modality is analyzed?	RQ2
D3	Data Sources	Which data sources are used for data analysis?	RQ2
D4	Dataset	Which dataset is used for the evaluation?	RQ3
D5	Methods & Algorithms	Which methods and algorithms are applied?	RQ4
D6	Unique Feature	Which unique feature is mentioned?	RQ4
D7	Application Field	Which application field is covered?	RQ5
D8	Use Case	Which use case is covered?	RQ5

3.1. RQ1: bibliometrical key facts

First, *D1 (Publication Year)* is extracted to analyze the bibliometric key facts. Here also the five publications without access were used since the bibliometrical key facts could be extracted even without access. In Fig. 3, the distribution of the publications over the years within the included time horizon is drawn.

*A clear trend is seen, where most publications (*ca. *68 % of the 41 publications) are published in 2020 and 2021. More and more articles are published related to machine learning based data analysis of heterogeneous data in the industrial automation domain, which shows the increasing significance of the topic and also proves the relevance of this contribution.*

3.2. RQ2: data modalities and data sources

D2 (Data Modality) and *D3 (Data Sources)* give an insight into which data modalities and sources are used for machine learning based data analysis. We categorize the data modalities for *D2* into different categories. In the investigated publications, the following categories were identified: *image data*, *time-series data*, *single sensor signals* (which are not used as time-series), *parameters* of a system that do not vary during runtime including metadata (describing data and information), (non-numerical) *tags* including diagnosis data (e.g. diagnosis codes), *text*, and *radar*. Fig. 4 shows the distribution of the analyzed data modalities within the investigated publications.

Mostly, time-series data and image data are analyzed within the publications. Time-Series data is essential to describe a system property over time, which is of huge interest for industrial automation systems. A specific value of a sensor can have a lot of different meanings, depending on the circumstancing values. This behavior can best be described in a time-series. For this reason, time-series data and time-series prediction or classification are of great interest within the domain. In addition, image data occurs in a wide range of applications (e.g. object detection or product quality inspection). This and the great progress and potential of data-driven approaches for image processing lead to an increasing interest in using image data for analysis in different applications. Further data modalities are also used, but just with a small fraction. So future focus shall be on time-series and image data for industrial automation.

Fig. 5 *shows the distribution of the number of analyzed data modalities within the publications. Most publications (about 63 %) deal with heterogeneous data from one modality (mostly time-series data, as seen in* Fig. 4*).*

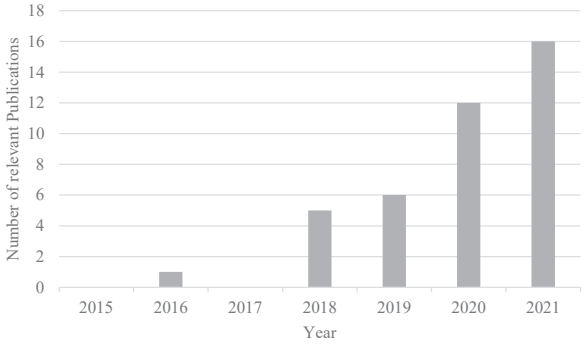

Fig. 3. Number of identified relevant publications over the years.

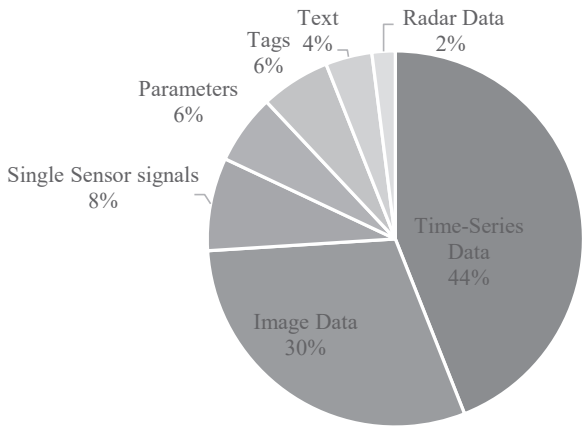

Fig. 4. Distribution of analyzed data modalities.

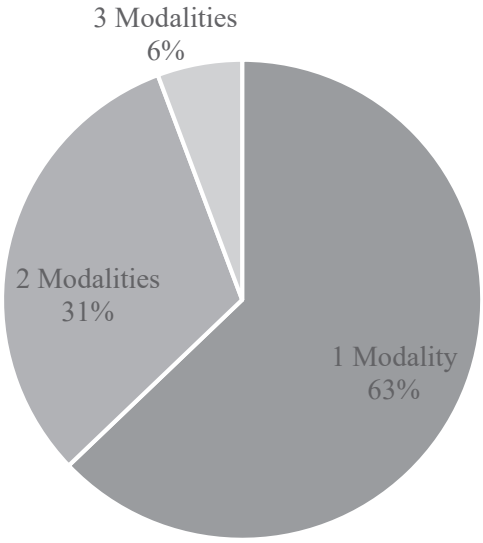

Fig. 5. Distribution of the number of analyzed data modalities.

However, there are also overall 37 % of publications dealing with two or three modalities.

3.3. RQ3: (public) datasets

D4 (Dataset) extracts the information, which dataset was used within the publication for evaluating the proposed approach. Overall, 45 different datasets were used within the full publications. Only two of them were used multiple times over the publications, which are: The bearing dataset from Case Western Reserve University (Case School of Engineering, 2022) (five times) and the LMT haptic texture database (two times) (Strese et al., 2016). These two datasets are shortly described in the following:

Bearing dataset from Case Western Reserve University (Case School of Engineering, 2022):

The bearing dataset from Case Western Reserve University provides ball bearing test data. Good and faulty bearings are included. Different single-point faults are contained in the data with five fault diameters and three possible fault depths at three different locations (inner raceway, outer raceway, and ball) for bearings from two different manufacturers. Drive end and fan end bearings are used within the set-up. During testing, the vibration data was collected with accelerometers. In addition, speed and horsepower data were collected. All signals were recorded over time, resulting in a time-series data modality. For more details about the setup and the data, we refer to the dataset website: https://engineering.case.edu/bearingdatacenter.

LMT haptic texture database (Strese et al., 2016):

The LMT haptic texture database contains multi-modal signals for the material classification of overall 69 different surfaces. Everyday materials (e.g. stones or fabrics) are used as categories, and multiple surfaces within the categories are contained in each category. Acceleration, friction force, and sound signals as well as surface images are recorded for each surface. The current dataset including comments and further explanations can be found on the corresponding website: https://zeus.lmt.ei.tum.de/downloads/texture/.

Further used datasets were publicly available datasets (for instance Target Detection Dataset (Nabritt et al., 2015), Air Compressor Dataset (Verma et al., 2015) or MIR Flickr Dataset (Huiskes and Lew, 2008)), custom datasets from industry (e.g. quality inspection data for the production of microfluidic chips (Baghbanpourasl et al., 2019)), own simulated data (e.g. power system simulation dataset for fault diagnosis (Chen et al., 2021a)) or own setups for data generation (e.g. motor test bench system (Wang et al., 2019), set up for optical tracking (Dai et al., 2020b).

Overall, there exists no commonly applied dataset for analyzing heterogeneous data within the industrial automation domain to properly compare and evaluate results. Therefore, there is a need to further define a dataset that covers the needs (e.g. different data modalities) and with it gives the chance to properly apply and compare newly developed approaches.

3.4. RQ4: applied algorithms

With the analysis of *D5 (Methods & Algorithms)* and *D6 (Unique Feature)*, the different applied algorithms and methods shall be identified, which are used to analyze heterogeneous data within the publications. Fig. 6 shows the number of uses for algorithms and methods which were applied at least two times within the publications. 18 other approaches are applied once, which are summarized as “others” in the given figure.

