# Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0

Juan Pablo Usuga Cadavid[1] · Samir Lamouri[1] · Bernard Grabot[2] · Robert Pellerin[3] · Arnaud Fortin[4]

## Abstract
Because of their cross-functional nature in the company, enhancing Production Planning and Control (PPC) functions can lead to a global improvement of manufacturing systems. With the advent of the Industry 4.0 (I4.0), copious availability of data, high-computing power and large storage capacity have made of Machine Learning (ML) approaches an appealing solution to tackle manufacturing challenges. As such, this paper presents a state-of-the-art of ML-aided PPC (ML-PPC) done through a systematic literature review analyzing 93 recent research application articles. This study has two main objectives: contribute to the definition of a methodology to implement ML-PPC and propose a mapping to classify the scientific literature to identify further research perspectives. To achieve the first objective, ML techniques, tools, activities, and data sources which are required to implement a ML-PPC are reviewed. The second objective is developed through the analysis of the use cases and the addressed characteristics of the I4.0. Results suggest that 75% of the possible research domains in ML-PPC are barely explored or not addressed at all. This lack of research originates from two possible causes: firstly, scientific literature rarely considers customer, environmental, and human-in-the-loop aspects when linking ML to PPC. Secondly, recent applications seldom couple PPC to logistics as well as to design of products and processes. Finally, two key pitfalls are identified in the implementation of ML-PPC models: the complexity of using Internet of Things technologies to collect data and the difficulty of updating the ML model to adapt it to the manufacturing system changes.

**Keywords** Machine learning · Industry 4.0 · Smart manufacturing · Production planning and control · State-of-the-art · Systematic literature review

## Introduction

The current manufacturing environment is characterized by high complexity, dynamic production conditions and volatile markets. Additionally, companies must offer customized products while engaging low costs and reducing the time-to-market if they want to remain competitive in a globalized world (Schuh et al. 2017b; Carvajal Soto et al. 2019). This situation poses tremendous challenges for manufacturers who seek to implement new technologies to meet their objectives while expecting a return on investment. Several countries have developed projects that aim to help companies adapt their industries to new production technologies. For instance, Germany created Industry 4.0 (I4.0), the United States proposed the Smart Manufacturing Leadership Coalition, and China introduced the plan called China Manufacturing 2025 (Wang et al. 2018a). This has led to significant financial support for manufacturing research;

✉ Juan Pablo Usuga Cadavid
  juan_pablo.usuga_cadavid@ensam.eu

  Samir Lamouri
  samir.lamouri@ensam.eu

  Bernard Grabot
  bernard.grabot@enit.fr

  Robert Pellerin
  robert.pellerin@polymtl.ca

  Arnaud Fortin
  a.fortin@ifakt.com

[1] LAMIH, Arts et Métiers ParisTech, Paris, France

[2] LGP, INP/ENIT, Tarbes, France

[3] Mathematics and Industrial Engineering Department, Polytechnique de Montreal, Montreal, Canada

[4] iFAKT France SAS, Toulouse, France

for example in the European Union around €7 billion will be invested by 2020 in Factories of the Future (Kusiak 2017).

Among the Industry 4.0 groups of technologies (Ruessmann et al. 2015), Big Data and Analytics (BDA) allows the constantly growing mass of produced data to be harnessed to generate added value. In fact, data generation in modern manufacturing has undergone explosive growth, reaching around 1000 Exabytes per year (Tao et al. 2018). However, the potential of this data has been found to be insufficiently exploited by companies (Manns et al. 2015; Moeuf et al. 2018). As BDA enables the exploitation of data, the scope of this review will focus on this technology, and more specifically ML applied in Production Planning and Control.

In the context of I4.0, Production Planning and Control (PPC) can be defined as the function determining the global quantities to be produced (production plan) to satisfy the commercial plan and to meet the profitability, productivity and delivery time objectives. It also encompasses the control of production process, allowing real-time synchronization of resources as well as product customization (Tony Arnold et al. 2012; Moeuf et al. 2018). In this review, I4.0 is considered a synonym of Smart Manufacturing, as they both refer to technological advances that value data to draw improvements in production. For example, Ruessmann et al. (2015) proposed nine technologies for I4.0 while Kusiak (2019) suggested six, but for Smart Manufacturing. Both proposals tend to refer to similar technologies and variations depend on the authors' focus. Hence, as the PPC is a core function of manufacturing, this paper regards its improvement through I4.0 technologies, namely ML, which concerns BDA. Regarding ML, the definition that will be retained is the one of a computer program capable of learning from experience to improve a performance measure at a given task (Mitchell 1997).

Classical approaches to performing PPC include analytical methods and precise simulations, providing solutions that may rapidly become unfeasible in the execution phase due to the stochastic nature of the production system and uncertainties such as machine breakdowns, scrap rate, delayed deliveries, etc. Moreover, Enterprise Resource Planning (ERP) systems perform poorly at the operative level (Gyulai et al. 2015). To tackle this issue, ML can endow the PPC with the capacity of learning from historical or real-time data to react to predictable and unpredictable events. Even though this may suggest that organizations must invest in data warehousing to handle the mass amount of collected data, studies have reported that enterprises successfully implementing data-driven solutions have experienced a payback of 10–70 times their investment in data warehousing (Rainer 2013).

Having introduced the synergism between ML and PPC, this study aims to provide an analysis of its state-of-the-art through a systematic literature review. This will contribute to the definition of a methodology to implement a ML-PPC

and to the proposal of a map to classify scientific literature. This paper analyzes research produced in the context of the I4.0 and is guided by five research questions:

1. Which are the activities employed to perform a ML-PPC?
2. Which are the techniques and tools used to implement a ML-PPC?
3. Which are the currently harnessed data sources to implement a ML-PPC?
4. Which are the addressed use cases by the recent scientific literature in ML-PPC?
5. Which are the characteristics of the I4.0 targeted by the recent scientific literature in ML-PPC?

The first three questions are related to the first objective of this research. They will contribute to the definition of a methodology to implement a ML-PPC. The last two questions address the second objective, as they will provide the basis to create a classification map.

The reminder of this paper is organized as follows: section "Research methodology and contribution" will explain the systematic literature review methodology employed to search and choose the sample of scientific articles. Additionally, the contribution of this paper with respect to similar studies will be briefly highlighted and a short bibliometric analysis is presented to assess the keywords used as string chains. The "Analytical framework" section will explain the four axes encompassed by the analytical framework. Afterwards, the "Results" section will focus on the results of the systematic literature review and an analysis of it. Finally, the "Conclusion and further research perspectives" section will conclude this study and provide further research perspectives.

## Research methodology and contribution

To meet the two objectives of this study, a systematic literature review was carried out following the method proposed by Tranfield et al. (2003) who extended research methods from the medical sector to the management sciences. This method has been successfully employed by other authors to draw insights from the scientific literature (Garengo et al. 2005; Moeuf et al. 2018). This literature review focuses exclusively on applications of ML in PPC in the context of I4.0.

In another domain, Zhong et al. (2016), proposed a bibliometric analysis of big data applications on different sectors such as healthcare, supply chain, finance, etc. but its focus on manufacturing was limited. (Kusiak 2017; Tao et al. 2018; Wang et al. 2018a) provided a literature analysis of data-driven smart manufacturing, citing representative references.

However, these references were not chosen through a systematic literature review. Finally, (Sharp et al. 2018) could be considered as a study close to this paper as the authors used a pre-defined methodology to select the articles to analyze. Nevertheless, they employed Natural Language Processing (NLP) to analyze around 4000 unique articles and provide insights about the scientific literature of ML applied in I4.0. The use of NLP can be useful to identify important trends, but it does not allow the authors to analyze the detail of the reviewed papers, where it is likely to find interesting research gaps and insights. On the other hand, a systematic review allows the authors to both follow a rigorous methodology and perform a detailed study of each chosen article.

Even though the PPC is closely related to the domain of supply chain, the latter is not included in the scope of this review as its vastness would increase the risk of straying from the focus on PPC. Therefore, to learn about recent trends on this topic, the authors invite readers to refer to Hosseini et al. (2019), who performed a comprehensive review of quantitative methods, technologies, definitions, and key drivers of supply chain resilience. In fact, supply chain resilience is a growing research area that examines the ability of a supply chain to respond to disruptive events (Hosseini et al. 2019). Applications of this topic have been done by Hosseini and Barker (2016), who applied Bayesian networks to perform supplier selection based on primary, green, and resilience criteria; and Hosseini and Ivanov (2019), who proposed a method using Bayesian networks to assess the resilience of suppliers in identifying critical links in a supply network.

The queries were performed between 10/10/2018 and 24/03/2019 in two scientific databases: ScienceDirect and SCOPUS. The following keywords conducted the research:

- ("Deep Learning" OR "Machine Learning") AND ("Production scheduling")
- ("Deep Learning" OR "Machine Learning") AND ("Production planning")
- ("Deep Learning" OR "Machine Learning") AND ("Production control")
- ("Deep Learning" OR "Machine Learning") AND ("Line balancing")

To consider the context of I4.0, only papers published since 2011 were considered because this year corresponds to the formal introduction of I4.0 at the Hannover Fair. Additionally, only communications labeled as "Research Articles" in ScienceDirect and "Conference paper" OR "Article" in SCOPUS were included to solely capture articles presenting application models. Subsequently, a review of titles and abstracts allowed for the exclusion of articles not related to ML-PPC. After the removal of duplicates, a full text analysis allowed a final selection that excluded papers that did not fit

with research questions. The sample size obtained encompasses 93 scientific papers. The article selection methodology with its Restrictions (R) is described in Fig. 1.

## A brief focus on the query keywords

The used string chains represent a core strategic choice for review. Therefore, this sub section aims to provide an analysis of the employed keywords.

Concerning the keywords used in the first parenthesis of the string chains, the use of "Deep Learning" and "Machine Learning" was done for two reasons: firstly, they are relatively new terms, which eases the identification of recent trends in the literature; and secondly, they are directly related to one of the two core subjects in this study, which is ML. Other terms such as "Data Mining" or "Statistical Learning" could have been sensible choices too, as they are often used interchangeably with "Machine Learning" and "Deep Learning". Nevertheless, using these two terms might have deviated this study from its core topic. In fact, a recent study suggests that the differences between ML and Data Mining are not consistently defined in the literature. Thus, Data Mining is mostly considered to be the process of generating useful knowledge from data (Schuh et al. 2019). To do so, it draws from other fields such as Artificial Intelligence, Statistics, ML, and Data Analytics. Therefore, Data Mining can be a vast topic, and does not exclusively concern ML, which could potentially affect the focus of this study. As there seems to be no clear boundary between these terms, a short bibliometric analysis was performed to assess the chosen keywords. The analysis was done using VOSviewer, software developed by the University of Leiden to draw insights from scientific literature. Furthermore, using keywords related to specific ML techniques such as "Random Forest" or "k-means" did not seem appropriate due to the risk of introducing a bias when answering the second research question. In fact, this could have artificially boosted the results of the queried techniques.

The bibliometric analysis followed a similar methodology to that used to choose the final article sample (cf. Fig. 1). The objective was to briefly assess the influence of different keywords on the queries' results. For the analysis, three different string chains were considered: "Deep Learning" OR "Machine Learning", "Data Mining", and "Statistical Learning". The queries were performed on 06/10/2019 and the detail of the search strategy can be found in "Appendix I". Finally, as the aim is to analyze the available literature when querying with a certain string chain, no Title and abstract review was done as this could introduce a bias into the results due to the authors' influence.

The bibliometric analysis focused on the keywords defined by the authors for all of the papers of each of the three samples. To represent the results, the network

**Fig. 1** Search strategy used to capture the scientific literature
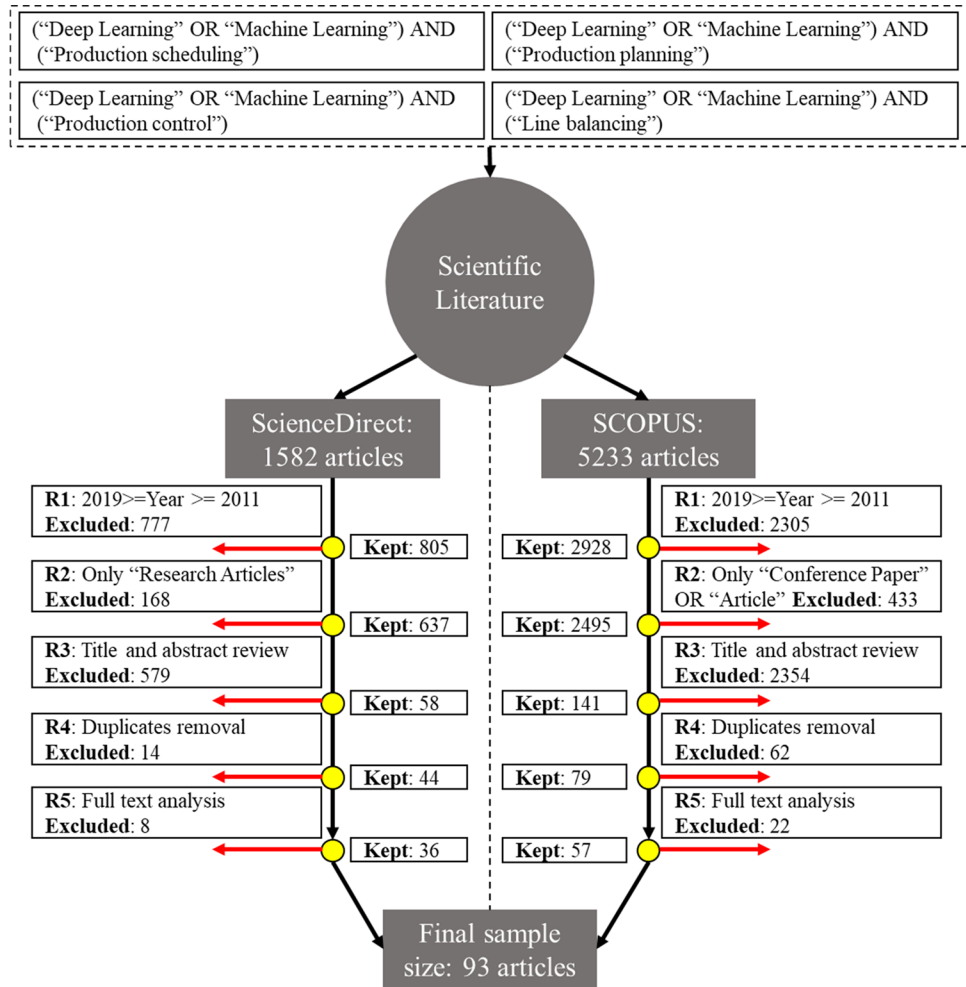
visualization from VOSviewer was employed. In such a network, the nodes represent the keywords or items, their sizes represent the keyword importance determined by the number of occurrences, and the links between the nodes represent their co-occurrence. Furthermore, the relatedness between two terms is represented through their spatial distance in the network: two keywords closely related will be spatially closer. For this review, the obtained networks were displayed under the "overlay visualization," which shows the average publication year for each of the keywords through a color scale. For clarity reasons, a filter was applied on the minimum number of occurrences to display; at most, 50 items per graph. Also, the queried keywords were highlighted with a red frame to assist in their identification. The networks are presented on Figs. 2, 3, 4.

Results from the bibliometric study suggest that "Statistical Learning" may not be a common keyword to find in ML-PPC research because the size of the obtained article sample (241 articles) is far below the results obtained with the other two queries. In fact, "Deep Learning" OR "Machine Learning" and "Data Mining" provided 2862 and 2166 articles,

respectively (cf. "Appendix I"). This is also stated in all of the networks, in which the item "Statistical Learning" does not appear, probably due to the filter excluding keywords with a low number of occurrences.