As can be seen, neural network based approaches are mostly applied. Different architectures of neural networks, such as convolutional neural networks (CNNs), deep neural networks with fully connected layers (DNNs), autoencoders (AEs) with their variants, generative adversarial networks (GANs), long-short term memory (LSTM) networks, or deep belief networks (DBNs) are applied, where CNNs are mostly used. CNNs are often applied to images (2D-CNN) or time-series data (1D-CNN), which are the most occurring data modalities, as shown in Fig. 4. *Ensemble Learning and support vector machines are two non-neural network based algorithms and methods in this overview, which are applied four and two times.* A more detailed discussion of the models and algorithms will be done in chapter 4.

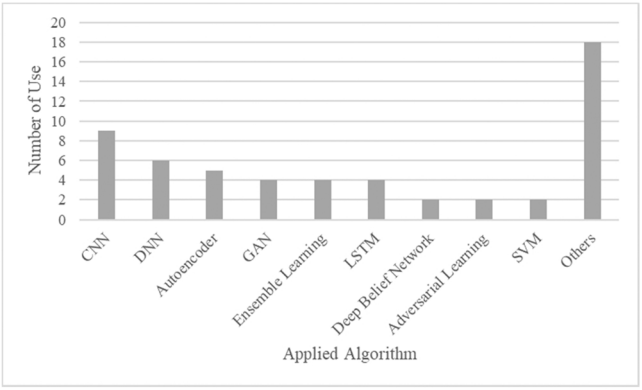

Fig. 6. Number of Algorithm Usage within full publications.

3.5. RQ5: application scenarios and use cases

In *D7 (Application Field)* and *D8 (Use Case)*, the application field and use cases are documented. The classification is done in a bottom-up approach, where based on the data fields *D7* and *D8*, classification and grouping were performed. The application field is, for instance, robotics or manufacturing while the use case can be fault diagnosis, condition monitoring, etc. 17 different application fields were identified, where twelve were covered only once ("others"). These twelve were for instance a wastewater treatment plant, human activity detection, a magnesium melting process, or general object detection. The remaining five were mentioned multiple times, and their distribution is shown in Fig. 7. For better readability, only those application fields that occurred more than once are visualized individually.

As Fig. 7 shows, manufacturing is the dominant application field as it is chosen in 14 publications. Manufacturing covers all fields, where the machinery for the production or a resulting product is in focus. Within such a manufacturing application, typically a variety of data is occurring due to multiple mounted sensors, e.g. *a camera that captures a final product or sensors which measure the environment, such as temperature, humidity, or acoustic sensors. Bearings are also commonly used with its publicly available dataset from Case Western Reserve University. Further, the analysis of heterogeneous data is applied four times to robotics, where e.g. the environment is sensed with the help of multiple sensors (for instance radar, camera, or lidar). Motor diagnosis and general computer vision tasks are other application fields covered two times. Some application fields, such as robotics or computer vision could be part of the manufacturing application but are not part of a manufacturing application within the identified publications.* Within these 17 application fields, different use cases are realized.

19 use cases were identified, whereas eleven of them are covered only once (grouped in "others"). Quality control, sensor defect detection, or object recognition are examples. Fig. 8 shows the different use cases with more than one occurrence.

The main use case, where heterogeneous data is analyzed is fault diagnosis, with 15 occurrences. Condition monitoring follows with five nominations. The further use cases have two nominations, which are e.g. *soft sensor modeling or anomaly detection.*

Following these two observations, fault diagnosis and condition monitoring within manufacturing, bearings or robotics are the main applications for analyzing heterogeneous data within the industrial automation domain.

4. Introduction of the requirements and SLR results

To enable a machine learning based analysis of heterogeneous data in industrial automation, the previously introduced challenge groups (CG1 – CG3, see chapter 1) need to be addressed. To identify the research gap, different requirements (**R**) were derived based on the challenge groups. They are briefly introduced in the following and serve as a basis for the subsequent evaluation of the approaches reviewed within the SLR. The respective challenge group for the individual requirements is given in brackets.

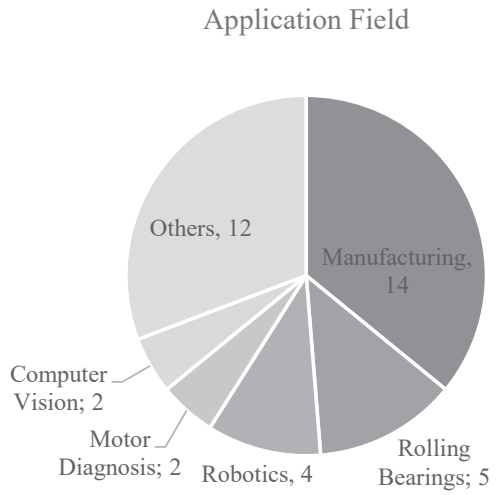

Fig. 7. Distribution of application fields with more than one occurrence.

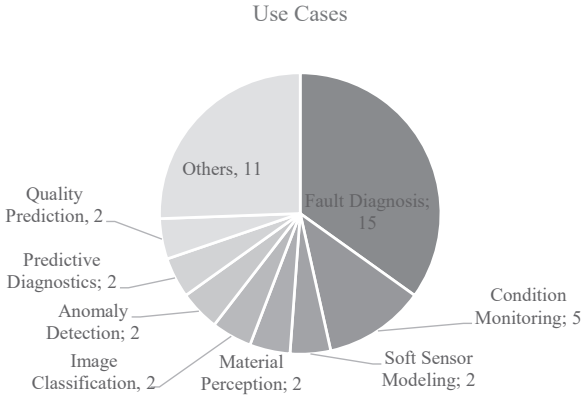

Fig. 8. Distribution of use cases with more than one occurrence.

R1 The heterogeneous data with different syntax and semantics shall be integrated, where the interfaces for the data integration shall be flexible for existing data sources. In addition, the data integration shall be extendable, when new data sources are added to the system. (CG1)

R2 Once the data is integrated, central data access in a unique, machine-readable format preserving the existing information shall be given. Within that, the known relations within the data shall be modeled and the data shall be stored with a unique semantic. (CG1)

R3 The data variety shall be used for the data analysis with machine learning algorithms, so the heterogeneous data with its variety shall be exploratively evaluated and intelligently combined. (CG2)

R4 The developed machine learning model shall be robust and adaptive, to address dynamic changes during runtime. In detail, the model shall perform despite possible failures concerning data sources or data integration (robust) and shall be extendable if new data sources are added. (CG2)

R5 Existing knowledge shall be integrated within the data analysis with machine learning models. Therefore, existing analytical knowledge (e.g. physical equations) shall be usable as well as expert knowledge. (CG3)

The investigated publications were checked concerning their fulfillment of the defined requirements. We classified the publications for each requirement into three levels (○ – not fulfilling the requirement or just discussing the aspect but not covering it within their approach,◑ – partially fulfilling the requirement and covering aspects of it within the approach,● – Fully covering the requirement within their approach). The overview of all publications and their classification can be found in Table 6.

In the following chapters, the most relevant publications for the areas of *integration and storage of heterogeneous data* (R1 and R2, chapter 4.1), *analysis of heterogeneous data with machine learning* (R3, chapter 4.2), *adaptiveness and robustness of the model* (R4, chapter 4.3) and the *integration of existing knowledge* (R5, chapter 4.4) are introduced with a focus on the specific methods which are used. Following that, the results of the different sub-chapters are discussed in chapter 4.5.