Analyzing the relatedness between "Data Mining" and "Machine Learning" by their spatial distance on the networks provides an idea of how these concepts are associated: they are spatially closer on the "Data Mining" Network (Fig. 3) than on the "Deep Learning" OR "Machine Learning" network (Fig. 2). This suggests that Data Mining tends to relate more often to ML, rather than ML to Data Mining. Such a relation may support what is said in (Schuh et al. 2019), in which Data Mining is considered a field drawing from ML, Artificial Intelligence, Statistics, etc. to produce useful insights.

Findings from the network visualizations show that the item "Machine Learning" is always associated with a more recent average publication year than "Data Mining". This supports the idea that "Machine Learning" is a relatively new term, which can lead to the identification of recent trends in literature. Furthermore, querying with "Deep
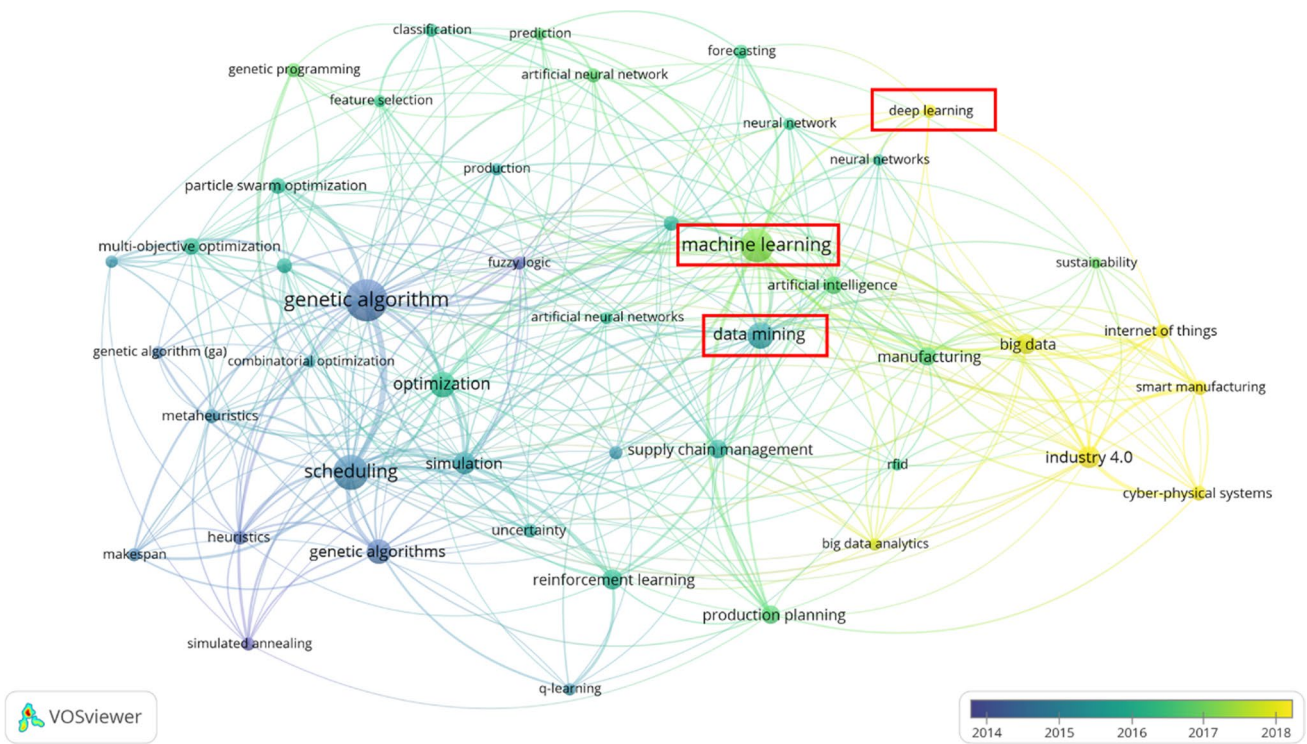
**Fig. 2** Network visualization with the average publication year for "Deep Learning" OR "Machine Learning"
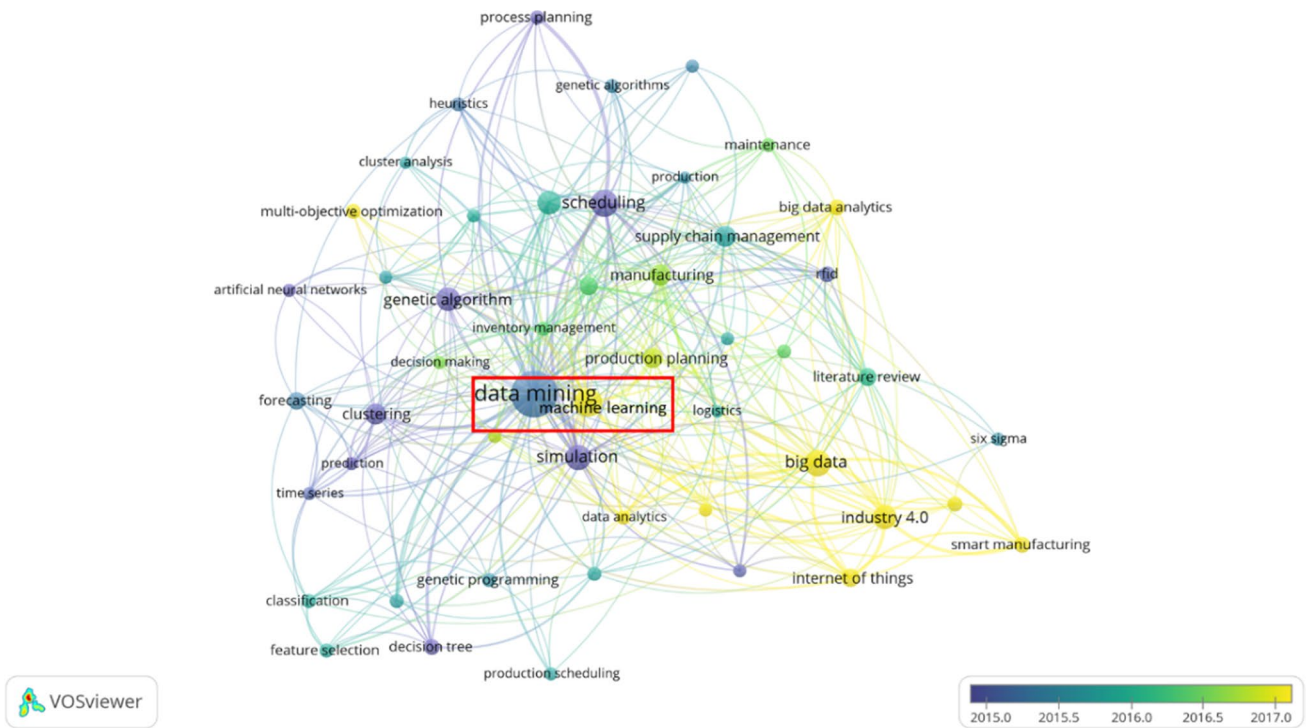


**Fig. 3** Network visualization with the average publication year for "Data Mining"
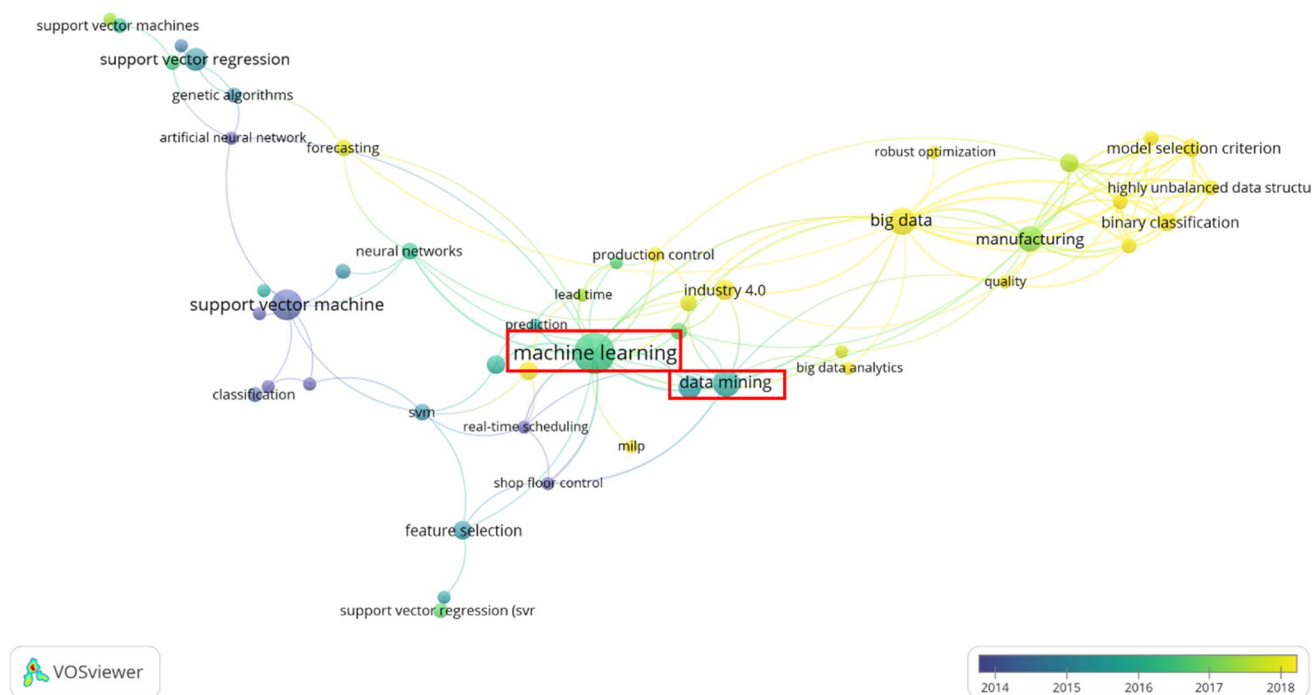
**Fig. 4** Network visualization with the average publication year for "Statistical Learning"

Learning" OR "Machine Learning" provides a more recent average publication year (2017.06) for the item "Machine Learning" than with the other two queries: 2016.95 when querying with "Data Mining" and 2016.41 when querying with "Statistical Learning". Finally, "Deep Learning" OR "Machine Learning" was the only query enabling the inclusion of the item "Deep Learning" with enough occurrences (25) to pass the filter, which is a recent research topic with an average publication year of 2018.28.

From the bibliometric analysis, it could be concluded that using "Deep Learning" OR "Machine Learning" as part of the query keywords is appropriate enough, as this allows for the identification of a big sample of recent papers, enabling the identification of new trends. It seems that "Statistical Learning" does not provide enough recent results to be considered. Finally, even if "Data Mining" is closely related to "Machine Learning," it covers a vast domain that can deviate from the focus of this review.

Regarding the keywords employed in the second parenthesis of the string chains, the objective was to represent the main functions of the PPC under the definition provided in the introduction section. Consequently, a determination of the global production quantities was represented by "Production Planning" and the aspect of the main objectives (i.e. profitability, productivity, and delivery time) was depicted by "Production Control". Finally, the real time synchronization of resources as well as product customizations were represented by both "Production Scheduling" and "Line

Balancing," given the fact that companies should be able to perform balanced scheduling even when facing customized client orders.

As the PPC is a transverse topic tangled with other functions such as maintenance, quality control, logistics, etc., the challenge was to decide whether or not these related subjects should be included as explicit keywords for the queries. The final choice was to not include them through keywords, as this would broaden the perimeter of the research too much, losing a focus on PPC. Nevertheless, it was decided to include, in the final article sample, the studies dealing with other functions only if they were related to the PPC.

## Analytical framework

This section presents the four axes that build the analytical framework that will be employed to harness knowledge and insight from the final sample of 93 scientific articles.

### First axis of the analytical framework: the elements of a method

This axis concerns the first and second research questions: the activities, techniques, and tools to implement a ML-PPC model. To link these three elements, the concept of "Mandatory Elements of a Method" (MEM) proposed by Zellner (2011) is used. In fact, this concept has been successfully

employed by other authors to propose methodologies in research domains such as product development (Lemieux et al. 2015) and lean in hospitals (Curatolo et al. 2014). Moreover, (Talhi et al. 2017) suggested its use to develop a methodology in the context of cloud manufacturing applied to product lifecycle management. Thus, the MEM suits the first objective of this study, which concerns the definition of a methodology to implement a ML-PPC. There are five elements in the MEM:

1. A procedure: order of activities to be followed when the method is employed.
2. Techniques: the means to generate the results. Activities from the procedure are supported by techniques, while the latter is supported by tools.
3. Results: they correspond to the output of an activity.
4. Role: the point of view adopted by the person who performs the activity and is responsible for it.
5. Information model: this refers to the relation between the first four mandatory elements.

In the scope of this study, the first two elements are the concern. Firstly, to evaluate the procedure, the activities used to perform a ML-PPC implementation will be recognized and their use will be measured. By activities, this research refers to tasks such as "model comparison and selection" or "data cleaning". Secondly, to address the techniques, ML models and tools will be identified, and their use will be measured. ML models point to elements such as Support Vector Machines or Neural Networks, while tools relate to programming languages or software used to implement these ML models.

To provide further insight concerning the ML techniques, the learning types will also be measured. This will be used to summarize the information regarding the techniques as well as to ease the identification of trends and research perspectives. Additionally, the learning types will serve as a bridge between the first and second objectives of this study, as they will be used in the mapping to classify scientific literature. Based on the work of Jordan and Mitchell (2015), three main learning types can be identified:

1. Supervised Learning (SL), which concerns ML techniques approximating a function $f(X) = Y$ by learning the relationship between the inputs $X$ and the outputs $Y$. For instance, learning the mapping between the Red Green, and Blue (RGB) codes (input $X$) in an image and the objects in it (output $Y$) to determine if a certain picture contains a misplaced product in a stock rack.
2. Unsupervised Learning (UL), which encompasses techniques allowing data exploration to find patterns and hidden structures in a given dataset $X$. For instance, finding categories in maintenance reports by using the

description of the problem and the duration of the maintenance intervention.
3. Reinforcement Learning (RL), which are techniques allowing the learning of actions to be performed by an agent interacting with a certain environment to maximize a reward. For example, teaching an Automated Guided Vehicle (AGV) in a warehouse how to avoid obstacles to maximize the number of delivered packages.

## Second axis of the analytical framework: employed data sources

This axis addresses the third research question: the harnessed data sources. Identifying which are the data sources used to perform a ML-PPC is capital. In fact, data could be considered as the raw material allowing ML models to develop autonomous computer knowledge gain (Sharp et al. 2018). Moreover, the quality of the final model will depend to a great extent on the quality and appropriateness of the used data. Therefore, the choice of the data source is an important decision when training a ML model. To address this axis of the analytical framework, the data source types proposed by Tao et al. (2018) will be used. They mention that there are five main data sources used in the data-driven smart manufacturing:

1. Management data (M): historical data coming from company's information systems such as the ERP, Manufacturing Execution System (MES), Customer Relationship Management system (CRM), etc. M data will concern production planning, maintenance, logistics, customer information, etc.
2. Equipment data (E): data coming from Internet of Things (IoT) technologies implemented in the factory. It refers to sensors installed in physical resources such as machines, places such as workstations or human resources such as workers. In the case of workers, data is collected passively, such as by RFID sensors installed on helmets.
3. User data (U): consumer information collected from e-commerce platforms, social media, etc. It also encompasses feedback given by workers or experts that will be used to train the ML-PPC model. User data coming from workers is collected actively, for example through interviews or questionnaires.
4. Product data (P): data originating from products or services either during the production process or from the final consumer.
5. Public data (Pb): data available in public databases from universities, governments or from other researchers.