Table 6
Analysis Results of the SLR.

Publication	R1	R2	R3	R4	R5
(Wang et al., 2019)	○	○	●	○	○
(Hu et al., 2018)	○	○	◑	○	○
(Zheng et al., 2020)	○	○	○	○	○
(Zhu, 2022)	◑	◑	●	○	●
(Yan et al., 2020)	○	○	○	○	○
(Jayaratne et al., 2019)	○	◑	●	○	◑
(Michau et al., 2020)	○	○	●	○	◑
(Yan et al., 2019)	○	○	●	○	○
(Chunfeng et al., 2018)	○	○	○	○	○
(Liu et al., 2019)	○	○	○	○	○
(Zheng et al., 2019)	○	○	●	○	○
(Verma et al., 2018)	○	○	◑	◑	◑
(Langenberg et al., 2019)	○	○	●	○	◑
(Baghbanpourasl et al., 2019)	○	○	◑	○	○
(Romeo et al., 2018)	○	○	●	○	○
(Zheng et al., 2021)	●	●	◑	○	◑
(Wang et al., 2021)	○	○	○	○	○
(Li et al., 2021)	○	○	○	○	○
(Hayashi et al., 2021)	○	○	◑	○	○
(Lee et al., 2021)	○	○	◑	○	○
(Dai et al., 2020b)	○	○	○	○	◑
(Zhou et al., 2021a)	○	○	●	○	○
(Kebisek et al., 2020)	●	●	◑	○	○
(Li and Li, 2020)	○	○	●	◑	○
(Liu et al., 2020)	○	○	◑	○	○
(Wu et al., 2019)	○	○	◑	○	○
(Zhou et al., 2021b)	○	○	●	○	○
(Hsu et al., 2019)	○	○	●	○	○
(Beyca et al., 2015)	○	○	◑	○	○
(Tang et al., 2020)	○	○	◑	○	◑
(Wei et al., 2021)	○	○	◑	○	○
(Xu et al., 2021)	○	○	◑	○	○
(Roheda et al., 2020)	○	○	●	◑	○
(Yan et al., 2021)	○	○	●	◑	◑
(Chen et al., 2021a)	●	◑	◑	○	○
(Bai et al., 2021)	◑	○	●	○	◑

4.1. Integration and management of heterogeneous data

As introduced, R1 and R2 define the need for special methods for *heterogeneous data integration* and *storage*, to enable a unique, machine-readable format for further analysis of the data. Publications that aim for analyzing heterogeneous data in the industrial automation domain, often poorly address these issues, although especially in this domain there is a huge demand for it. Many machines, devices, or (software) systems are not directly linked to a database or data analytics system. Nevertheless, to analyze the data of those systems an infrastructure for integrating, storing, and managing this data coming from a variety of sources is mandatory.

Only six of the reviewed publications cover the aspects of integrating, storing, and managing heterogeneous data, with four publications covering (at least partially) both requirements. In the following, these six publications will be shortly introduced concerning their fulfillment of requirements 1 and 2.

The authors of (Zhu, 2022) present an overall concept for a big data oriented smart tool condition monitoring system, where a CNC machining center and tool monitoring unit are utilized as the research object. Within this machining center, different machine components generate a variety of data structures, thus a huge number of heterogeneous data exists (1D signals – e.g. force, vibration, acoustics, 2D images, possible 3D point clouds, and textual process data). OPC UA and the MTConnect protocol are adapted to solve the problem of multi-source heterogeneous data collection. The acquired data is then stored in the database of the overall machine tool system. This enables data analytics for the tool monitoring of a CNC machining center. With that, both requirements are partially fulfilled, since no further information about the adaptation of OPC UA and the MTConnect protocol or the database for storing the heterogeneous data is given.

In (Jayaratne et al., 2019), only R2 regarding data storage and management is partially fulfilled, since they included Apache Hadoop into their architecture to parallelize and thus speed up the process of accessing the stored data and analyzing it further. They mainly used Apache Hadoop for the interface to the data analytics algorithms with the MapReduce Algorithm and not taking care of the data storage or integration.

The research provided in (Zheng et al., 2021) covers an approach to fulfill R1 and R2, where heterogeneous data is first integrated into and then stored and managed in an industrial knowledge graph (IKG). Three groups of data are identified (documentation text, IoT sensory data, and image/video) to realize an IKG-based self-configurable manufacturing network. Based on the identified knowledge sources, the relevant entities are recognized and the associated relations and attributes are extracted based on the knowledge sources and available (external) knowledge of the domain. For each knowledge source, a specific extraction method needs to be defined (e.g. object detection algorithm based on a CNN for images). Then the schema is constructed (e.g. in the form of an ontology) to finally build and populate the knowledge graph. This knowledge graph can further be checked for completion and its quality (consistency, validity, conflicts, and uniqueness). Based on this knowledge graph, the analytics algorithm can then use the structured and stored data and query it.

Another approach, which fulfills R1 and R2 is presented in Kebisek et al. (2020). Two main data sources are used there, namely the manufacturing process data and quality measurements for the paint structure, which should later be analyzed. The manufacturing process data in this scenario originates from sensors, IoT devices, PLCs, or other field control level systems. The data is collected using the company's existing field level bus platform, where various protocols, such as MQTT, OPC UA, and ODBC, are provided. Therefore, different production systems can be interconnected to access the data. This acquired data is collected inside an MES. Since the data inside this MES is stored for a maximum of six months, historical manufacturing data is stored inside an independent Hadoop-based Data Lake. The second data source is stored in an internal application database and then exported as.xlsx-files for further usage. All the data is combined into one dataset within the company's discovery platform for further data analytics.

Data from multivariate time-series coming from different sensors are integrated and stored in a graph in Chen et al. (2021a). Each sensor is seen as one node in the graph and the sliced time-series signals are the features for each node. With an interaction learning layer, the explicit graph structure, as well as the interaction relationship between the nodes is learned. Different independent interaction-aware layers are used to learn independent node interactions. This forms a heterogeneous sensor network, where the data is stored and the relationships between the sensors are stored within the learned relations. This can be further used for analyzing the data. However, the storage of the data within the heterogeneous sensor network was specifically designed for further graph analytics algorithms, so there is no general interface for data querying.

In Bai et al. (2021), multi-source information is used for a data-driven prediction of the sinter composition within the ironmaking industry. The data acquisition and preprocessing for temperature and vibration data (time-series) and images are shortly described. Since the data is directly analyzed in this approach, no further data storage or management is performed or discussed. Moreover, there is no concept for the data acquisition, but just the raw data acquisition from the sensors/cameras.

4.2. Analysis of heterogeneous data with machine learning

In this chapter, the most relevant publications with their approaches regarding R3 are discussed. The publications address the data analysis of heterogeneous data with machine learning algorithms. The heterogeneous data with its variety shall be exploratively evaluated and used and

intelligently combined. This chapter answers the main research question which was the baseline for conducting this SLR. Overall 29 publications (partially) fulfill the requirement R3. For brevity, not all of them are discussed in detail. Rather, some selected ones with the highest scoring for R3 will be introduced. The different algorithms are often used in a multi-modal machine learning setup, where multiple algorithms or networks are used to extract the features or decisions for the different modalities.

LSTM networks, CNNs, DBNs, and sparse coding are utilized in Zhu (2022) to extract features from different signals. LSTM networks are used to extract features from force signals over time. This signal was converted into an image by wavelet analysis and features extracted by a CNN. The deep belief network extracts features from vibration signals and sparse coding features from images and processes are extracted. The benefit of the attention method is applied in the feature fusion step followed by a so-called deep connection layer for predicting the desired output. Here, the tool estimates the remaining useful life.

In Yan et al. (2019), the different rotor unbalanced fault conditions are classified based on shaft orbit images and vibration signals. A multi-DBN approach is proposed where multiple DBNs are used to classify the different signals independently to obtain a corresponding result. The output vectors are connected to get a final prediction.

A fault diagnosis for multi-source heterogeneous data by extracting inherent common features is proposed by Zhou et al. 2021a). With the proposed method, a "*comprehensive utilization of heterogeneous data*" shall be achieved. They want to go beyond existing feature fusion approaches, where several networks learn the feature extraction of different data sources independently. Thus, the extracted features are "*based on their own intrinsic information, and it does not make full use of the complementary ability between heterogeneous data*". Therefore, a feature fusion network is designed which uses an alternating optimization algorithm for training. A stacked autoencoder (SAE) is proposed for 1D feature extraction from time-series data and a CNN for feature extraction from 2D images.

A generic multimodal fusion approach with a co-attention mechanism is proposed in Li and Li (2020). The authors argue, that there are still two problems with the feature fusion based on neural networks, although many advantages came up. On the one side, it is hard to build a generic neural network for this task, since the feature representation is quite diverse for different modalities (e.g. images, time-series). In addition, it is challenging to describe the correlation between the different multimodal features in a mathematical form, so it can not be captured and described accurately by a formula. To resolve this, a model is proposed consisting of three parts: the overall model, the basic co-attention model, and the multi-head structure. With that, a generic feature extraction shall be realized. Therefore, the attention mechanism is used to weigh the values of the features before feature fusion. Following that, the multi-head structure projects the inputs into different representation subspaces without increasing the number of parameters. It is worth mentioning, that the current co-attention mechanism only works for two modalities. When more than two modalities shall be fused, multiple overall blocks need to be stacked.

A combination of deep learning and multisource information fusion for the identification of abnormal conditions in a fused magnesium melting process is described in Zhou et al. (2021b). Images of the smelting process are used in combination with data from the smelting process such as current and change rates. The approach uses a CNN for the feature extraction from the images of the smelting condition. A Wasserstein GAN is used to generate synthetic images of the abnormal condition state to avoid overfitting. The trained CNN can classify one abnormal condition and extract features, which are further used in combination with the standardized process data (current and change rates) by an SVM classifier for working condition identification and more detailed classification of abnormal conditions if they exist.

In Roheda et al. (2020), data from sensors with different modalities are learned using a special architecture that tries to learn a structured hidden space between sensors. The hidden space between the sensors is learned by GANs, where one GAN is built for each sensor. Following that, a selection matrix selects the most relevant features from each modality enabling an event-driven fusion. Within the publication, object detection for two datasets is realized, where an event is defined as the existence of an object. The events are defined for each feature independently based on existing expert knowledge (e.g. different speeds of objects are probably caused by different object classes).

An approach to classify tool wear with the help of multimodal signals is outlined in Yan et al. (2021). Thermal images and time-series signals (acoustic, power, and vibration signals) are used. The time-series signal is directly forwarded to the concatenation layer for fusion. In addition, the time-series signals are converted into a 2D time-frequency spectrum by a short-time Fourier transform. These spectra are processed by CNNs, as well as thermal images. The extracted features are fused with the time-series data by a concatenation layer followed by a fully connected layer for final classification.

The authors of Chen et al. (2021a) apply graph neural networks for the fault diagnosis of complex industrial processes, which is evaluated on a three-phase flow facility simulation and a power system simulation dataset. As described in the previous chapter 4.1, the time-series sensor data are integrated into a heterogeneous sensor network with multiple edge types. For each edge type, a subgraph is extracted. Based on each subgraph, a graph convolutional block is applied to extract the node information. Finally, the created graph embeddings are extracted and either simply concatenated or grouped by a weighted summation algorithm. Fully connected layers are then used to finally predict the fault labels.

An LSTM network based on multi-source information for the data-driven prediction of sinter composition is proposed in Bai et al. (2021). Images, vibration, and temperature data are used for the prediction. Features from the images are extracted based on defined image preprocessing algorithms (calculating the average brightness, area, and thickness). The outliers from the vibration and temperature are removed first and then the relevant parameters are selected by the calculation of the Pearson correlation coefficient. The constructed features are then fused in an LSTM network.

4.3. Adaptiveness and robustness of the machine learning model

For the adaptiveness and robustness of the machine learning model (R4), overall only four publications are found in the full publication screening phase, which partially covers the requirement. None of them tackles both, the adaptive and robustness aspect, therefore none of them has full coverage here. The publications partially covering R4 are shortly discussed in the following.

The robustness of different autoencoder variants concerning added noise, which can be seen as a faulty sensor input, is explored in Verma et al. (2018). The noise was added at the input layer during training and shall enforce the autoencoders to learn robust features. A stacked denoising sparse autoencoder is stable in the results of the noisy input experiment. This architecture contains stacked autoencoders with sparsity constraints on each layer where corrupted input was passed to the input.

The robustness of the proposed algorithm is also investigated in Li and Li (2020), where one modality (here image) is masked with an increasing ratio, and the influence on the final accuracy is checked. The multimodal fusion with co-attention mechanism seems to be robust to noise within one modality. Even with a 100 % mask ratio in one modality, the accuracy drops to 71,01 % from the initial 97,84 % (0 % mask ratio) for the MNIST dataset (ten classes) extended to two modalities.

The approach in Roheda et al. (2020) already incorporates robustness to sensor failures within the architecture. First, a damaged sensor can either be detected by cross-sensor tracking, where a global space estimation based on erroneous observations will significantly differ from the estimates for normal observations, so a damaged sensor (anomaly)

can be detected. In addition, hierarchical clustering is proposed to detect damaged sensors. First, clusters in the global hidden space are defined during training with normal operations data. If the cluster score for test data is above or below a threshold, the data is erroneous. If a sensor is detected as damaged, representative features within the learned hidden space can be generated for that damaged sensor. Based on the degree of confidence in these generated features, the contribution of this generated feature to the final decision can be controlled.

Deep transfer learning is proposed in Yan et al. (2021) to transfer knowledge from extracting features in a source domain into the target domain. With that, the training of feature extractors for newly added data sources shall be simplified and existing knowledge shall be reused. Each data source has its feature extractor to extract the features for the following concatenation layer.

4.4. Integrating existing knowledge

In this chapter, the approaches which are covering R5 (at least partially) are shortly introduced. Ten publications address R5.

In Zhu (2022), a physical model for the description of the basic physical functions is implemented to support the data-driven model, since the "*pure-data oriented approach could not meet the high precision demand of micro-milling alone, and the physical model should also be referred for decision*". The physical model as existing knowledge gives a first estimate of the tool state. This estimation is further used in combination with features coming from the process monitoring to perform the data-driven online tool condition monitoring. Thus, this approach fulfills R5.

The authors of Jayaratne et al. (2019) performed manual feature extraction from the raw data to improve the accuracy of the classification. Here, triaxial acceleration jerk and acceleration magnitude is extracted from the acceleration modality. With the engineered features, additional information from the raw data is extracted and incorporated implicitly together with the training data into the algorithm. Knowledge is necessary to extract these features from the raw data. This leads to a partially fulfilled R5 since the knowledge is not directly incorporated into the algorithm, but implicitly with the training data. There are further approaches, where features are extracted from the raw data with the help of expert knowledge, such as in (Michau et al., 2020; Verma et al., 2018; Yan et al., 2021), or Bai et al., (2021).