The analysis of the 93 shortlisted articles suggested that some of them did not fit into the five data sources proposed by Tao et al. (2018): these communications used artificially generated data through computer simulations. Therefore, a sixth data source is proposed, which corresponds to the first contribution of this paper to the scientific literature:

6. Artificial data (A): data generated by computers (e.g. simulations) to assess ML-PPC implementations.

### Third axis of the analytical framework: the use cases of the ML-PPC in the I4.0

This axis concerns the fourth question: it aims to show which applications can be achieved when applying a ML-PPC. Moreover, identifying the use cases and quantifying their use frequency is important to detect trends as well as further research gaps. By use cases, this study refers to the different possible applications in a certain domain, such as maintenance, quality control, distribution, etc. In fact, as the PPC is entwined with several manufacturing subjects, is difficult to perform a complete review on PPC if these topics are ignored. For example, if there were a predictive maintenance study meant to enable a more robust production scheduling, such application would be directly related to the PPC through maintenance. To start this analysis, the use cases of I4.0 initially proposed by Tao et al. (2018) were considered. They identified six of them:

1. Smart Maintenance: harnessing data to perform preventive and predictive maintenance. For instance, monitoring machine components to estimate the best date to perform a maintenance intervention.
2. Quality Control: applying BDA to supervise the manufacturing process or products, seeking for possible quality problems and/or allowing the identification of root causes.
3. Process Control and Monitoring: constantly analyzing data coming from the shop floor to perform a smart adjustment of the functioning parameters of physical resources (machines, AGVs, etc.). The objective is to automatically control these physical resources and/or optimize their parameters with respect to the working conditions.
4. Inventory and Distribution Control: stock management, parts and tools tracking, and distribution control with the use real-time and/or historical data.
5. Smart Planning and Scheduling: considering production uncertainties to perform a production planning and scheduling closer to the current state of the production system. For instance, considering unexpected maintenance problems to reschedule a production order and minimize the delay.

6. Smart Design of Products and Processes: using BDA to support new products and processes development. For instance, using NLP to analyze the technical requirements of a new product and then to propose the potentially suitable manufacturing process.

The analysis of the 93 scientific articles suggests that these six use cases are not enough to fully characterize the recent publications. Additionally, papers not fitting in the initially proposed use cases shared the same application: time estimation (cycle time, operation time, etc.). Consequently, a seventh use case is proposed:

7. Time estimation: adaptation of different manufacturing related times to current working conditions. For instance, estimating the operation times to the actual work rate of each employee instead of using the data from the Method Time Measurement (MTM) approach.

### Fourth axis of the analytical framework: the characteristics of I4.0

The I4.0 aims to transform the collected data during the product's lifecycle into "intelligence" to enhance the manufacturing process (Tao et al. 2018). With this transformation, the objective is to reduce costs while improving the quality, productivity, sustainability of the production system (Wang et al. 2018a). However, what specific benefits could be expected when embracing the I4.0? To answer this question, the characteristics of I4.0 need to be identified. Tao et al. (2018) argue that I4.0 enables the following paradigms:

1. Customer-Centric Product Development: production systems in the I4.0 should be able to adjust their parameters by considering variables coming from customers such as their behavior, their needs, the way they use the products, *inter alia*. It is the case of manufacturing personalized products, designing processes from the customer requirements or proposing a target manufacturing cost for each consumer profile.
2. Self-Organization of Resources: I4.0 should endow production systems with the capacity of considering data coming from the manufacturing process to better engage the available resources. Additionally, this data should also be used to plan capital and operational expenditures. For example, updating the scheduling of machines the shop floor after new urgent order is released.
3. Self-Execution of Resources and Processes: in the I4.0, resources should become "smart" by providing them a real-time awareness and interaction capacity with the manufacturing environment (Huang et al. 2019). Therefore, the self-execution of resources concerns their faculty of making decisions depending on the received

information or measured data. It is the case of machines automatically adapting their functioning parameters to work optimally or trolleys automatically replenishing workstations when these reach a certain level of security stock.

4. Self-Regulation of the Production Process: unexpected events should be effectively handled in the I4.0. Thus, this characteristic concerns the capability to perform the required adjustments to respond to unpredicted problems. For example, relaunching the scheduling process for a certain production line when one of the machines experienced a breakdown.

5. Self-Learning of the Production Process: this characteristic follows a similar logic as the self-regulation of processes in terms of adjustability. However, it relates to the capacity of the production system to adapt to predicted events. It is the case of predictive maintenance, which uses BDA to estimate the remaining useful life of machine's components. Afterwards, the manufacturing system can adapt to the results of this prediction.

After concluding the analysis of the 93 articles, three characteristics seem to be overlooked: the environmental dimension, the knowledge generation, and the inclusion of the human being. To consider these dimensions that seem to not be explicitly raised in the work of Tao et al. (2018), three new characteristics are proposed:

6. Environment-Centric Processes: estimations suggest that the electronics and home appliances industry scrapped around 100 million goods in China in 2012 (Tian et al. 2013). As exemplified, the environmental impact of industry is far from being negligible, which is the reason why industrialized countries have started to tighten regulations and engage environmentally friendly practices in manufacturing (Tuncel et al. 2014). Research done in the context of I4.0 must not overlook this aspect. Therefore, this characteristic concerns the use of new technologies to create environment-centric processes. For example, optimizing the disassembly scheduling process to maximize the number of components that can be recycled.

7. Knowledge Discovery and Generation: most of the companies have been computerized for a long time, which has eased the collection of data. Despite the access to a plethora of information systems, generating knowledge from raw data still supposes a major industrial and academic challenge. Besides, the generation of knowledge is a mandatory step to improve the adoption of BDA by companies (Grabot 2018). In fact, knowledge could be considered as one of the most valuable assets in manufacturing (Harding et al. 2006), the reason why generating it represents an important gain behind the adoption of BDA. Therefore, as I4.0 is characterized by allowing

knowledge creation, research efforts must include it to generate value. One example of this is harnessing data from maintenance reports to provide the production of responsible real-time information about the root causes of machine breakdowns.

8. Smart Human Interaction: even with the advent of multiple I4.0 technologies, its adoption would be significantly hindered by not keeping humans in the loop or not considering their interaction with the proposed solutions. For instance, Thomas et al. (2018a) experienced the case of a company that was not willing to introduce an improved version of a quality control system because it somehow excluded the person from the process. Therefore, this characteristic concerns the consideration and/or inclusion of a human being when implementing new technologies. Examples of this would be a worker behavior recognition system based on computer vision or software interacting with operators through NLP.

Figure 5 summarizes this section. It also presents the relationship between the Research Questions (RQ), the analytical framework axes, the research objectives, and the expected outputs of this study.

# Results

## First research question: activities employed in ML-PPC

To identify the activities, the tasks used to implement a ML-PPC in each of the 93 communications were identified. Afterwards, these tasks were grouped into categories to ease the information analysis. These groups of activities were analyzed by two experts to keep the most meaningful ones. Results suggest eleven standard and recurrent activities:

1. Data Acquisition system design and integration (DA): design and implementation of IoT systems to collect data. This activity also encompasses the data storage and communication protocols.

2. Data Exploration (DE): use of data visualization techniques, inferential statistics, and others to derive initial insights and conclusions about the dataset.

3. Data Cleaning and formatting (DC): data preparation from the raw data to make it exploitable by the ML-PPC model. It concerns tasks such as outlier removal or missing values handling.

4. Feature Selection (FS): choice of the most suitable inputs to the ML-PPC model. It can be done through statistical techniques, e.g. stepwise regression or by means of expert insight.
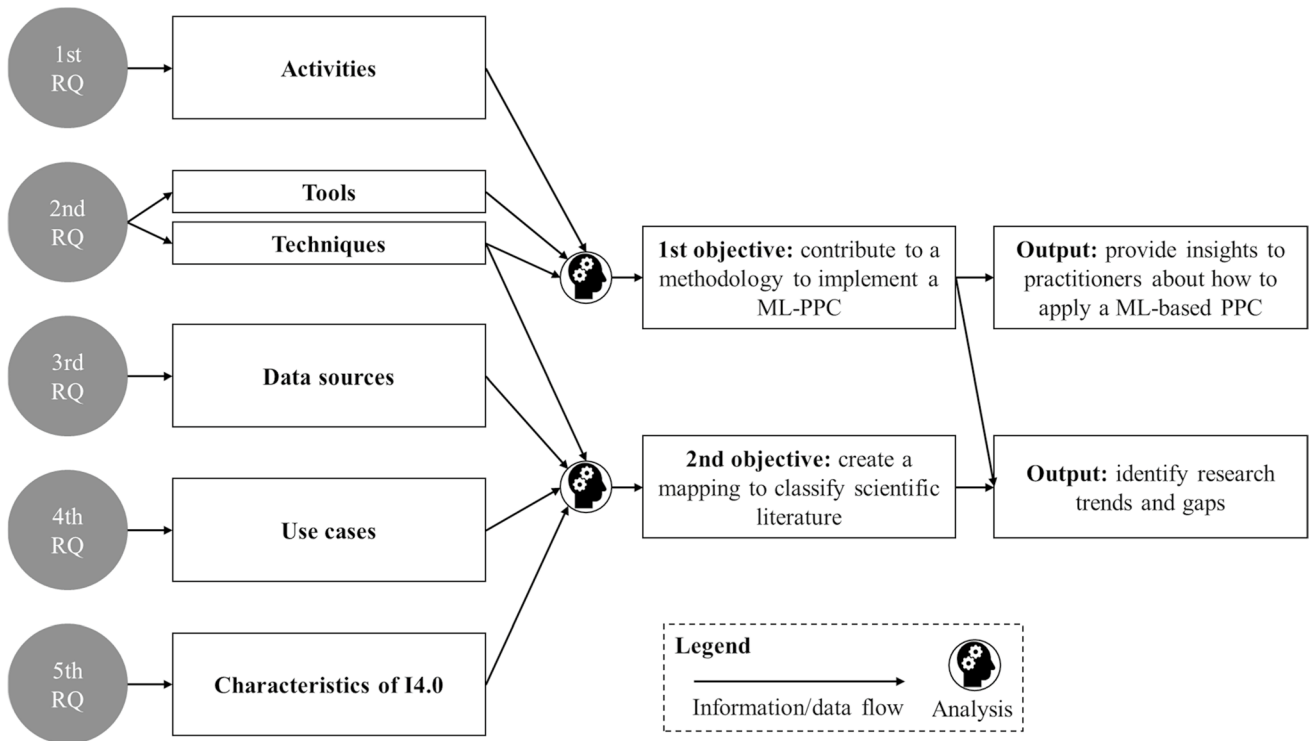
**Fig. 5** Relationship between the building blocks, research objectives and expected outputs of this study

5.  Feature Extraction (FE): use of variables from the initial dataset to calculate more meaningful features.
6.  Feature Transformation (FT): representation of the initial features into different spaces or scales using techniques such as normalization, standardization or kernel transformations.
7.  Hyperparameter Tuning and architecture design (HT): definition of the ML model architecture and adjustment of its hyperparameters to improve the performance. For instance, optimizing the learning rate and defining the activation function in a neural network.
8.  Model Training, validation, testing, and assessment (MT): using the data to perform the training, validation and testing process. It can be done through techniques such as k-fold cross-validation. It also encompasses the choice of the training/validation/testing set split and the model's performance assessment.
9.  Model Comparison and selection (MC): several ML techniques can be used to achieve a certain task. This activity concerns the comparison of multiple ML models to choose the one that better suits the needs.
10. Contextualized Analysis or application (CA): going further than just assessing the model's performance. It concerns the actual implementation of the ML-PPC model or the analysis of its results in the context of the problem that is addressed by the study.

11. Model Update (MU): data used to train ML models represents the context of the studied environment at a given moment. However, this context is dynamic, hence the ML-PPC model must be adapted. Therefore, this task concerns the model update through new data.

To address this research question, the percentage of papers using each activity was measured. These results are summarized in Fig. 6. Findings suggests that four groups of activities can be proposed following their usage:
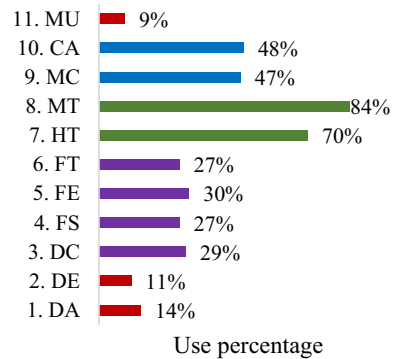


**Fig. 6** Use percentage by activity. CUAs in green, OUAs in blue, MUAs in purple, and SUAs in red

1. Commonly Used Activities (CUAs), applied in more than 60% of the analyzed papers (no. 7, 8).
2. Often Used Activities (OUAs), harnessed in 40 to 60% of the communications (no. 9, 10).
3. Medium Use Activities (MUAs), employed in 20 to 40% of the cases (no. 3–6).
4. Seldom Use Activities (SUAs), used in less than 20% of the reviewed articles (no. 1, 2, 11).

These groups show that a considerable amount of research papers only focus on the architecture design, training, and assessment of ML-PPC models (CUAs cluster), while not employing or documenting the use of other activities. Considering OUAs, it is surprising to find that only half of the communications used the CA, which corresponds to an actual implementation of the proposed model in the context of the study. This suggests that half of the studies go no further than just training and evaluating the performance of the model.

MUAs group encompasses data pre-processing tasks, which are capital to any ML implementation. Even if these activities are frequently employed in practice, their low usage is probably because researchers do not mention them, implying a lack of documentation. Moreover, as one of the characteristics of big data is the variety (in type, nature, format, etc.) (Zhou et al. 2017), it is crucial to employ data pre-processing activities to ensure the quality of the final models. Consequently, this lack of documentation can represent a pitfall to practitioners willing to apply ML-PPC based on research papers.

Finally, SUAs cluster highlights the most important research gaps in scientific literature. Three key findings can be inferred from activities in this group: firstly, the low usage of DA highlights the challenge of coupling IoT technologies with ML-PPC. This is a major pitfall to deploy ML-PPC in companies, as they normally need real-time data or statuses from their manufacturing systems. Secondly, the lack of DE utilization could mean that ML-PPC applications tend to jump directly to activities in the CUAs cluster while overlooking descriptive and basic inferential statistics techniques. This represents an obstacle to generating knowledge from data, as DE can draw conclusions easily interpretable by non-ML specialists. Finally, the rare use of MU implies that adapting the ML-PPC model to a dynamic manufacturing context is seldom addressed. This unpredictable change of the statistical properties and relationships between variables over time is known as *concept drift* (Hammami et al. 2017). Not addressing this issue can be harmful for the model reliability in the long term.

## Second research question: techniques and tools used in ML-PPC

Concerning the techniques, results present the number of times a given ML model is used. In the case of communications comparing several techniques, only the one chosen by the authors because of its better performance was considered. If this best-performing model employs several techniques, each of them is counted as used once.

There are numerous ML techniques in scientific literature. Therefore, to ease the analysis of results, a grouping of techniques in families is proposed is Table 1. These families were determined with the help of a ML expert. It is important to mention that the column "Concerned techniques" in Table 1 is not an exhaustive list, it is limited to techniques found in the systematic literature review.