In Tang et al. (2020), the features are not extracted with the help of a-priori knowledge but grouped into meaningful and related groups based on expert knowledge. Based on this, feature selection is proposed where expert knowledge is incorporated into selecting a proper threshold value for the feature selection to keep the relevant features and neglect redundant information.

Knowledge in the form of feature extractors designed by experts is used in Langenberg et al. (2019) to extract relevant metadata of traffic lights (green or yellow traffic lights, traffic lights with left direction arrow, etc.) and fuse this with available image data.

An industrial Knowledge Graph is built in Zheng et al. (2021), where expert knowledge is incorporated into the creation process of the graph. Experts are necessary to define and extract the relevant attributes and values of the identified knowledge sources. Based on that, an ontology or schema can be constructed, which brings the existing knowledge into a formal specification.

For the optical tracking system based on multiple cameras developed in Dai et al. (2020b), the geometrical principles and the condition state of each camera are included in the method as prior knowledge. With that, it shall be ensured, that just valid and proper information is used for tracking the robot arm to ensure high accuracy.

4.5. Discussion of results

In this sub-chapter, the results shown in the previous sub-chapters 4.1–4.4 are discussed. In chapter 4.1 the approaches for the integration, storage, and management of heterogeneous data are introduced, which relates to requirements R1 and R2. Just a small number of publications address these aspects, with the two most convincing approaches using graph-based approaches for the integration and storage of heterogeneous data. The advantages are, that (complex) relationships within the data coming from multiple sources can be modeled and stored with the data. The data integration is possible through previously defining a data schema, e.g. in the form of an ontology. This leads to ontology-based data integration (OBDI) and access (OBDA), both known from the computer science domain, where the different data sources are modeled semantically and the data can be stored through known interfaces of the data sources. Industrial automation has the specific challenge of complex systems, which have to be modeled within such an ontology. So a full-fledged ontology (often named heavy-weight ontology) may not always be appropriate for an industrial automation system (Müller et al., 2022). Depending on the concrete scenario and expected benefit of a semantically modeled data schema, the trade-off between effort for defining a semantically rich ontology or maybe a more lightweight ontology with lower effort in the development needs to be considered.

Chapter 4.2 addresses requirement R3, with R3 requiring machine learning approaches for the analysis of heterogeneous data. Most of the approaches which are analyzing two or more data modalities are using multi-modal neural network architectures. These approaches can be split into the three classical fusion approaches: early, late, and joint fusion. The early fusion has the advantage, that the features and their correlation are learned within the learning process since the fusion takes place at an early stage within the network. This fusion is often referred to as data or feature fusion. However, the overall network architecture becomes more complex. Late fusion is often denoted as decision-level fusion, where the decisions of the individual previous decision networks are fused. The individual networks can be simple and weak (weak learners), while the ensemble of all networks then boosts the final performance. The networks are independent, so there is no side effect between the networks in case one of the networks is updated or if one of the data sources is failing. However, since the data is independently handled, the correlation between the features is not learned during the training process. Joint fusion combines both approaches, where decisions of single networks are used with the features of other data sources. This approach is rarely used within the reviewed publications. Based on the survey results multi-modal machine learning models constitute a proper choice for analyzing heterogeneous data. Within the computer science domain, a lot of research regarding multi-modal machine learning is ongoing. In addition to open challenges in this field in general, such as representation, translation, alignment, fusion, or co-learning (Liang et al., 2022; Baltrušaitis et al., 2018), we see domain-specific challenges which need to be resolved. This are, for instance, the lack of publicly available (multi-modal) datasets within the domain to further improve the algorithms on common use cases in the domain (such as anomaly detection, failure diagnosis, or failure prognostics). Other domain-specific challenges are the (often) small datasets, since data acquisition within a running system is costly and time-consuming. In use cases like failure classification, highly imbalanced datasets often exist, since a failure class rarely occurs. Furthermore, most computer science approaches focus on classification tasks, whereas regression tasks are rarely covered. However, regression tasks are of interest within the industrial automation domain, e.g. for predicting the remaining useful lifetime of a production machine to enable predictive maintenance.

Chapter 4.3 shows the approaches for the adaptiveness and robustness of the model related to R4. Just a small amount of publications are covering this requirement. Most of them cover added noise to a data source. So there is still data available, but the trustworthiness of the data source is decreased. One approach (Roheda et al., 2020) covers damaged sensors, where the hidden space features of the damaged sensor are replaced by generated values. Therefore, the following network can stay

unchanged (e.g. a classification network). The knowledge for a newly added sensor is transferred utilizing deep transfer learning in (Yan et al., 2021). A network or feature extractor for the new sensor, which might correlate with an existing sensor, can re-use existing knowledge and thus enables the network to become more adaptive. However, there is still a significant need for research to enable adaptive and robust machine learning models within the industrial automation domain. Further approaches from the literature outside the industrial automation domain exist, which for instance use missing value imputation or feature reconstruction to increase the robustness of networks for failed data sources (Lin et al., 2020; Lee et al., 2019; Ma et al., 2021). These approaches can become very complex when for each data source an additional model needs to be trained to impute the missing data of this data source in case of failure. Within an industrial automation system with multiple sensors, the training complexity becomes huge. So further developments for more lightweight algorithms and approaches are necessary. Transfer learning is a promising approach for the dynamic aspect, which is also partially applied within the industrial automation domain, as discussed above. However, these approaches are usually limited in complexity and further elaborated tasks should be investigated as well as the lack of approaches for regression tasks. Overall, more development for the industrial usage of transfer learning is mandatory as discussed in detail in (Maschler and Weyrich, 2021).

The previous chapter 4.4 covers the approaches for incorporating existing knowledge into the machine learning based analysis. Building upon that, four groups are identified, and how existing knowledge can be included. The first group is the *feature extraction group*, where experts of the domain extract meaningful features from the raw data or define rules (e.g. in the form of a mathematical formula) on how the features shall be extracted from the raw data. This should enrich the data with expert knowledge, which is used for the analysis and thus improves the final performance. A second group is the *direct inclusion of analytical equations into the analysis process*. There are multiple interactions possible between the machine learning model and the analytical equations. The analytical equations can be either used within the training process to enforce the machine learning to learn a special behavior or calculate some special parameters for the machine learning model or vice versa (the machine learning model calculates intermediate results as input for the analytical equation). In addition, simulation models can be used to integrate existing analytical knowledge into the machine learning model. The machine learning model can be trained based on the simulation model or even mirror its behavior (surrogate model) and then be fine-tuned on a reduced amount of real-world data. The third identified group is *using knowledge in the form of information models*, e.g. knowledge graphs. This information can be fed as an additional feature into the following analysis task, for instance by applying knowledge graph embeddings to convert the graph structure, which covers knowledge, into a numerical vector. This is beneficial as it enables the application of classical machine learning algorithms since they rely on vectors and matrices as input data.

For the first group, *feature extraction*, multiple approaches are found. However, there is only limited coverage of the other two groups (*direct inclusion of analytical equations into the analysis process* and *knowledge usage in the form of information models*), indicating the need for further research within the industrial automation domain. Analytic equations within the industrial automation domain are often highly complex and hard to model. Current research in using analytical equations and combining them with neural networks (named physics-informed neural networks – PINN) often incorporates simpler equations (e.g. partial differential equations of a known physical process (Raissi et al., 2019); Tod et al., 2021). Further research is mandatory to incorporate complex system-describing analytical equations in a PINN. In addition, within the industrial automation domain, there is a lot of knowledge in the form of experts, which first needs to be formalized to use this knowledge for the analysis, e.g. within an information model such as a knowledge graph. The step to formalize the knowledge and then use it for analysis may be done with the help of knowledge graph embeddings. First studies for the usage of graph embeddings in the industrial automation domain are conducted (Chen et al., 2021b), but further research is necessary to develop methods for formalizing the expert knowledge and using it for analysis.