Results are presented in Fig. 7. They suggest that NN, Q-Learning, and DT are the most used techniques in ML-PPC. The extensive use of NN is probably due to their ability to learn complex non-linear relationships between variables, often delivering good performance when compared to other techniques. Even if Q-Learning remains, by far, the most used RL technique, other RL models such as Sarsa or R-Learning are used, which points an interest in agent-based modeling in ML-PPC. Finally, the attention drawn by DT techniques is probably linked to their excellent trade-off between accuracy and interpretability, allowing knowledge generation.

The high use of Clustering techniques could be explained by the fact that data in manufacturing systems is normally unlabeled and can contain meaningful unknown patterns. Therefore, clustering can be employed to discover groups as well as hidden structures in datasets.

The usage evolution of the six most used technique families was also measured. Figures representing this can be found in "Appendix II". Due to an imbalance in the amount of articles over the different years, results are presented as relative frequencies. For example, if the NN achieved a usage of 27% in 2018, it means that 27% of all the techniques used in that year corresponded to such a model. Results suggest that there is a strong growth in the use of NN since 2015, this is possibly due to the growing computing power, recent findings in terms of architectures such as CNNs or LSTMs, and the development of specialized frameworks like PyTorch, TensorFlow, Keras, etc. which ease the task of implementing such models. Moreover, results show a growing interest on Ensemble learning techniques which evolved from not being used between 2011 and 2013 to accounting for 14% of applications in 2018. This can possibly explain the loss of interest on DT since 2017, as Random forests (a type of Ensemble learning) can achieve better performance by using committees of decision trees.

**Table 1** Technique families with their respective ML models

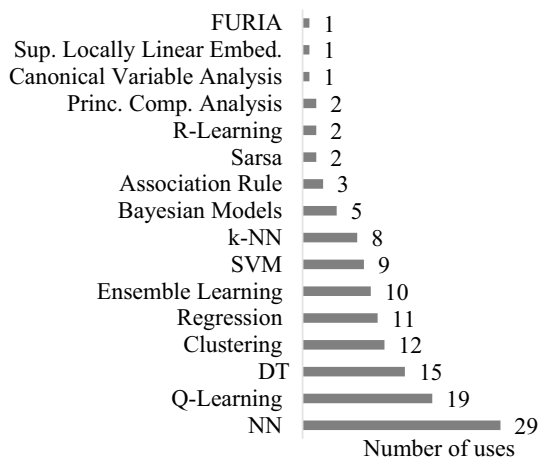| Family | Concerned techniques |
|---|---|
| Association rule | Association rule |
| Bayesian models | Bayesian networks, Naïve Bayes |
| Canonical variable analysis | Canonical variable analysis |
| Clustering | c-Means, density peak clustering, hierarchical clustering, k-means |
| Ensemble learning | Bagging, gradient boosting, machine learner fusion-regression, random forests, stacking |
| FURIA | FURIA |
| k-NN | K-nearest neighbors (k-NN), neighborhood component feature selection |
| Neural network (NN) | Artificial neural agent, autoencoders, convolutional neural network (CNN), deep belief networks, extreme learning machine, long short-term memory (LSTM), multi-layer perceptron, self-organizing maps, stacked denoising autoencoders |
| Principal component analysis (Princip. Comp. analysis) | Principal component analysis |
| Q-learning | Q-learning |
| Regression | Gaussian process regression, linear regression, logistic regression, polynomial regression, radial basis function approximation |
| R-learning | R-learning |
| Sarsa | Sarsa |
| Supervised locally linear embedding (Sup. Locally Linear Embed.) | Supervised locally linear embedding |
| Support vector machines | Support vector machines (SVM) |
| Decision trees (DT) | Decision trees |



**Fig. 7** Number of uses by technique family

As NN and Ensemble learning families seem to be recently attracting the research community, a detailed view of their encompassed techniques is presented in "Appendix III". Concerning NN, the most used technique is the Multi-layer perceptron, which is the classic architecture of a NN. However, more specialized architectures belonging to deep learning are starting to appear in PPC research. Such is the case of the CNNs, LSTMs, and Deep Belief Networks.

These techniques have presented good performance when dealing with specific problems, such as image recognition for CNNs, time series analysis for LSTMs or feature extraction for Deep Belief Networks. In the case of Ensemble learning, the most used technique is, by far, the Random forests. They seem to provide excellent results while enabling knowledge generation. In fact, they allow the most meaningful variables to be easily identified in the SL task, which is the reason why researchers tend to use them to both attain accuracy and model interpretability.

To measure the utilization of the learning types, each paper was analyzed, and the learning types used were identified and counted. As a given model can use several ML techniques, it can refer to several learning types at the same time. Hence, the different synergies between these were also considered. Results are presented in Fig. 8.

Findings show that the most used learning type is SL. This is probably because SL addresses two recurrent needs in applied research: classification and regression. In fact, SL can be used to learn the relationship between an input *X* and an output *Y* that can be either discrete in the case of classification or continuous for regression. Furthermore, it was found that RL techniques are extensively used, which confirms the interest behind agent-based models.

Concerning UL, it seems to be especially used with SL (SL-UL), which suggests a strong synergy between these
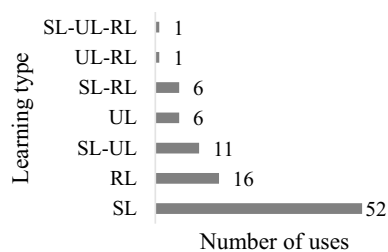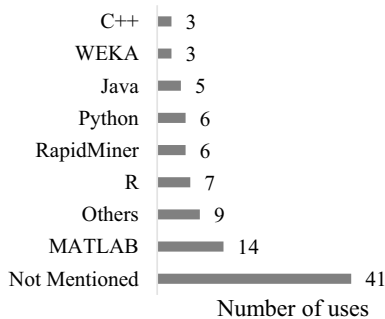
**Fig. 8** Number of uses by learning type



**Fig. 9** Number of uses by tool

two learning types. The reason behind this could be that UL techniques are normally used to perform data pre-processing, as with Principal Component Analysis, or discovery of hidden patterns in datasets, e.g. with Clustering. There are 6 papers using just UL, however, this learning type seems to unlock all of its potential when used in synergies, allowing for the design of more complex models.

Even if there are some SL-RL synergies, they are not very common. This is probably because SL is normally coupled with RL when there is a need of performing rapid estimations of functions to save computing time. However, it seems that most of the applications do not reach a scale that needs this kind of configuration. Finally, it was found that using UL-RL and SL-UL-RL is rare in the scientific literature. This does not mean that their synergy does not provide advantages, it is just that there may not be a current need for it. Also, it could be that coupling these learning types over-complexifies the model design, which prevents its use.

Concerning the tools, only programming languages or software used to implement the ML model were considered. Therefore, other tools such as discrete event simulation software are out of the scope of this research. Results are presented in Fig. 9.

For clarity sake, tools being used only once were grouped in the category denominated as "Others". These tools were: ACE Datamining System, C#, Clementine, GeNIe Modeler, Hugin 8.1, NetLogo, Neural-SIM, Visual C++, and Xelopes Library. Additionally, it is important to mention that most

of the researchers do not mention the tool the use to implement the models.

It could be said that MATLAB is, by far, the most used tool to perform ML-PPC in research. Besides its robust calculation capacity, the reason behind this could be that universities often invest in licenses for this software; therefore, they expect their researchers to use this tool. R is the second most used tool, which may be because it is a free software targeting statistical applications, including ML. Finally, the third most used tools are both RapidMiner and Python. The former eases the implementation of ML models thanks to its visual programming logic, while the latter is a multipurpose programming language recently characterized by its ML libraries and frameworks such as Scikit-learn, PyToch, Keras, etc.

### Third research question: used data sources to implement a ML-PPC

To answer this question, the data sources used by each of the analyzed papers were identified. These results are summarized in Table 2. The column "Identification" (ID) will assign a number to each communication. This will be used later to establish a mapping of the scientific literature.

Results show that "Artificial data" is the most used data source in recent scientific literature. This probably highlights the difficulty of accessing data coming from companies. Additionally, it is important to remember the extensive use of RL techniques. These models normally require constant access to data concerning the real-time status of the production system, which can be difficult to find in real factories. Therefore, researchers normally use Artificial or Public data to test their models. This issue could be addressed by creating digital twins, but this still represents a research challenge.

The extensive use of artificial data suggests that there are data availability issues. This poses two main challenges: firstly, dealing with highly unbalanced datasets when training, for instance, SL algorithms for classification, and secondly, accessing enough data to enable good generalization capacity, especially in deep learning models.

The first challenge is common when training ML models to identify disruptions. In fact, disruptive events in PPC such as machine breakdowns or quality problems tend to be scarce when compared to the total size of the dataset. Thus, ML techniques struggle to learn these events. To tackle this issue, some authors have proposed solutions such as data augmentation, a common practice in computer vision that consists of artificially creating new training examples by modifying existent observations (Perez and Wang 2017; Mikołajczyk and Grochowski 2018). Another approach is to use crafted algorithms adapted to class-imbalance. Bi and Zhang (2018) performed a comprehensive comparison of

**Table 2** Data sources used by each of the analyzed scientific articles

| ID | References | M | E | U | P | Pb | A |
|----|-----------|---|---|---|---|----|----|
| 1 | Aissani et al. (2012) | | | | | | X |
| 2 | Altaf et al. (2018) | | | | X | | |
| 3 | Bergmann et al. (2016) | | | | | | X |
| 4 | Cai et al. (2016) | | X | X | | | |
| 5 | Cao et al. (2019) | | X | | | | |
| 6 | Carvajal Soto et al. (2019) | | | | | X | |
| 7 | Chen et al. (2015) | | | | | | X |
| 8 | Diaz-Rozo et al. (2017) | | X | | | | |
| 9 | Ding and Jiang (2018) | X | X | | X | | |
| 10 | Dinis et al. (2019) | X | | X | | | |
| 11 | Dolgui et al. (2018) | X | | | | | |
| 12 | Doltsinis et al. (2014) | | X | | | | |
| 13 | Fotuhi et al. (2013) | | | | | X | X |
| 14 | Gao et al. (2014) | | | | | | X |
| 15 | Gyulai et al. (2014) | X | | | | | X |
| 16 | Gyulai et al. (2015) | X | X | | | | X |
| 17 | Gyulai et al. (2018a) | X | | | | | X |
| 18 | Gyulai et al. (2018b) | X | | | X | | |
| 19 | Habib Zahmani and Atmani (2018) | | | | | | X |
| 20 | Hammami et al. (2016) | | | | | | X |
| 21 | Heger et al. (2016) | | | | | | X |
| 22 | Huang et al. (2019) | | X | | X | | X |
| 23 | Ji and Wang (2017) | | | | | | X |
| 24 | Jiang et al. (2016) | X | | | | | X |
| 25 | Jurkovic et al. (2018) | | X | | | | |
| 26 | Kartal et al. (2016) | X | | | | | |
| 27 | Khader and Yoon (2018) | | X | | | | |
| 28 | Kho et al. (2018) | | X | | | | |
| 29 | Kim and Lim (2018) | | | | | | X |
| 30 | Kim and Nembhard (2013) | | | | | X | |
| 31 | Kosmopoulos et al. (2012) | | X | | | | |
| 32 | Kretschmer et al. (2017) | X | | | X | | |
| 33 | Kruger et al. (2011) | | X | | | | |
| 34 | Kumar et al. (2018) | | | | | X | |
| 35 | Lai and Liu (2012) | X | | | | | |
| 36 | Lai et al. (2018) | X | | | | | |
| 37 | Leng et al. (2018) | X | | | X | | |
| 38 | Li et al. (2012a) | | | | X | | |
| 39 | Li et al. (2012b) | | | | | | X |
| 40 | Li et al. (2013) | | | | | | X |
| 41 | Li et al. (2018) | | X | | | | |
| 42 | Liao (2018) | | | | | | X |
| 43 | Lieber et al. (2013) | | X | | X | | |
| 44 | Lingitz et al. (2018) | X | | | | | |
| 45 | Lubosch et al. (2018) | | | | | X | X |
| 46 | Lv et al. (2018b) | X | | | | | |
| 47 | Lv et al. (2018a) | X | | | | X | X |
| 48 | Ma et al. (2017) | X | | | | | X |
| 49 | Maghrebi et al. (2016) | X | | | | | |
| 50 | Manns et al. (2015) | X | | | X | | |
| 51 | Manupati et al. (2013) | | | | | | X |

**Table 2** (continued)

| ID | References | M | E | U | P | Pb | A |
|---|---|---|---|---|---|---|---|
| 52 | Mori and Mahalec (2015) | X | | | | | |
| 53 | Ou et al. (2018) | X | | | | | X |
| 54 | Ou et al. (2019) | | | | | | X |
| 55 | Palombarini and Martínez (2012) | | | | | X | |
| 56 | Priore et al. (2018) | | | | | X | X |
| 57 | Qu et al. (2015) | | | | | | X |
| 58 | Qu et al. (2016a) | | | | | | X |
| 59 | Qu et al. (2016b) | | | | | | X |
| 60 | Reboiro-Jato et al. (2011) | X | | | | | |
| 61 | Reuter et al. (2016) | X | | | | | |
| 62 | Rostami et al. (2018) | | X | | | | |
| 63 | Sahebjamnia et al. (2016) | X | | | | | X |
| 64 | Schuh et al. (2017a) | X | | | | | |
| 65 | Shahzad and Mebarki (2012) | | | | | | X |
| 66 | Shiue et al. (2012) | | | | | X | |
| 67 | Shiue et al. (2018) | | | | | X | X |
| 68 | Solti et al. (2018) | | | | X | | X |
| 69 | Stein et al. (2018) | | | | X | | |
| 70 | Stricker et al. (2018) | | | | | | X |
| 71 | Thomas et al. (2018a) | X | | | | | |
| 72 | Thomas et al. (2018b) | | | | | | X |
| 73 | Tian et al. (2013) | | | | | X | |
| 74 | Tong et al. (2016) | X | | | | | |
| 75 | Tuncel et al. (2014) | | | | | X | |
| 76 | Wang and Jiang (2018) | | X | | | | |
| 77 | Wang and Jiang (2019) | | X | X | | | |
| 78 | Wang and Yan (2016) | | | | | | X |
| 79 | Wang et al. (2015) | | | | | | X |
| 80 | Wang et al. (2017) | X | | | | | |
| 81 | Wang et al. (2018b) | X | | | | X | |
| 82 | Wang et al. (2018c) | | | | | X | |
| 83 | Waschneck et al. (2018) | | | | | | X |
| 84 | Wauters et al. (2012) | X | | | | | |
| 85 | Wu et al. (2015) | X | | | | | X |
| 86 | Xanthopoulos et al. (2017) | | | | | | X |
| 87 | Yang et al. (2016) | | X | | | | |
| 88 | Yeh et al. (2011) | X | | | | | |
| 89 | Yuan et al. (2014) | | | | | | X |
| 90 | Zhang et al. (2011) | X | | | | | X |
| 91 | Zhang et al. (2012) | | | | | | X |
| 92 | Zhong et al. (2014) | | X | | | | |
| 93 | Zhou et al. (2018) | | X | | | | X |
| | Totals | 33 | 20 | 2 | 12 | 14 | 43 |

state-of-the-art ML techniques adapted to this issue. The second challenge normally concerns the training of deep learning models as they need voluminous data to learn meaningful representations. This issue is normally tackled by transfer learning, which is the use of models already trained on a source task to perform another related task (Wang et al. 2018a), for instance, using a CNN trained to recognize pedestrians in the street to recognize operators on the shop floor. A comprehensive survey of transfer learning can be found in Pan and Yang (2010).