5. Conclusion & outlook

To examine how comprehensive the stated research question of *how can heterogeneous data be analyzed with the help of machine learning in the industrial automation domain* is answered by the literature between 01.01.2015 and 31.12.2021, a systematic literature review was conducted. The following **key insights** have been derived within the course of our literature review:

- The formulated research question was not sufficiently answered by any single investigated approach.
- The aspects of the analysis of heterogeneous data are the subject of numerous ongoing research activities. Thus, the importance of heterogeneous data analysis is evident.
- Most approaches solely tackle one data modality (e.g. time-series or images). Moreover, in about 75 % of the publications, time-series and/or image data are used. Thus, it is concluded that these two are the most relevant data modalities within the industrial automation domain.
- Multi-modal machine learning models constitute a proper choice for analyzing heterogeneous data.
- Only a limited number of approaches address the adaptiveness and robustness of the machine learning model. Most approaches investigate the effect of noise, but whole sensor failures or newly added sensors are rarely investigated.
- For incorporating existing knowledge into the machine learning based analysis, three groups are identified: Feature extraction with the help of expert knowledge; Incorporating analytical equations into the machine learning algorithm or its training process (e.g. by FE simulations); Utilizing information models as further data sources within the analysis.

For this paper, we assessed the machine learning based analysis of heterogeneous data, taking into account the mandatory data integration, storage, and management as well as relevant aspects for the adaptiveness and robustness of the models. Furthermore, the integration of existing knowledge into the analysis process, which is often available in the industrial automation domain, was examined.

Based on this SLR, we identified some open research challenges, which offer great potential to enable or even improve the machine learning based analysis of heterogeneous data within industrial automation and thus provide added value to its application:

5.1. Research challenge 1 – using multiple data modalities for data analytics

As seen in Fig. 5, most approaches cover only one data modality (e.g. time-series data from different sources). A fraction covers two modalities, while a very limited number covers three modalities. Future research activities should investigate and develop concepts to handle heterogeneous data of two and more data modalities and tackle the introduced specific challenges for the industrial domain. Image and time-series data are the most applied modalities within industrial automation and should be covered. Beyond this, different modalities are relevant, depending on the concrete application and use case.

5.2. Research challenge 2 – adaptiveness and robustness of the machine learning model

In an industrial application, data sources (e.g. sensors) in general can

always fail or become noisy/faulty during their lifetime. Therefore, the analytics model should also be robust to this. Additionally, new data sources can be added over a lifetime. The effort for updating the model should be as low as possible. Thus, the model shall be adaptive. Only a limited amount of approaches tackle this aspect. Further research is necessary on how the used machine learning model can be made more adaptive and robust.

5.3. Research challenge 3 – integrating existing knowledge into the algorithm

Within industrial automation, processes, and devices are engineered and designed with huge expert knowledge. This knowledge is available in different forms (experts, analytical equations, simulation models, etc.). Machine learning algorithms are usually applied to learn complex patterns within the data and thus are applied as data-driven approaches. Mostly none of this available knowledge is incorporated into the algorithms, although this knowledge can be beneficial, e.g. by improving the models' performance, and reducing the training time and the cost of data generation. However, this is still an open research field and should be extensively investigated in the future, to use the benefits of both, the data-driven approaches which discover (unknown) patterns within the data, and also the existing knowledge of the application. These approaches are named in many different ways, such as hybrid or grey-box modeling or (physics-)informed machine learning.

5.4. Research challenge 4 – holistic approach for the analysis of heterogeneous data

Most approaches reviewed in this SLR are covering only one aspect of the previously introduced and mentioned requirements for analyzing heterogeneous data in the industrial automation domain. They focus mainly on applying machine learning algorithms to a specific use case. To enable the application of the developed algorithms in an industrial setting, further components, namely data integration, and storage are necessary. Therefore, further holistic approaches are necessary, which take care of all necessary components to apply a developed algorithm to an industrial use case and gain benefits from applying the algorithm.

Further research activities shall focus on the derived open research challenges which are an outcome of the conducted systematic literature review. The authors currently focus on research challenge 1 and research challenge 2. Therefore, a multi-modal machine learning approach is under development that is adaptive and robust to dynamic changes in the environment. As a following step, existing knowledge shall be integrated (research challenge 3) into this approach and finally a holistic approach for the analysis of heterogeneous data in the industrial automation domain (research challenge 4) is the goal of our future research activities. We encourage readers to tackle the derived research challenges in their research activities to further promote machine learning based analysis of heterogeneous data in industrial automation.

Funding

This work was supported by the German "Bundesministerium für Bildung und Forschung (BMBF)" in the project FA4.0. This project is a common initiative with Czech Republic, French, and Swedish consortia in the European EUREKA Clusters EURIPIDES2 and PENTA.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Baghbanpourasl, A., Lughofer, E., Meyer-Heye, P., Zörrer, H., Eitzinger, C. , Virtual Quality control using bidirectional LSTM networks and gradient boosting. In: Proceedings of the Seventeenth International Conference on Industrial Informatics (INDIN), IEEE, 2019, 1638–1643.

Bai, X. , Chen, C. , Liu, W., Zhang, H., Data-driven prediction of sinter composition based on multi-source information and LSTM network. In: Proceedings of the Fortieth Chinese Control Conference (CCC), 2021, 3311–3316.

Baltrušaitis, T., Ahuja, C., Morency, L.-P., 2018. Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41 (2), 423–443.

Beyca, O.F., Rao, P.K., Kong, Z., Bukkapatnam, S.T.S., Komanduri, R., 2015. Heterogeneous sensor data fusion approach for real-time monitoring in ultraprecision machining (UPM) process using non-parametric Bayesian clustering and evidence theory. IEEE Trans. Autom. Sci. Eng. 13 (2), 1033–1044.

2022 Case School of Engineering, Case Western Reserve University Bearing Data Center (Online). https://engineering.case.edu/bearingdatacenter. (Accessed 22 November 2022) 2022.

Chen, D., Liu, R., Hu, Q., Ding, S.X., 2021a. Interaction-aware graph neural networks for fault diagnosis of complex industrial processes. IEEE Trans. Neural Netw. Learn. Syst.

Chen, Z., Liu, Y., Valera-Medina, A., Robinson, F., 2021b. Multi-sourced modelling for strip breakage using knowledge graph embeddings. Procedia CIRP 104, 1884–1889.

Chunfeng, W., Zheng, L., Jun, Z., Wei, W., 2018. Heterogeneous transfer learning based on stack sparse auto-encoders for fault diagnosis. In: Proceedings of the Chinese Automation Congress (CAC) 4277–4281.

Dai, H., et al., 2020b. Prior knowledge-based optimization method for the reconstruction model of multicamera optical tracking system. IEEE Trans. Autom. Sci. Eng. 17 (4), 2074–2084.

Dai, H.-N., Wang, H., Xu, G., Wan, J., Imran, M., 2020a. Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies. Enterp. Inf. Syst. 14 (9–10), 1279–1303.

Damoulas, T., Girolami, M.A., 2009. Combining feature spaces for classification. Pattern Recognit. 42 (11), 2671–2683.