Management is the second most used data source. Hence, there seems to be a strong interest in valuing enterprise data

stored in information systems by making it available for researchers and practitioners. Furthermore, the use of Equipment and Product data suggests that recent applications are starting to employ data coming from IoT technologies installed in machines or semi-finished products. However, there are still tremendous research gaps when harnessing user data to implement ML-PPC models. Two studies used this data source, but only under the form of expert feedback to train the ML model. No study included consumer feedback from e-commerce platforms or social media to influence the PPC.

## Fourth research question: addressed use cases by recent scientific literature

To answer this question, each analyzed article was allocated to one of the seven proposed types of use cases. This allows to measure their importance in the scientific literature (Fig. 10).

Results point out that Smart Planning and Scheduling is the most addressed use case in recent scientific literature, with nearly half of the communications discussing it. This result may come from two main reasons: firstly, the string chains used in the methodology are closely related to this use case; secondly, it normally uses structured data relatively easy to get from information systems, which eases the task of implementing a data-driven approach. The strong use of Time Estimation in ML-PPC (14% of the papers) suggests that classical time measurement methods are not compliant with the growing complexity of the manufacturing systems, which may represent a pitfall to perform a reliable planning. Therefore, ML models considering more diverse variables

as inputs are being adopted. Moreover, some researchers have addressed the coupling of Smart Maintenance, Process Control and Monitoring, and Quality Control with the PPC. However, there is still effort to be made, as the share of these use cases was no higher than 10%.

Finally, two use cases are targeted as critical: The Inventory and Distribution Control (6%) and the Smart Design of Products and Processes (4%). These findings suggest two things: first, a lack of integration of the logistic functions into the ML-PPC, and secondly, a difficulty for harnessing insights from data to serve product and process design. This difficulty is probably because data employed in design is highly unstructured (text data, image data, etc.) and greatly depends on people's experience.

## Fifth research question: the characteristics of I4.0

To quantify their usage, the addressed characteristics in each of the 93 analyzed papers were identified and counted. Results are summarized in Fig. 11. In this figure, the sum of all the totals is higher than 93 as one ML-PPC model can satisfy several characteristics.

Findings show that the Self-Organization of Resources is, by far, the most addressed characteristic (56 uses) in ML-PPC applications. This result was expected, as this characteristic can be achieved through production planning and scheduling, two functions directly related to the PPC and found to be extensively employed in the use cases. Therefore, it can be concluded that the ML-PPC based models effectively enable this characteristic.

The Self-Regulation of the Production Process (33 papers), the Self-Learning of the Production Process (26

**Fig. 10** Share of the analyzed sample by proposed use case

Share of papers by use case

Legend:
- Time Estimation
- Smart Design of Products and Processes
- Smart Planning and scheduling
- Inventory and Distribution Control
- Process Control and Monitoring
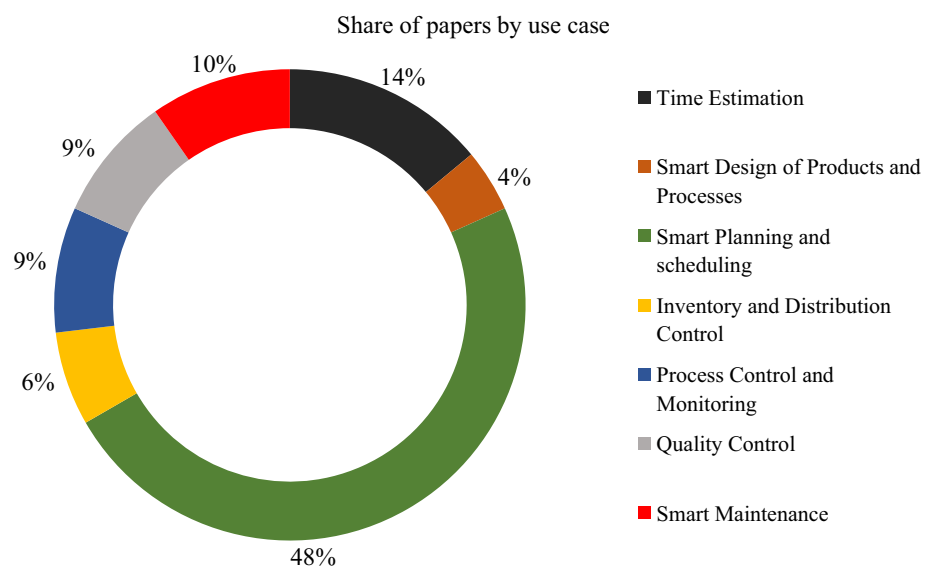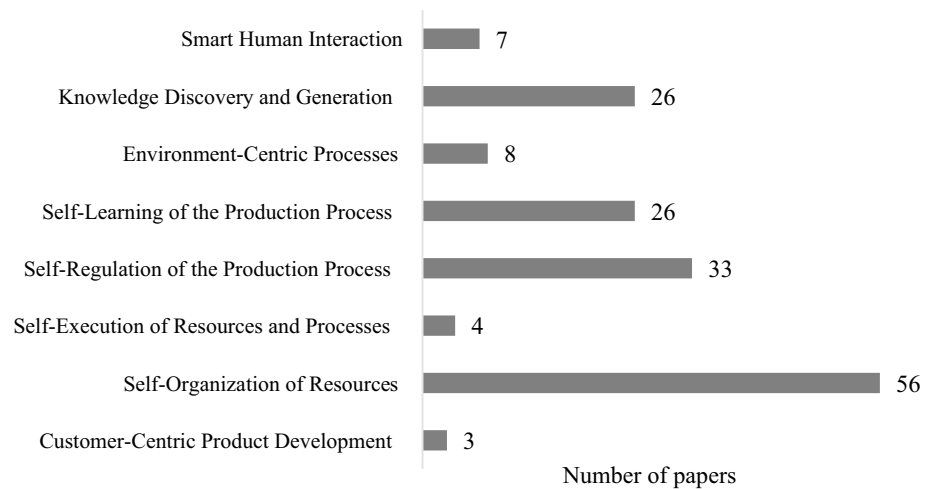- Quality Control
- Smart Maintenance

**Fig. 11** Number of papers by
I4.0 characteristic



papers), as well as the Knowledge Discovery and Genera-
tion (26 papers) appear to be moderately boarded. This leads
to two main conclusions: first, ML-PPC models effectively
endow manufacturing systems with the capacity of adapting
to unexpected events and predicting production problems.
This is suitable to handle the stochastic nature of production
environments. Secondly, ML is suitable to generate knowl-
edge from PPC data, which is crucial in I4.0, where data is
abundant, and it can provide useful guidelines to improve
the company's know-how.

Four characteristics were rarely satisfied: The Customer-
Centric Product Development (3 papers), the Self-Execution
of Resources and Processes (4 papers), the Smart Human
Interaction (7 papers), and the Environment-Centric Pro-
cesses (8 papers), which points to strong research perspec-
tives of ML-PPC applications enabling these features. Con-
cerning the Customer-Centric Product Development, it was
rare to find papers including customer-related variables into
their PPC. This can be due to the difficulty to access data
from customers or end users. For instance, as observed in
the data sources section, user data was seldom employed.

The low number of papers dealing with Self-Execution
of Resources and Processes suggests that it is unusual to
couple the PPC with autonomous physical resources. This
can be due to the complexity of such systems as they require
important capital investments as well as multi-disciplinary
knowledge in production systems, mechatronics, and control
theory.

It was very surprising to find that the Smart Human
Interaction (7 papers) and the Environment-Centric Pro-
cesses (8 papers) are rarely addressed. Indeed, manufactur-
ing systems can be human based in several steps such as
during the execution in the shop floor or during the tactical
planning definition. Not considering the interaction of the

proposed ML-PPC models with humans can be harmful for
the deployment of the proposed system, as it may worsen
the working conditions. Therefore, thinking about this
human-ML interaction is the cornerstone for a successful
adoption. Concerning the Environment-Centric Processes,
scarce applications tried to minimize the environmental
impact of production processes through ML-PPC. In a
world where natural resources are becoming rare, this is
a non-negligible aspect that must be considered, not only
because of the tightening of environmental laws by gov-
ernments but also because of the ethical responsibility of
companies.

## Cross-axes analysis: mapping the scientific literature through use cases, I4.0 characteristics, and learning types

To address the second objective of this study, a mapping
of the scientific literature in ML-PPC is proposed. This is
achieved through a cross-analysis employing the use cases,
characteristics of I4.0, and learning types. Results are repre-
sented via a cross-matrix having the use cases in the vertical
axis and the characteristics of I4.0 in the horizontal axis.
This matrix also allows the maturity of a given use case to
be assessed. For instance, a mature use case in the scientific
literature will tend to satisfy more I4.0 characteristics. From
this point of view, the crossing between a characteristic of
I4.0 and a use case will be referred as a *domain*.

The ID numbers defined on Table 2 are employed to place
the analyzed articles in the matrix. Additionally, the learn-
ing types employed by each communication are represented
using a color code. Figure 12 provides a summarized view of
this matrix, allowing for a high-level analysis that will help
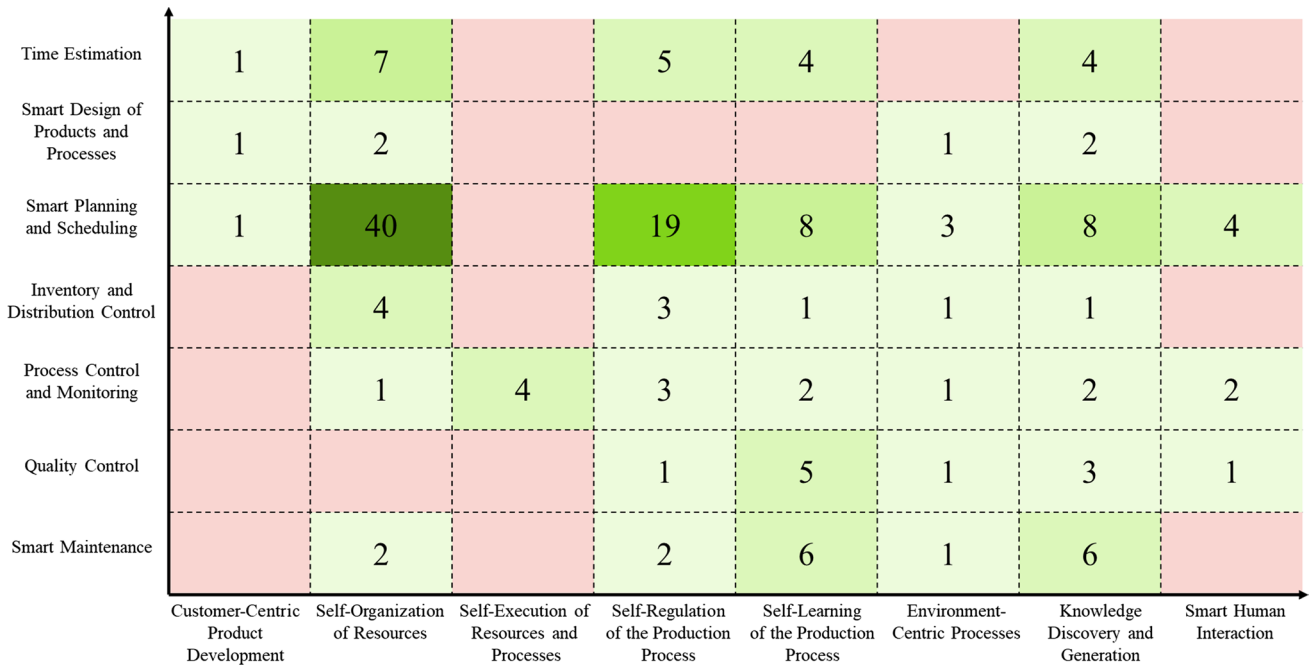to identify research gaps and trends in ML-PPC. Figure 13 is

**Fig. 12** Summarized view of the cross-matrix: number of papers by domain
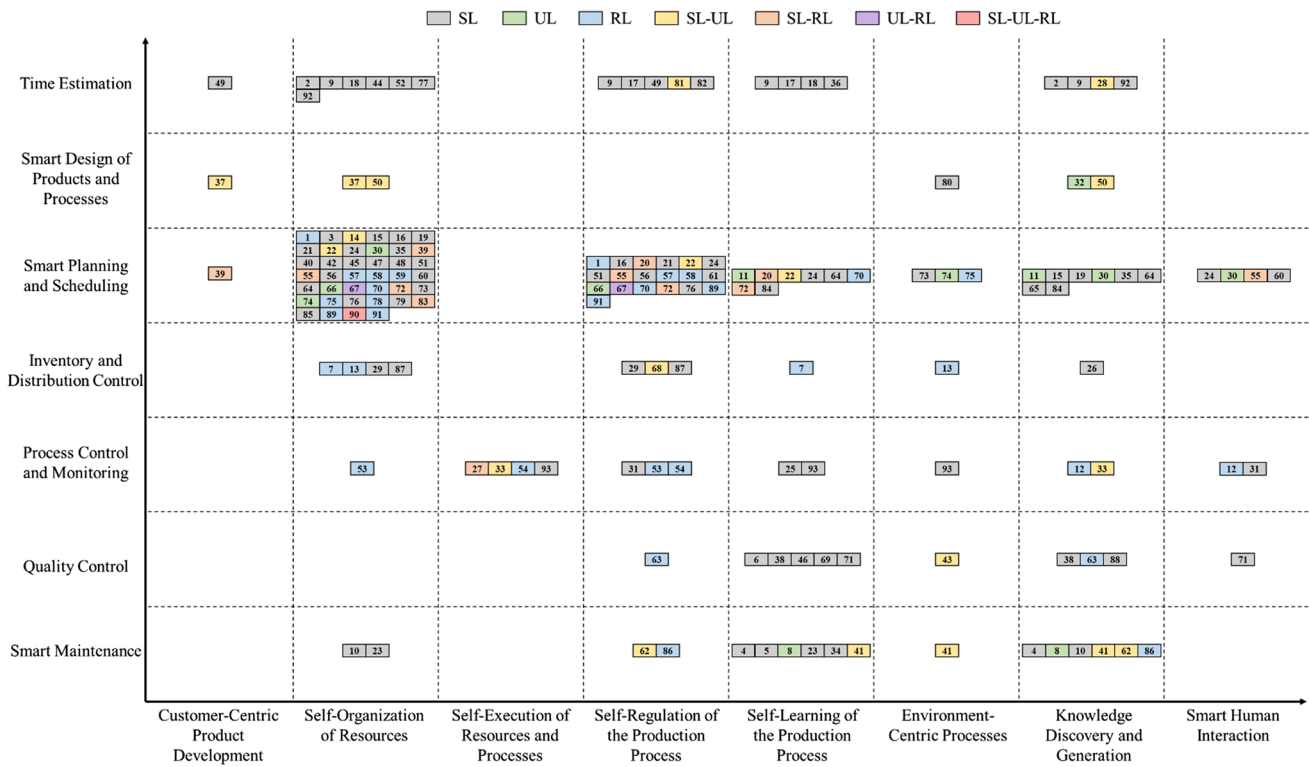


**Fig. 13** Detailed view of the cross-matrix for use cases, characteristics of I4.0, and learning types

a detailed view of the matrix indicating the scientific articles with their respective learning types found in each domain.

Figure 12 shows that among the 56 possible domains, 18 (32%) were not addressed at all. Furthermore, 24 (43%) domains lie in the range of 1 to 3 papers. This means that nearly half of the domains are in an exploration phase. These two remarks lead to conclude that ML-PPC in the I4.0 is still an active research topic with strong perspectives.