Desai, P.V., 2018. A survey on big data applications and challenges. In: Proceedings of the Second International Conference on Inventive Communication and Computational Technologies. ICICCT, pp. 737–740.

van Eck, N., Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84 (2), 523–538.

Faul, A., Jazdi, N., Weyrich, M., 2016. Approach to interconnect existing industrial automation systems with the Industrial Internet. In: Proceedings of the IEEE Twenty First International Conference on Emerging Technologies and Factory Automation. ETFA, pp. 1–4.

I. Goodfellow, Y. Bengio, A. Courville, and Safari, an O'Reilly Media Company, Deep Learning - Grundlagen, aktuelle Verfahren und Algorithmen, neue Forschungsansätze: mitp Verlag, 2018. (Online). ⟨https://books.google.de/books?id=uFVRzQEACAAJ⟩.

Hayashi, K., Zheng, W., El Hafi, L., Hagiwara, Y., Taniguchi, T., 2021. Bidirectional generation of object images and positions using deep generative models for service robotics applications. In: Proceedings of the IEEE/SICE International Symposium on System Integration. SII, pp. 325–329.

Henkel, R., Wolkenhauer, O., Waltemath, D., 2015. Combining computational models, semantic annotations and simulation experiments in a graph database. Database 2015.

Hildebrandt, C., et al., 2020. Ontology building for cyber-physical systems: application in the manufacturing domain. IEEE Trans. Autom. Sci. Eng. 17 (3), 1266–1282.

Hsu, Y.-C., Kuo, P.-Y., Huang, W.-S., 2019. A novel feature-spanning machine learning technology for defect inspection. In: Proceedings of the Fourteenth International Microsystems, Packaging, Assembly and Circuits Technology Conference. IMPACT, pp. 54–57.

Hu, G., Li, H., Xia, Y., Luo, L., 2018. A deep Boltzmann machine and multi-grained scanning forest ensemble collaborative method and its application to industrial fault diagnosis. Comput. Ind. 100, 287–296.

Huiskes, M.J., Lew, M.S., 2008. The mir flickr retrieval evaluation. In: Proceedings of the First ACM International Conference on Multimedia Information Retrieval. ACM, pp. 39–43.

Jayaratne, M., de Silva, D., Alahakoon, D., 2019. Unsupervised machine learning based scalable fusion for active perception. IEEE Trans. Autom. Sci. Eng. 16 (4), 1653–1663.

Jirkovsky, V., Obitko, M., 2014. Semantic heterogeneity reduction for big data in industrial automation. ITAT 1214.

Jirkovsky, V., Obitko, M., Mavrik, V., 2016. Understanding data heterogeneity in the context of cyber-physical systems integration. IEEE Trans. Ind. Inform. 13 (2), 660–667.

Kamm, S., Bickelhaupt, S., Sharma, K., Jazdi, N., Kallfass, I., Weyrich, M., 2022a. Simulation-to-reality based transfer learning for the failure analysis of SiC power transistors. In: Proceedings of the IEEE Twenty Seventh International Conference on Emerging Technologies and Factory Automation. ETFA, pp. 1–8.