From Fig. 13, it can be said that there is a strong trend of using multiple synergies between learning types across all of the different use cases. However, there are no applications of RL in Time Estimation and in Smart Design of Products and Processes. The reason for this may be that these use cases have strong strategic impacts. Therefore, current ML implementations in such applications aim to support decisions rather than automating them such as with agent-based systems driven by RL.

There are two use cases achieving a high maturity: Smart Planning and Scheduling and Process Control and Monitoring. They both cover all but one of the characteristics of I4.0. In the case of Smart Planning and Scheduling, it fails to address the Self-Execution of Resources and Processes, which suggests that there are research perspectives in coupling the production planning and scheduling with autonomous physical resources. For the Process Control and Monitoring, there is a lack of applications satisfying the Customer-Centric Product Development, which would be an automatic optimization of physical resources from the analysis of customer-related variables.

Knowledge Discovery and Generation is the only characteristic addressed by all the use cases, which denotes an intense interest in knowledge creation from data. Furthermore, there is a strong presence of SL, UL, and SL-UL in this characteristic. This implies an important affinity between these learning types and the generation of useful information from raw data. Following a similar trend, there seems to be a generalized interest in Environment-Centric Processes, a characteristic that is addressed by almost all of the use cases. However, its low number of papers implies that there are strong research avenues to be explored.

Communications addressing the Self-Execution of Resources and Processes focused exclusively on Process Control and Monitoring applications. This shows that the dynamic optimization of working parameters of the machines allows data-driven intelligent resources to be created. However, this characteristic has further potential to be explored in PPC research with other use cases, such as in Inventory and Distribution Control with autonomous AGVs to serve logistic needs or in quality, by automating processes.

## Conclusion and further research perspectives

This state-of-the-art analysis studied 93 research articles chosen through the logic of a systematic literature review. These papers were analyzed by means of an analytical framework composed of four axes. First, the elements of a method were reviewed, which enabled an analysis of activities, techniques, and tools to perform a ML-PPC. Secondly, the data sources employed to implement a ML-PPC model were recognized and assessed. Thirdly, an analysis of the use cases enabled the recognition of the applications of data-driven models in the 4.0. Fourthly, the characteristics of I4.0 were identified and assessed through their usage. Additionally, a mapping of the scientific literature was proposed by means of the use cases, characteristics of I4.0 and ML learning types.

Results concerning the activities allowed the recognition of eleven recurrent tasks that are employed to create a ML-PPC model. They were grouped in four clusters following their use percentage: CUAs (Commonly Used Activities), OUAs (Often Used Activities), MUAs (Medium Use Activities), and SUAs (Seldom Use Activities). From these clusters, it can be concluded that activities belonging to the CUAs and OUAs clusters are well documented in the scientific literature. MUAs activities mainly contain data pre-processing tasks, which are necessary but not commonly documented by researchers. Finally, the SUAs cluster suggests that there are three activities rarely addressed in literature: the design and implementation of data acquisition methods from the manufacturing system, the exploration of data to get insights, and the constant adaptation of the proposed ML-PPC model to the environment dynamics.

An extensive review of the techniques identified the most used families in scientific literature. These were found to be the NN, Q-Learning, DT, Clustering, Regression, and Ensemble learning. From these results, a temporal evolution analysis of the top 6 most used families was performed. Findings suggested a growing interest in NN and Ensemble learning, which motivated a focused study on the detailed techniques encompassed by these families. Concerning the NN, the Multi-layer perceptron was the most used technique. Nevertheless, more specialized deep learning techniques such as CNNs, LSTMs, and Deep Belief Networks are

starting to be employed. With respect to Ensemble learning, the most used technique was Random forests.

The ML learning types were also reviewed. Findings showed that scientific literature mainly focused on the individual use of SL and RL. However, synergies between learning types are also employed. For instance, the most used synergy was SL-UL, which allows to explore and pre-process the data through UL to improve the SL training. The UL-RL and SL-UL-RL synergies had only one use each, which could be considered as a research gap, advising improvements in its integration. In fact, each learning type has its advantages and limitations. Hence, it is important to explore more synergy possibilities, as they may help overcome individual limits.

Other than increasing data availability, one option to encourage the utilization of UL-RL and SL-UL-RL is to boost the development of specialized libraries to build complex models coupling several learning types. Examples of this are deep learning frameworks such as TensorFlow, Keras, PyTorch, etc. which have eased the implementation of deep learning applications. This has allowed researchers to spend more time on the addressed problem than on the coding stage.

Results concerning the tools showed that MATLAB, R, Python, and RapidMiner are the most used tools in developing ML-PPC models in research. However, most authors did not mention the tool used, which is a limit of this study. Furthermore, it is important to mention that these results come from a sample of scientific articles, meaning that results are mainly valid in an academic context. If there are practitioners willing to implement ML-PPC models in companies, other aspects need to be analyzed such as the cost of the software, its scalability, skill availability in the labor market, compatibility with existing information systems, etc.

The current horizon of data sources used is dominated by Artificial and Management data. The former points to a difficulty in collecting all of the data required to implement ML-PPC models, while the latter suggest that companies are interested in valuing their data stored in information systems. Data coming from IoT sources such as Equipment and Product data was moderately used, nevertheless showing an interest in these technologies to collect data. Finally, ML-PPC models failed to integrate User data, probably because it is complex to collect and it engages an important responsibility concerning data privacy.

The most addressed use cases were Smart Planning and Scheduling and Time Estimation, probably because they are directly concerned by the PPC, which may lead to its high utilization. The fact that there are research articles in all of the use cases suggests that the PPC is a transversal function that benefits from several applications. Therefore, when designing a ML-PPC system for a company, the impact on all of the use cases must be assessed. Finally, it was found that Inventory and Distribution Control, as well as Smart Design of Products and Processes, are seldom addressed. This suggests that there is still a lot of progress to be made when coupling the PPC to logistics as well as product and process design through ML.

Concerning the characteristics of I4.0, results suggest that scientific literature in ML-PPC is extremely focused on satisfying the Self-Organization of Resources, which was expected, as one of the main goals of the PPC is resource management to satisfy the commercial plan. At a second level, the Self-Regulation of the Production Process, the Self-Learning of the Production Process, and the Knowledge Discovery and Generation seem to be more frequently addressed. However, Fig. 13 showed that they are mainly employed for Smart Planning and Scheduling, implying a lack of research in the other applications. Finally, there are three characteristics that are partially overlooked by researchers: Environment-Centric Processes, Smart Human Interaction, and Customer-Centric Product Development. The first two are essential characteristics of building more responsible production systems as they aim to include human beings and reduce the environmental impact of manufacturing processes. The latter relates to the alignment of the PPC to the customer's needs. Hence, it appears that recent ML-PPC research ignores the influence of the customer in the manufacturing process.

As illustrated in the proposed cross-matrix, 75% of the possible research domains are barely addressed or were not explored at all. This means that the ML-PPC is still a key topic for the enablement of I4.0, which presents strong research avenues. The main future research perspectives could be summarized in the following three key items:

1. Reinforce the role of IoT in ML-PPC: this would allow an improvement to the data acquisition system's design and would provide a means to perform a model update to tackle the concept drift issue. To do so, the ML mindset and workflow should be shifted from a linear to a circular process, considering the need to constantly retrain through new data. This way of thinking would enable the identification, from an early development stage, of the retraining policy and the necessary variables that could be measured again at a sensitive cost. By defining these two aspects, the data acquisition system design will be less complex to conceive, as the needs will be clearer.

This would avoid investment in sensors and resources and architecture that would not be exploited. Concerning the retraining policy, a review in the context of PPC reporting common practices, advantages and pitfalls seems to be missing in the scientific literature.

2. Improve the integration between the PPC, logistics, and design: it was stated that the PPC benefits from different use cases. However, recent literature seems to overlook logistics as well as product and process design applications coupled with the PPC. To tackle this challenge, it is necessary enable data availability, continuity and sharing over the design, logistics, and production departments. This could be achieved through interoperability as well as communication of intra-organizational systems such as the PLM, ERP, and MES. Even if projects that are meant to couple such systems are costly, they are necessary to ensure data availability and quality. One way to achieve this is the use of data lakes, which have been recognized as suitable to handle big data repositories of a structured and unstructured nature (Llave 2018; Lo Giudice et al. 2019). For instance, Llave (2018) concluded, through expert interviews, that one of the key purposes of data lakes is to serve as experimentation platforms for data scientists.

3. Set human interaction and environmental aspect as priorities to ensure the development of ethical manufacturing in I4.0: exploring the interaction of humans with the proposed ML-PPC models is paramount to building inclusive technologies at the service of society. To achieve this, the short- and long-term impact of ML-PPC systems on employees' working conditions must be assessed. If the system degrades them, it must be redesigned. Concerning the second aspect, seeking a reduction in the environmental impact of manufacturing through ML could provide important developments. This can be addressed from a purely PPC approach by optimizing, for instance, the scheduling of disassembly processes or by improving the prediction of production times to avoid energy waste. Other approaches could be the optimization of the supply chain. Even though the supply chain was not covered in this review, it is an appropriate domain for researchers to implement ML applications. For instance, by considering environmental criteria when choosing suppliers, as in (Hosseini and Barker 2016).

Some of the research gaps indicated in this review could motivate future work. Future work will be focused on the following aspects:

1. The proposed activities will be reviewed to determine an order between them, creating a procedure: this would help shift from a linear to a circular workflow when implementing ML-PPC models.

2. The most suitable techniques and tools will be linked to each of the activities with sectorial information: linking techniques, tools, and activities is the key to creating good practices that could be helpful to new practitioners, both in research and industry. Furthermore, according to Kusiak (2017, 2019), there are profound differences in the volume of data generation and usage across different industries. Therefore, future work will aim to identify trends categorized by sectorial information.

3. The current state of data availability solutions and workarounds will be explored: as data availability was found to be a main issue, a review of techniques to tackle the class-imbalance problem and the use of transfer learning in the context of PPC will be performed. Additionally, the utilization of data lakes for ML-PPC will also be explored.

4. Future research avenues will be proposed through an NLP analysis: NLP may enable the discovery of nontrivial trends present in the corpus of the 93 sampled articles. This will complement the results of the systematic literature review.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

# Appendix

## Appendix I: Detail on the search strategy for the bibliometric analysis

See Figs. 14, 15 and 16.

**Fig. 14** Search strategy detail for "Deep Learning" OR "Machine Learning"
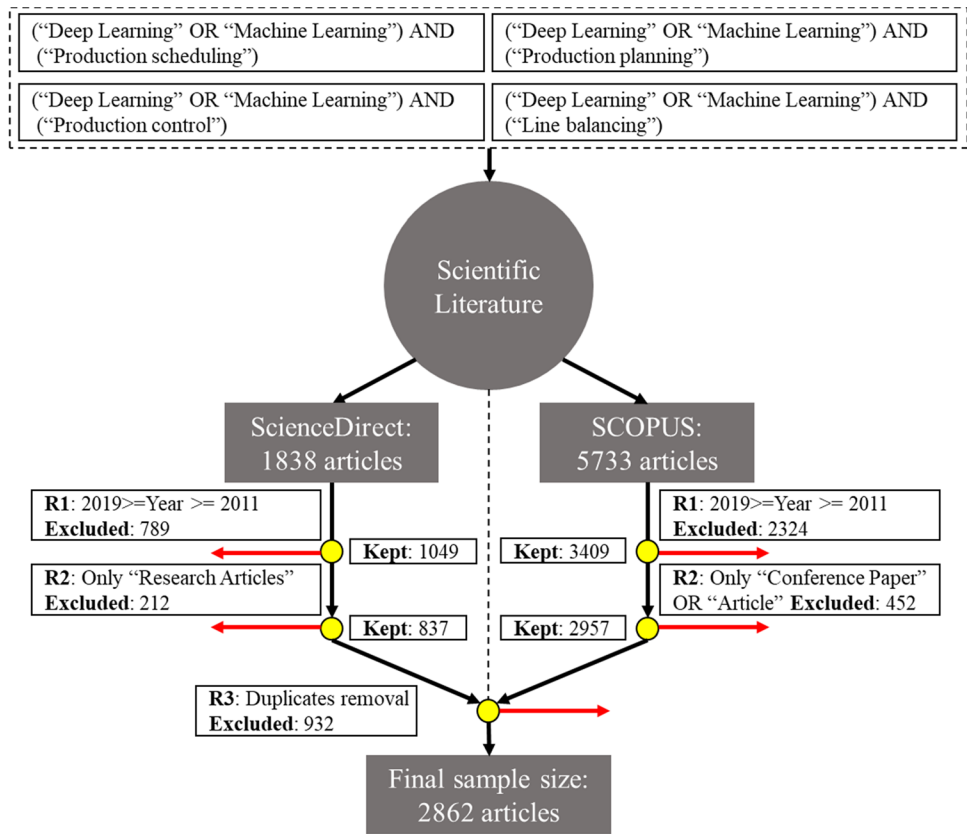


**Fig. 15** Search strategy detail for "Data Mining"
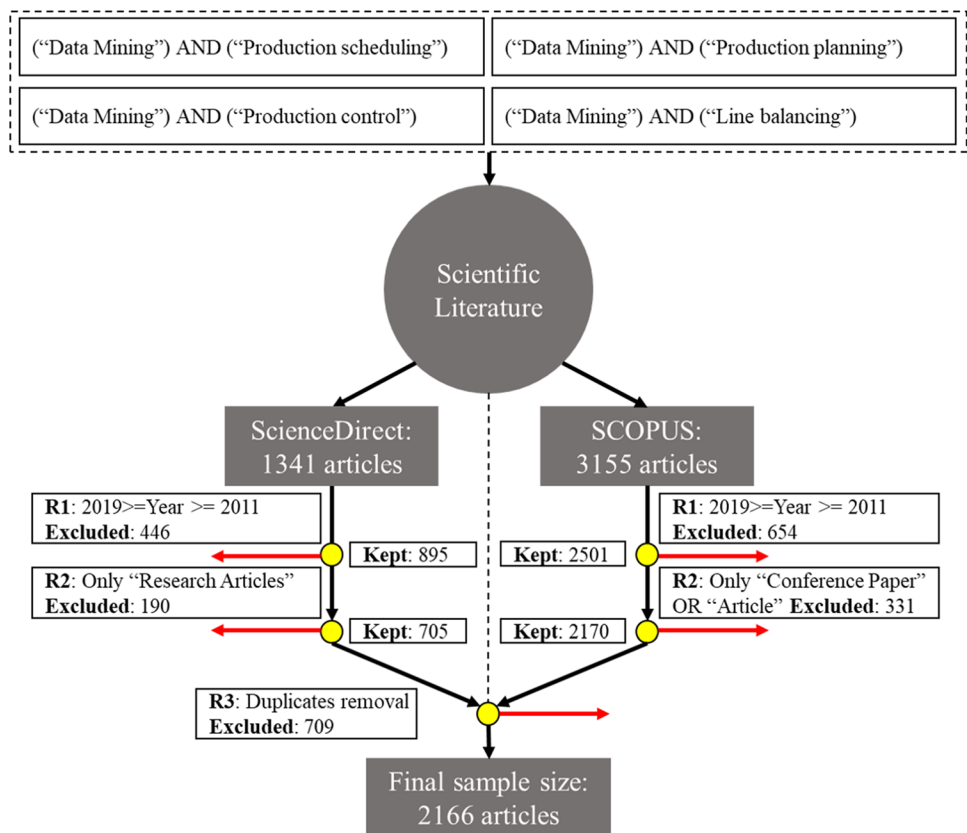
**Fig. 16** Search strategy detail for "Statistical Learning"

("Statistical Learning") AND ("Production scheduling")

("Statistical Learning") AND ("Production planning")

("Statistical Learning") AND ("Production control")

("Statistical Learning") AND ("Line balancing")

Scientific Literature

ScienceDirect: 126 articles

SCOPUS: 352 articles

**R1**: 2019>=Year >= 2011
**Excluded**: 31

**Kept**: 95

**Kept**: 282

**R1**: 2019>=Year >= 2011
**Excluded**: 70

**R2**: Only "Research Articles"
**Excluded**: 17

**Kept**: 78

**Kept**: 234

**R2**: Only "Conference Paper" OR "Article" **Excluded**: 48

**R3**: Duplicates removal
**Excluded**: 71

Final sample size: 241 articles

## Appendix II: Usage evolution of the top 6 most used techniques

See Figs. 17, 18, 19, 20, 21 and 22.
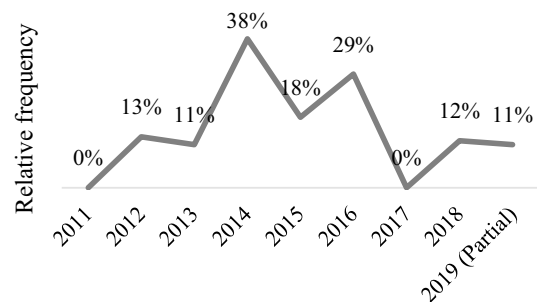


**Fig. 18** Usage evolution for Q-Learning



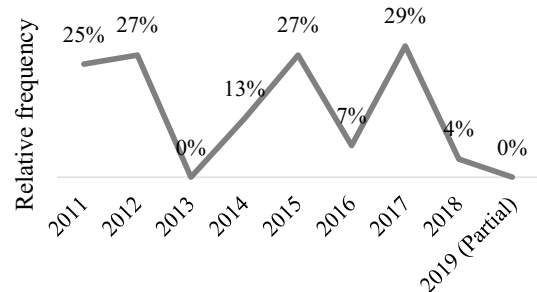**Fig. 17** Usage evolution for Neural Networks



**Fig. 19** Usage evolution for decision trees
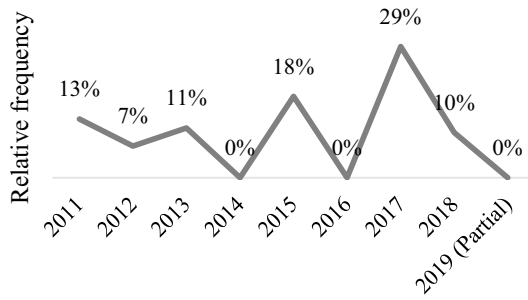
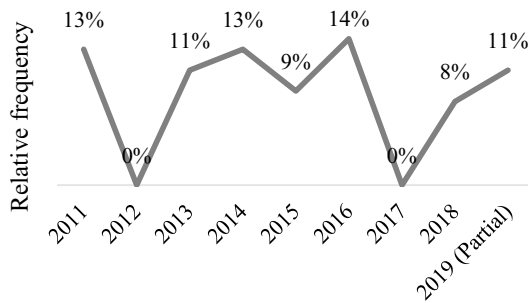**Fig. 20** Usage evolution for clustering



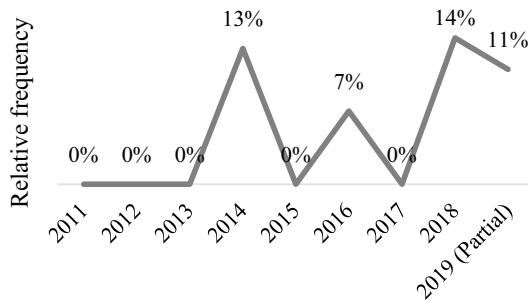**Fig. 21** Usage evolution for regression



**Fig. 22** Usage evolution for ensemble learning

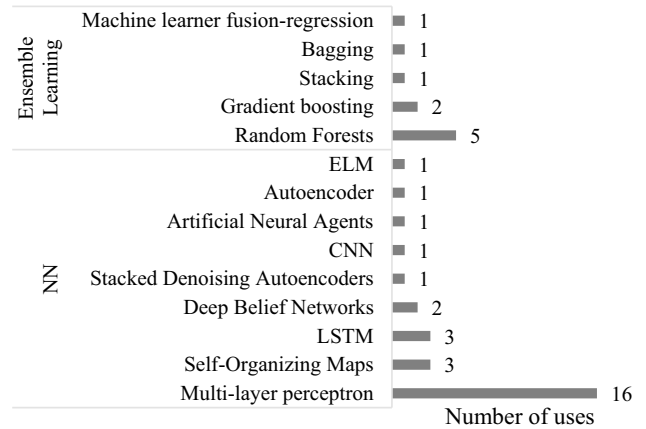## Appendix III: Detail on NN and Ensemble learning techniques

See Fig. 23.



**Fig. 23** Detail on the techniques of the NN and ensemble learning families

## References

Aissani, N., Bekrar, A., Trentesaux, D., & Beldjilali, B. (2012). Dynamic scheduling for multi-site companies: A decisional approach based on reinforcement multi-agent learning. *Journal of Intelligent Manufacturing, 23*(6), 2513–2529.

Altaf, M. S., Bouferguene, A., Liu, H., Al-Hussein, M., & Yu, H. (2018). Integrated production planning and control system for a panelized home prefabrication facility using simulation and RFID. *Automation in Construction, 85,* 369–383.

Bergmann, S., Feldkamp, N., & Strassburger, S. (2016). Approximation of dispatching rules for manufacturing simulation using data mining methods. In *2015 winter simulation conference* (pp. 2329–2340). Huntington Beach, USA.

Bi, J., & Zhang, C. (2018). An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. *Knowledge-Based Systems, 158,* 81–93.

Cai, B., Liu, H., & Xie, M. (2016). A real-time fault diagnosis methodology of complex systems using object-oriented Bayesian networks. *Mechanical Systems and Signal Processing, 80,* 31–44.

Cao, X. C., Chen, B. Q., Yao, B., & He, W. P. (2019). Combining translation-invariant wavelet frames and convolutional neural network for intelligent tool wear state identification. *Computers in Industry, 106,* 71–84.

Carvajal Soto, J. A., Tavakolizadeh, F., & Gyulai, D. (2019). An online machine learning framework for early detection of product failures in an Industry 4.0 context. *International Journal of Computer Integrated Manufacturing, 32*(4–5), 452–465. https://doi.org/10.1080/0951192X.2019.1571238.

Chen, C., Xia, B., Zhou, B. H., & Xi, L. (2015). A reinforcement learning based approach for a multiple-load carrier scheduling problem. *Journal of Intelligent Manufacturing, 26*(6), 1233–1245.

Curatolo, N., Lamouri, S., Huet, J. C., & Rieutord, A. (2014). A critical analysis of Lean approach structuring in hospitals. *Business Process Management Journal, 20*(3), 433–454.

Diaz-Rozo, J., Bielza, C., & Larrañaga, P. (2017). Machine learning-based CPS for clustering high throughput machining cycle conditions. In: *45th SME North American manufacturing research conference* (pp. 997–1008). Los Angeles, USA.

Ding, K., & Jiang, P. (2018). RFID-based production data analysis in an IoT-enabled smart job-shop. *IEEE/CAA Journal of Automatica Sinica, 5*(1), 128–138.

Dinis, D., Barbosa-Póvoa, A., & Teixeira, Â. P. (2019). Valuing data in aircraft maintenance through big data analytics: A probabilistic approach for capacity planning using Bayesian networks. *Computers & Industrial Engineering, 128,* 920–936.

Dolgui, A., Bakhtadze, N., Pyatetsky, V., Sabitov, R., Smirnova, G., Elpashev, D., & Zakharov, E. (2018). Data mining-based prediction of manufacturing situations data mining-based. In: *16th IFAC symposium on information control problems in manufacturing* (pp. 316–321). Bergamo, Italy: Elsevier B.V.

Doltsinis, S., Ferreira, P., & Lohse, N. (2014). An MDP model-based reinforcement learning approach for production station ramp-up optimization: Q-learning analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44*(9), 1125–1138.

Fotuhi, F., Huynh, N., Vidal, J. M., & Xie, Y. (2013). Modeling yard crane operators as reinforcement learning agents. *Research in Transportation Economics, 42*(1), 3–12.

Gao, X., Shang, C., Jiang, Y., Huang, D., & Chen, T. (2014). Refinery scheduling with varying crude: A deep belief network classification and multimodel approach. *AIChE Journal, 60*(7), 2525–2532.

Garengo, P., Biazzo, S., & Bititci, U. S. (2005). Performance measurement systems in SMEs: A review for a research agenda. *International Journal of Management Reviews, 7*(1), 25–47.

Grabot, B. (2018). Rule mining in maintenance: Analysing large knowledge bases. *Computers and Industrial Engineering, 139,* 1–15.

Gyulai, D., Kádár, B., & Monosotori, L. (2015). Robust production planning and capacity control for flexible assembly lines. In *15th IFAC symposium on information control problems in manufacturing* (pp. 2312–2317). Ottawa, Canada: Elsevier Ltd.

Gyulai, D., Kádár, B., & Monostori, L. (2014). Capacity planning and resource allocation in assembly systems consisting of dedicated and reconfigurable lines. In *8th international conference on digital enterprise technology* (pp. 185–191). Stuttgart, Germany: Elsevier B.V.

Gyulai, D., Pfeiffer, A., Bergmann, J., & Gallina, V. (2018a). Online lead time prediction supporting situation-aware production control. In *6th CIRP global web conference—Envisaging the future manufacturing, design, technologies and systems in innovation era* (pp. 190–195).

Gyulai, D., Pfeiffer, A., Nick, G., Gallina, V., Sihn, W., & Monostori, L. (2018b). Lead time prediction in a flow-shop environment with analytical and machine learning approaches. In *16th IFAC symposium on information control problems in manufacturing* (pp. 1029–1034). Bergamo, Italy.

Habib Zahmani, M., & Atmani, B. (2018). Extraction of dispatching rules for single machine total weighted tardiness using a modified genetic algorithm and data mining. *International Journal of Manufacturing Research, 13*(1), 1–25.

Hammami, Z., Mouelhi, W., & Ben Said, L. (2017). On-line self-adaptive framework for tailoring a neural-agent learning model addressing dynamic real-time scheduling problems. *Journal of Manufacturing Systems, 45,* 97–108.

Hammami, Z., Mouelhi, W., & Said, L. B. (2016). A self adaptive neural agent based decision support system for solving dynamic real time scheduling problems. In *10th international conference on intelligent systems and knowledge engineering* (pp. 494–501). Taipei, Taiwan.

Harding, J. A., Shahbaz, M., Srinivas, & Kusiak, A. (2006). Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering-Transactions of the ASME, 128*(4), 969–976.

Heger, J., Branke, J., Hildebrandt, T., & Scholz-Reiter, B. (2016). Dynamic adjustment of dispatching rule parameters in flow shops with sequence-dependent set-up times. *International Journal of Production Research, 54*(22), 6812–6824.

Hosseini, S., & Barker, K. (2016). A Bayesian network model for resilience-based supplier selection. *International Journal of Production Economics, 180,* 68–87.

Hosseini, S., & Ivanov, D. (2019). A new resilience measure for supply networks with the ripple effect considerations: A Bayesian network approach. *Annals of Operations Research*, 1–27. https://doi.org/10.1007/s10479-019-03350-8.

Hosseini, S., Ivanov, D., & Dolgui, A. (2019). Review of quantitative methods for supply chain resilience analysis. *Transportation Research Part E: Logistics and Transportation Review, 125*(December 2018), 285–307.

Huang, B., Wang, W., Ren, S., Zhong, R. Y., & Jiang, J. (2019). A proactive task dispatching method based on future bottleneck prediction for the smart factory. *International Journal of Computer Integrated Manufacturing, 32*(3), 278–293.

Ji, W., & Wang, L. (2017). Big data analytics based fault prediction for shop floor scheduling. *Journal of Manufacturing Systems, 43,* 187–194.

Jiang, S. L., Liu, M., Lin, J. H., & Zhong, H. X. (2016). A prediction-based online soft scheduling algorithm for the real-world steelmaking-continuous casting production. *Knowledge-Based Systems, 111,* 159–172.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.

Jurkovic, Z., Cukor, G., Brezocnik, M., & Brajkovic, T. (2018). A comparison of machine learning methods for cutting parameters prediction in high speed turning process. *Journal of Intelligent Manufacturing, 29*(8), 1683–1693.

Kartal, H., Oztekin, A., Gunasekaran, A., & Cebi, F. (2016). An integrated decision analytic framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. *Computers & Industrial Engineering, 101,* 599–613.

Khader, N., & Yoon, S. W. (2018). Online control of stencil printing parameters using reinforcement learning approach. In *28th international conference on flexible automation and intelligent manufacturing* (pp. 94–101). Columbus, USA: Elsevier B.V.

Kho, D. D., Lee, S., Zhong, R. Y. (2018). Big data analytics for processing time analysis in an IoT-enabled manufacturing shop floor. In *46th SME North American manufacturing research conference* (pp. 1411–1420). Texas, USA: Elsevier B.V.

Kim, H., & Lim, D.-E. (2018). Deep-learning-based storage-allocation approach to improve the AMHS throughput capacity in a semiconductor fabrication facility. In *Communications in computer and information science* (pp. 232–240). Springer Singapore.

Kim, S., & Nembhard, D. A. (2013). Rule mining for scheduling cross training with a heterogeneous workforce. *International Journal of Production Research, 51*(8), 2281–2300.

Kosmopoulos, D. I., Doulamis, N. D., & Voulodimos, A. S. (2012). Bayesian filter based behavior recognition in workflows allowing for user feedback. *Computer Vision and Image Understanding, 116*(3), 422–434.

Kretschmer, R., Pfouga, A., Rulhoff, S., & Stjepandić, J. (2017). Knowledge-based design for assembly in agile manufacturing by using data mining methods. *Advanced Engineering Informatics, 33,* 285–299.

Kruger, G. H., Shih, A. J., Hattingh, D. G., & Van Niekerk, T. I. (2011). Intelligent machine agent architecture for adaptive control optimization of manufacturing processes. *Advanced Engineering Informatics, 25*(4), 783–796.

Kumar, A., Shankar, R., & Thakur, L. S. (2018). A big data driven sustainable manufacturing framework for condition-based maintenance prediction. *Journal of Computational Science, 27,* 428–439.

Kusiak, A. (2017). Smart manufacturing must embrace big data. *Nature, 544*(7648), 23–25.

Kusiak, A. (2019). Fundamentals of smart manufacturing: A multi-thread perspective. *Annual Reviews in Control, 47,* 214–220.

Lai, L. K. C., & Liu, J. N. K. (2012). WIPA: Neural network and case base reasoning models for allocating work in progress. *Journal of Intelligent Manufacturing, 23*(3), 409–421.

Lai, X., Shui, H., & Ni, J. (2018). A two-layer long short-term memory network for bottleneck prediction in multi-job manufacturing systems. In *13th international manufacturing science and engineering conference* (p. V003T02A014). Texas, USA.

Lemieux, A. A., Lamouri, S., Pellerin, R., & Tamayo, S. (2015). Development of a leagile transformation methodology for product development. *Business Process Management Journal, 21*(4), 791–819.

Leng, J., Chen, Q., Mao, N., & Jiang, P. (2018). Combining granular computing technique with deep learning for service planning under social manufacturing contexts. *Knowledge-Based Systems, 143,* 295–306.