Kamm, S., Jazdi, N., Weyrich, M., 2021. Knowledge discovery in heterogeneous and unstructured data of industry 4.0 systems: challenges and approaches. Procedia CIRP 104, 975–980.
S. Kamm, N. Sahlab, N. Jazdi, M. Weyrich, 2022b. A concept for dynamic and robust machine learning with contex modeling for heterogeneous manufacturing data, Procedia CIRP.
Karpatne, A., et al., 2017. Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. 29 (10), 2318–2331.
Kebisek, M., Tanuska, P., Spendla, L., Kotianova, J., Strelec, P., 2020. Artificial intelligence platform proposal for paint structure quality prediction within the industry 4.0 concept. IFAC-Pap. 53 (2), 11168–11174.
B. Kitchenham , S. Charters, Guidelines for performing systematic literature reviews in software engineering, Technical Report, ver. 2.3 Ebse Technical Report, ebse, 2007.
L'heureux, A., Grolinger, K., Elyamany, H.F., Capretz, M. am, 2017. Machine learning with big data: challenges and approaches. IEEE Access 5, 7776–7797.
Langenberg, T., Lüddecke, T., Wörgötter, F., 2019. Deep metadata fusion for traffic light to lane assignment. IEEE Robot. Autom. Lett. 4 (2), 973–980.
Lee, J., Qu, S., Kang, Y., Jang, W., 2021. Multimodal machine learning for display panel defect layer identification. In: Proceedings of the Thirty Second Annual SEMI Advanced Semiconductor Manufacturing Conference. ASMC, pp. 1–7.
Lee, M., An, J., Lee, Y., 2019. Missing-value imputation of continuous missing based on deep imputation network using correlations among multiple iot data streams in a smart space. IEICE Trans. Inf. Syst. 102 (2), 289–298.
Li, H., Fan, R., Shi, Q., Du, Z., 2021. Class imbalanced fault diagnosis via combining K-means clustering algorithm with generative adversarial networks. J. Adv. Comput. Intell. Intell. Inform. 25 (3), 346–355.
Li, P., Li, X., 2020. Multimodal fusion with co-attention mechanism. In: Proceedings of the IEEE Twenty Third International Conference on Information Fusion. FUSION, pp. 1–8.
Liang, P.P., Zadeh, A., Morency, L.-P., 2022. Foundations and recent trends in multimodal machine learning: principles, challenges, and open questions. arXiv Prepr. arXiv 2209, 03430.
Lin, J., Li, N., Alam, M.A., Ma, Y., 2020. Data-driven missing data imputation in cluster monitoring system based on deep neural network. Appl. Intell. 50, 860–877.
Lindemann, B., Jazdi, N., Weyrich, M., 2020. Anomaly detection and prediction in discrete manufacturing based on cooperative LSTM networks. In: Proceedings of the IEEE Sixteenth International Conference on Automation Science and Engineering. CASE, pp. 1003–1010.
Liu, Y., Gao, H., Guo, L., Qin, A., Cai, C., You, Z., 2019. A data-flow oriented deep ensemble learning method for real-time surface defect inspection. IEEE Trans. Instrum. Meas. 69 (7), 4681–4691.
Liu, Y., Lv, Z., Zhao, J., Liu, Y., Wang, W., 2020. Scheduling knowledge retrieval based on heterogeneous feature learning for byproduct gas system in steel industry. IFAC-Pap. 53 (2), 11938–11943.
Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X., 2021. Smil: Multimodal learning with severely missing modality. In: Proceedings of the AAAI Conference on Artificial Intelligenc. AAAI, pp. 2302–2310.
Maschler, B., Kamm, S., Jazdi, N., Weyrich, M., 2020. Distributed cooperative deep transfer learning for industrial image recognition. Procedia CIRP 93, 437–442.
Maschler, B., Weyrich, M., 2021. Deep transfer learning for industrial automation: a review and discussion of new techniques for data-driven machine learning. IEEE Ind. Electron. Mag. 15 (2), 65–75.
Michau, G., Hu, Y., Palmé, T., Fink, O., 2020. Feature learning for fault detection in high-dimensional condition monitoring signals. Proc. Inst. Mech. Eng., Part O J. Risk Reliab. 234 (1), 104–115.
Müller, T., et al., 2022. Architecture and knowledge modelling for self-organized reconfiguration management of cyber-physical production systems. Int. J. Comput. Integr. Manuf. 1–22.
Munappy, A., Bosch, J., Olsson, H.H., Arpteg, A., Brinne, B., 2019. Data management challenges for deep learning. In: Proceedings of theForty Fifth Euromicro Conference on Software Engineering and Advanced Applications. SEAA, pp. 140–147.
S.M. Nabritt, T. Damarla, G. Chatters, Personnel and vehicle data collection at aberdeen proving ground (apg) and its distribution for research, Army Research Lab Adelphi MD Sensors and Electron Devices Directorate, 2015.
Page, M.J., et al., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst. Rev. 10 (1), 1–11.
M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations, arXiv preprint arXiv: 1711.10561 (Titel anhand dieser ArXiv-ID in Citavi-Projekt übernehmen), 2017.
Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–707.
Roheda, S., Krim, H., Riggan, B.S., 2020. Robust multi-modal sensor fusion: an adversarial approach. IEEE Sens. J. 21 (2), 1885–1896.
Romeo, L., Paolanti, M., Bocchini, G., Loncarski, J., Frontoni, E., 2018. An innovative design support system for industry 4.0 based on machine learning approaches. In: Proceedings of the Fifth International Symposium on Environment-Friendly Energies and Applications. EFEA, pp. 1–6.
Rowley, J., 2007. The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. 33 (2), 163–180.
Sahlab, N., Kamm, S., Müller, T., Jazdi, N., Weyrich, M., 2021. Knowledge graphs as enhancers of intelligent digital twins. In: Proceedings of the Fourth IEEE International Conference on Industrial Cyber-Physical Systems. ICPS, pp. 19–24.
Strese, M., Schuwerk, C., Iepure, A., Steinbach, E., 2016. Multimodal feature-based surface material classification. IEEE Trans. Haptics 10 (2), 226–239.
Tang, J., Zhang, J., Yu, G., Zhang, W., Yu, W., 2020. Multisource latent feature selective ensemble modeling approach for small-sample high-dimensional process data in applications. IEEE Access 8, 148475–148488.
Tod, G., Ompusunggu, A.P., Struyf, G., Pipeleers, G., de Grave, K., Hostens, E., 2021. Physics-informed neural networks (PINNs) for improving a thermal model in stereolithography applications. Procedia CIRP 104, 1559–1564.
Verma, N.K., Dixit, S., Sevakula, R.K., Salour, A., 2018. Computational framework for machine fault diagnosis with autoencoder variants. In: Proceedings of the International Conference on Sensing, Diagnostics, Prognostics, and Control. SDPC, pp. 353–358.
Verma, N.K., Sevakula, R.K., Dixit, S., Salour, A., 2015. Intelligent condition based monitoring using acoustic signals for air compressors. IEEE Trans. Reliab. 65 (1), 291–309.
Wang, J., Fu, P., Zhang, L., Gao, R.X., Zhao, R., 2019. Multilevel information fusion for induction motor fault diagnosis. IEEE/ASME Trans. Mechatron. 24 (5), 2139–2150.
Wang, K., Guo, Z., Wang, Y., Yuan, X., Yang, C., 2021. Common and specific deep feature representation for multimode process monitoring using a novel variable-wise weighted parallel network. Eng. Appl. Artif. Intell. 104, 104381.
Wang, L., 2017. Heterogeneous data and big data analytics. Autom. Control Inf. Sci. 3 (1), 8–15.
Wei, J., Cui, S., Hu, J., Hao, P., Wang, S., Lou, Z., 2021. Multimodal unknown surface material classification and its application to physical reasoning. IEEE Trans. Ind. Inform. 18 (7), 4406–4416.
Wilcke, X., Bloem, P., de Boer, V., 2017. The knowledge graph as the default data model for learning on heterogeneous knowledge. Data Sci. 1 (1–2), 39–57.
Wu, H., Zhou, B., Zhu, P., Hu, Q., Shi, H., 2019. Multi-task Sparse Regression Metric Learning for Heterogeneous Classificationn. Int. Conf. Artif. Neural Netw. 543–553.
Xiao, Y., Watson, M., 2019. Guidance on conducting a systematic literature review. J. Plan. Educ. Res. 39 (1), 93–112.
Xu, D., Li, Y., Song, Y., Jia, L., Liu, Y., 2021. IFDS: an intelligent fault diagnosis system with multisource unsupervised domain adaptation for different working conditions. IEEE Trans. Instrum. Meas. 70, 1–10.
Yan, J., Hu, Y., Guo, C., 2019. Rotor unbalance fault diagnosis using DBN based on multi-source heterogeneous information fusion. Procedia Manuf. 35, 1184–1189.
Yan, J., Wang, X., Ali, A., 2021. Deep Transfer Learning Based Multi-source Heterogeneous data Fusion with Application to Cross-scenario Tool Wear monitoring. In: Proceedings of the Seventh International Conference on Mechanical Engineering and Automation Science. ICMEAS, pp. 96–101.
Yan, W., Xu, R., Wang, K., Di, T., Jiang, Z., 2020. Soft sensor modeling method based on semisupervised deep learning and its application to wastewater treatment plant. Ind. Eng. Chem. Res. 59 (10), 4589–4601.
Yoon, B.-H., Kim, S.-K., Kim, S.-Y., 2017. Use of graph database for the integration of heterogeneous biological data. Genom. Inform. 15 (1), 19–27.
Zhang, Q., Yang, L.T., Chen, Z., Li, P., 2018. A survey on deep learning for big data. Inf. Fusion 42, 146–157.
Zheng, P., Xia, L., Li, C., Li, X., Liu, B., 2021. Towards Self-X cognitive manufacturing network: an industrial knowledge graph-based multi-agent reinforcement learning approach. J. Manuf. Syst. 61, 16–26.
Zheng, T., Song, L., Wang, J., Teng, W., Xu, X., Ma, C., 2020. Data synthesis using dual discriminator conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearings. Measurement 158, 107741.
Zheng, W., Liu, H., Wang, B., Sun, F., 2019. Cross-modal material perception for novel objects: a deep adversarial learning method. IEEE Trans. Autom. Sci. Eng. 17 (2), 697–707.
Zhou, F., Yang, S., He, Y., Chen, D., Wen, C., 2021a. Fault diagnosis based on deep learning by extracting inherent common feature of multi-source heterogeneous data. Proc. Inst. Mech. Eng., Part I J. Syst. Control Eng. 235 (10), 1858–1872.
Zhou, P., Gao, B., Wang, S., Chai, T., 2021b. Identification of abnormal conditions for fused magnesium melting process based on deep learning and multisource information fusion. IEEE Trans. Ind. Electron. 69 (3), 3017–3026.
Zhu, K., 2022. Big data oriented smart tool condition monitoring system. Smart Machining Systems. Springer,, pp. 361–381.

Simon Kamm, M. Sc. is a research assistant at the Institute of Industrial Automation and Software Engineering at the University of Stuttgart. His research focusses on the analysis of heterogeneous data with machine learning within the industrial automation domain.

Sushma Sri Veekati, B. Tech. is a student assistant at the Institute of Industrial Automation and Software Engineering at the University of Stuttgart. Her work focusses on the management of heterogeneous data.

Timo Müller, M. Sc. is a research assistant at the Institute of Industrial Automation and Software Engineering at the University of Stuttgart. His research focusses on the reconfiguration of cyber-physical production systems.

Dr.-Ing. Nasser Jazdi is the deputy head of the Institute of Industrial Automation and Software Engineering at the University of Stuttgart. His research focusses on the Internet of Things as well as learning ability, reliability, safety and artificial intelligence in industrial automation.

Prof. Dr.-Ing. Dr. h. c. Michael Weyrich teaches at the University of Stuttgart and is head of the Institute of Industrial Automation and Software Engineering. His research focuses on intelligent automation systems, complexity control of cyber-physical systems and validation and verifcation of automation systems.