Li, D. C., Chen, C. C., Chen, W. C., & Chang, C. J. (2012a). Employing dependent virtual samples to obtain more manufacturing information in pilot runs. *International Journal of Production Research, 50*(23), 6886–6903.

Li, X., Duan, F., Loukopoulos, P., Bennett, I., & Mba, D. (2018). Canonical variable analysis and long short-term memory for fault diagnosis and performance estimation of a centrifugal compressor. *Control Engineering Practice, 72*(January), 177–191.

Li, X., Wang, J., & Sawhney, R. (2012b). Reinforcement learning for joint pricing, lead-time and scheduling decisions in make-to-order systems. *European Journal of Operational Research, 221*(1), 99–109.

Li, L., Zijin, S., Jiacheng, N., & Fei, Q. (2013). Data-based scheduling framework and adaptive dispatching rule of complex manufacturing systems. *International Journal of Advanced Manufacturing Technology, 66*(9–12), 1891–1905.

Liao, Q. (2018). Study of SVM-based intelligent dispatcher for parallel machines scheduling with sequence-dependent setup times. In *6th international conference on mechanical, automotive and materials engineering, CMAME 2018* (pp. 46–50). Hong Kong: IEEE.

Lieber, D., Stolpe, M., Konrad, B., Deuse, J., Morik, K. (2013). Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. In *46th CIRP conference on manufacturing systems 2013* (pp. 193–198). Setúbal, Portugal: Elsevier B.V.

Lingitz, L., Gallina, V., Ansari, F., Gyulai, D., Pfeiffer, A., & Sihn, W. (2018). Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer. In *51st CIRP conference on manufacturing systems* (pp. 1051–1056). Stockholm, Sweden.

Llave, M. R. (2018). Data lakes in business intelligence: Reporting from the trenches. In *CENTERIS/ProjMAN/HCist 2018* (pp. 516–524). Lisbon, Portugal: Elsevier B.V.

Lo Giudice, P., Musarella, L., Sofo, G., & Ursino, D. (2019). An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. *Information Sciences, 478,* 606–626.

Lubosch, M., Kunath, M., & Winkler, H. (2018). Industrial scheduling with Monte Carlo tree search and machine learning. In *51st CIRP conference on manufacturing systems* (pp. 1283–1287). Stockholm, Sweden: Elsevier B.V.

Lv, Y., Qin, W., Yang, J., & Zhang, J. (2018a). Adjustment mode decision based on support vector data description and evidence theory for assembly lines. *Industrial Management and Data Systems, 118*(8), 1711–1726.

Lv, S., Zheng, B., Kim, H., & Yue, Q. (2018b). Data mining for material feeding optimization of printed circuit board template production. *Journal of Electrical and Computer Engineering*. https://doi.org/10.1155/2018/1852938.

Ma, Y., Qiao, F., Zhao, F., & Sutherland, J. (2017). Dynamic scheduling of a semiconductor production line based on a composite rule set. *Applied Sciences, 7*(10), 1052.

Maghrebi, M., Shamsoddini, A., & Waller, S. T. (2016). Fusion based learning approach for predicting concrete pouring productivity based on construction and supply parameters. *Construction Innovation, 16*(2), 185–202.

Manns, M., Wallis, R., & Deuse, J. (2015). Automatic proposal of assembly work plans with a controlled natural language. In *9th CIRP conference on intelligent computation in manufacturing engineering* (pp. 345–350). Capri, Italy.

Manupati, V. K., Anand, R., Thakkar, J. J., Benyoucef, L., Garsia, F. P., & Tiwari, M. K. (2013). Adaptive production control system for a flexible manufacturing cell using support vector machine-based approach. *International Journal of Advanced Manufacturing Technology, 67*(1–4), 969–981.

Mikołajczyk, A., & Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop, IIPhDW 2018* (pp. 117–122). Swinoujście, Poland: IEEE.

Mitchell, T. (1997). *Machine learning* (Vol. 2). New York: McGraw-Hill.

Moeuf, A., Pellerin, R., Lamouri, S., Tamayo-Giraldo, S., & Barbaray, R. (2018). The industrial management of SMEs in the era of Industry 4.0. *International Journal of Production Research, 56*(3), 1118–1136.

Mori, J., & Mahalec, V. (2015). Planning and scheduling of steel plates production. Part I: Estimation of production times via hybrid Bayesian networks for large domain of discrete variables. *Computers & Chemical Engineering, 79,* 113–134.

Ou, X., Chang, Q., Arinez, J., & Zou, J. (2018). Gantry work cell scheduling through reinforcement learning with knowledge-guided reward setting. *IEEE Access, 6,* 14699–14709.

Ou, X., Chang, Q., & Chakraborty, N. (2019). Simulation study on reward function of reinforcement learning in gantry work cell scheduling. *Journal of Manufacturing Systems, 50,* 1–8.

Palombarini, J., & Martínez, E. (2012). SmartGantt—An intelligent system for real time rescheduling based on relational reinforcement learning. *Expert Systems with Applications, 39*(11), 10251–10268.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv e-prints, arXiv:1712.04621.

Priore, P., Ponte, B., Puente, J., & Gómez, A. (2018). Learning-based scheduling of flexible manufacturing systems using ensemble methods. *Computers & Industrial Engineering, 126,* 282–291.

Qu, S., Chu, T., Wang, J., Leckie, J., & Jian, W. (2015). A centralized reinforcement learning approach for proactive scheduling in manufacturing. In *IEEE international conference on emerging technologies and factory automation*. Luxembourg, Luxembourg.

Qu, S., Jie, W., & Shivani, G. (2016a). Learning adaptive dispatching rules for a manufacturing process system by using reinforcement learning approach. In *IEEE international conference on emerging technologies and factory automation*. Berlin, Germany.

Qu, S., Wang, J., Govil, S., & Leckie, J. O. (2016b). Optimized adaptive scheduling of a manufacturing process system with multi-skill workforce and multiple machine types: An ontology-based, multi-agent reinforcement learning approach. In *49th CIRP conference on manufacturing systems (CIRP-CMS 2016)* (pp. 55–60). Stuttgart, Germany: Elsevier B.V.

Rainer, C. (2013). Data mining as technique to generate planning rules for manufacturing control in a complex production system. In K. Windt (Ed.), *Robust manufacturing control*. Heidelberg: Springer.

Reboiro-Jato, M., Glez-Dopazo, J., Glez, D., Laza, R., Gálvez, J. F., Pavón, R., et al. (2011). Using inductive learning to assess compound feed production in cooperative poultry farms. *Expert Systems with Applications, 38*(11), 14169–14177.

Reuter, C., Brambring, F., Weirich, J., & Kleines, A. (2016). Improving data consistency in production control by adaptation of data mining algorithms. In *9th international conference on digital enterprise technology* (pp. 545–550). Nanjing, China.

Rostami, H., Blue, J., & Yugma, C. (2018). Automatic equipment fault fingerprint extraction for the fault diagnostic on the batch process data. *Applied Soft Computing, 68,* 972–989.

Ruessmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., et al. (2015). Industry 4.0: The future of productivity and growth in manufacturing. *The Boston Consulting Group, 9,* 54–89.

Sahebjamnia, N., Tavakkoli-Moghaddam, R., & Ghorbani, N. (2016). Designing a fuzzy Q-learning multi-agent quality control system for a continuous chemical production line—A case study. *Computers & Industrial Engineering, 93,* 215–226.

Schuh, G., Prote, J. P., Luckert, M., & Hünnekes, P. (2017a). Knowledge discovery approach for automated process planning. In *50th CIRP conference on manufacturing systems knowledge* (pp. 539–544). Taichung, Taiwan.

Schuh, G., Reinhart, G., Prote, J. P., Sauermann, F., Horsthofer, J., Oppolzer, F., & Knoll, D. (2019). Data mining definitions and applications for the management of production complexity. In *52nd CIRP conference on manufacturing systems* (pp. 874–879). Ljubljana, Slovenia: Elsevier B.V.

Schuh, G., Reuter, C., Prote, J. P., Brambring, F., & Ays, J. (2017b). Increasing data integrity for improving decision making in production planning and control. *CIRP Annals—Manufacturing Technology, 66*(1), 425–428.

Shahzad, A., & Mebarki, N. (2012). Data mining based job dispatching using hybrid simulation-optimization approach for shop scheduling problem. *Engineering Applications of Artificial Intelligence, 25*(6), 1173–1181.

Sharp, M., Ak, R., & Hedberg, T. (2018). A survey of the advancing use and development of machine learning in smart manufacturing. *Journal of Manufacturing Systems, 48,* 170–179.

Shiue, Y. R., Guh, R. S., & Tseng, T. Y. (2012). Study on shop floor control system in semiconductor fabrication by self-organizing map-based intelligent multi-controller. *Computers & Industrial Engineering, 62*(4), 1119–1129.

Shiue, Y. R., Lee, K. C., & Su, C. T. (2018). Real-time scheduling for a smart factory using a reinforcement learning approach. *Computers & Industrial Engineering, 125*(101), 604–614.

Solti, A., Raffel, M., Romagnoli, G., & Mendling, J. (2018). Misplaced product detection using sensor data without planograms. *Decision Support Systems, 112,* 76–87.

Stein, N., Meller, J., & Flath, C. M. (2018). Big data on the shop-floor: Sensor-based decision-support for manual processes. *Journal of Business Economics, 88*(5), 593–616.

Stricker, N., Kuhnle, A., Sturm, R., & Friess, S. (2018). Reinforcement learning for adaptive order dispatching in the semiconductor industry. *CIRP Annals—Manufacturing Technology, 67*(1), 511–514.

Talhi, A., Fortineau, V., Huet, J. C., & Lamouri, S. (2017). Ontology for cloud manufacturing based product lifecycle management. *Journal of Intelligent Manufacturing, 30*(5), 1–22.

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems, 48,* 157–169.

Thomas, T. E., Koo, J., Chaterji, S., & Bagchi, S. (2018b). Minerva: A reinforcement learning-based technique for optimal scheduling and bottleneck detection in distributed factory operations. In *10th international conference on communication systems and networks* (pp. 129–136). Bengaluru, India.

Thomas, A., Noyel, M., Zimmermann, E., Suhner, M.-C., Bril El Haouzi, H., & Thomas, P. (2018a). Using a classifier ensemble for proactive quality monitoring and control: The impact of the choice of classifiers types, selection criterion, and fusion process. *Computers in Industry, 99*(March), 193–204.

Tian, G., Zhou, M., & Chu, J. (2013). A chance constrained programming approach to determine the optimal disassembly sequence. *IEEE Transactions on Automation Science and Engineering, 10*(4), 1004–1013.

Tong, Y., Li, J., Li, S., & Li, D. (2016). Research on energy-saving production scheduling based on a clustering algorithm for a forging enterprise. *Sustainability, 8*(2), Article number 136.

Tony Arnold, J. R., Chapman, S. N., & Clive, L. M. (2012). *Introduction to materials management* (Vol. 118). Pearson: London.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14,* 207–222.

Tuncel, E., Zeid, A., & Kamarthi, S. (2014). Solving large scale disassembly line balancing problem with uncertainty using reinforcement learning. *Journal of Intelligent Manufacturing, 25*(4), 647–659.

Wang, C., & Jiang, P. (2018). Manifold learning based rescheduling decision mechanism for recessive disturbances in RFID-driven job shops. *Journal of Intelligent Manufacturing, 29*(7), 1485–1500.

Wang, C., & Jiang, P. (2019). Deep neural networks based order completion time prediction by using real-time job shop RFID data. *Journal of Intelligent Manufacturing, 30*(3), 1303–1318.

Wang, H., Jiang, Z., Zhang, X., Wang, Y., & Wang, Y. (2017). A fault feature characterization based method for remanufacturing process planning optimization. *Journal of Cleaner Production, 161,* 708–719.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018a). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems, 48,* 144–156.

Wang, C. L., Rong, G., Weng, W., & Feng, Y. P. (2015). Mining scheduling knowledge for job shop scheduling problem. In *15th IFAC symposium on information control problems in manufacturing* (pp. 800–805). Ottawa, Canada: Elsevier Ltd.

Wang, H. X., & Yan, H. Sen. (2016). An interoperable adaptive scheduling strategy for knowledgeable manufacturing based on SMGWQ-learning. *Journal of Intelligent Manufacturing, 27*(5), 1085–1095.

Wang, J., Yang, J., Zhang, J., Wang, X., & Zhang, W. (2018b). Big data driven cycle time parallel prediction for production planning in wafer manufacturing. *Enterprise Information Systems, 12*(6), 714–732.

Wang, J., Zhang, J., & Wang, X. (2018c). Bilateral LSTM: A two-dimensional long short-term memory model with multiply memory units for short-term cycle time forecasting in re-entrant manufacturing systems. *IEEE Transactions on Industrial Informatics, 14*(2), 748–758.

Waschneck, B., Bauernhansl, T., Knapp, A., Kyek, A. (2018). Optimization of global production scheduling with deep reinforcement learning. In *51st CIRP conference on manufacturing systems* (pp. 1264–1269). Stockholm, Sweden.

Wauters, T., Verbeeck, K., Verstraete, P., Vanden Berghe, G., & De Causmaecker, P. (2012). Real-world production scheduling for the food industry: An integrated approach. *Engineering Applications of Artificial Intelligence, 25*(2), 222–228.

Wu, W., Ma, Y., Qiao, F., & Gu, X. (2015). Data mining based dynamic scheduling approach for semiconductor manufacturing system. In *34th Chinese control conference* (pp. 2603–2608). Hangzhou, China.

Xanthopoulos, A. S., Kiatipis, A., Koulouriotis, D. E., & Stieger, S. (2017). Reinforcement learning-based and parametric production-maintenance control policies for a deteriorating manufacturing system. *IEEE Access, 6*, 576–588.

Yang, Z., Zhang, P., & Chen, L. (2016). RFID-enabled indoor positioning method for a real-time manufacturing execution system using OS-ELM. *Neurocomputing, 174*, 121–133.

Yeh, D. Y., Cheng, C. H., & Hsiao, S. C. (2011). Classification knowledge discovery in mold tooling test using decision tree algorithm. *Journal of Intelligent Manufacturing, 22*(4), 585–595.

Yuan, B., Wang, L., & Jiang, Z. (2014). Dynamic parallel machine scheduling using the learning agent. In *2013 IEEE international conference on industrial engineering and engineering management* (pp. 1565–1569). Bangkok, Thailand.

Zellner, G. (2011). A structured evaluation of business process improvement approaches. *Business Process Management Journal, 17*(2), 203–237.

Zhang, Z., Zheng, L., Hou, F., & Li, N. (2011). Semiconductor final test scheduling with Sarsa($\lambda$, k) algorithm. *European Journal of Operational Research, 215*(2), 446–458.

Zhang, Z., Zheng, L., Li, N., Wang, W., Zhong, S., & Hu, K. (2012). Minimizing mean weighted tardiness in unrelated parallel machine scheduling with reinforcement learning. *Computers & Operations Research, 39*(7), 1315–1324.

Zhong, R. Y., Huang, G. Q., Dai, Q. Y., & Zhang, T. (2014). Mining SOTs and dispatching rules from RFID-enabled real-time shopfloor production data. *Journal of Intelligent Manufacturing, 25*(4), 825–843.

Zhong, R. Y., Newman, S. T., Huang, G. Q., & Lan, S. (2016). Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Computers & Industrial Engineering, 101*, 572–591.

Zhou, P., Guo, D., & Chai, T. (2018). Data-driven predictive control of molten iron quality in blast furnace ironmaking using multi-output LS-SVR based inverse system identification. *Neurocomputing, 308*, 101–110.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing, 237*(January), 350–361.