Artificial intelligence for throughput bottleneck analysis – State-of-the-art and future directions

Citation for the original published paper (version of record):
Subramaniyan, M., Skoogh, A., Bokrantz, J. et al (2021). Artificial intelligence for throughput bottleneck analysis – State-of-the-art and future directions. Journal of Manufacturing Systems, 60: 734-751.
http://dx.doi.org/10.1016/j.jmsy.2021.07.021

Journal of Manufacturing Systems 60 (2021) 734–751

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

journal homepage: www.elsevier.com/locate/jmansys

Review

Artificial intelligence for throughput bottleneck analysis – State-of-the-art and future directions

Mukund Subramaniyan [a,*], Anders Skoogh [a], Jon Bokrantz [a], Muhammad Azam Sheikh [b], Matthias Thürer [c], Qing Chang [d,e]

[a] Department of Industrial and Materials Science, Chalmers University of Technology, Gothenburg, 41296, Sweden
[b] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, 41296, Sweden
[c] School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai, 519070, China
[d] Department of Mechanical and Aerospace Engineering, University of Virginia, Charlottesville, VA, 22904, USA
[e] Department of Engineering Systems and Environment, University of Virginia, Charlottesville, VA, 22904, USA

ARTICLE INFO

Keywords:
Throughput bottlenecks
Artificial intelligence
Production system
Data-driven
Manufacturing

ABSTRACT

Identifying, and eventually eliminating throughput bottlenecks, is a key means to increase throughput and productivity in production systems. In the real world, however, eliminating throughput bottlenecks is a challenge. This is due to the landscape of complex factory dynamics, with several hundred machines operating at any given time. Academic researchers have tried to develop tools to help identify and eliminate throughput bottlenecks. Historically, research efforts have focused on developing analytical and discrete event simulation modelling approaches to identify throughput bottlenecks in production systems. However, with the rise of industrial digitalisation and artificial intelligence (AI), academic researchers explored different ways in which AI might be used to eliminate throughput bottlenecks, based on the vast amounts of digital shop floor data. By conducting a systematic literature review, this paper aims to present state-of-the-art research efforts into the use of AI for throughput bottleneck analysis. To make the work of the academic AI solutions more accessible to practitioners, the research efforts are classified into four categories: (1) identify, (2) diagnose, (3) predict and (4) prescribe. This was inspired by real-world throughput bottleneck management practice. The categories, identify and diagnose focus on analysing historical throughput bottlenecks, whereas predict and prescribe focus on analysing future throughput bottlenecks. This paper also provides future research topics and practical recommendations which may help to further push the boundaries of the theoretical and practical use of AI in throughput bottleneck analysis.

1. Introduction

One of the grand themes that have permeated manufacturing is the endless pursuit of productivity [1] (p.341). Manufacturing companies must increase their factory floor productivity to remain cost-efficient and competitive. Productivity is often measured in terms of "throughput". Throughput is defined as the number of products produced in a unit time interval [2] (p. 3823). In real-world production systems, practitioners constantly devote their efforts to improving system throughput. However, they often find a sizable gap between target and actual throughput. Recent empirical studies show throughput losses in real-world production systems to be 20 %–30 % ([3] (p.831), [4] (p.7278)). These losses are partly due to the existence of "throughput bottlenecks". Throughput bottlenecks are machines which constrain throughput. They may occur in a system due to variability in the time duration of production flow disturbances, such as unplanned stops in machines, variable processing times of machines, setups, operator delays and so on [5,6].

To improve throughput, multiple operations management theories (cf. Theory of Constraints (ToC) [7], swift even flow [8](p.102)) recommend eliminating throughput bottlenecks. However, this is easier in theory than in practice. Academic researchers have helped practitioners identify and analyse throughput bottlenecks by developing analytical approaches (such as system-theoretic analysis based on recursive equations) [2,4,9,10] and discrete event simulation model-based approaches (building discrete event simulation models of

* Corresponding author.
E-mail address: mukunds@chalmers.se (M. Subramaniyan).

https://doi.org/10.1016/j.jmsy.2021.07.021
Received 18 December 2020; Received in revised form 3 June 2021; Accepted 19 July 2021
Available online 11 August 2021

production systems) [11–16]. Although these approaches have helped analyse throughput bottlenecks, they were better suited to static analysis and more useful for early configuration of production systems [17]. Each time the production system changes, new equations may have to be manually derived (in the case of analytical approaches) and existing models updated, or sometimes new models created (in the case of simulation approaches). This is a time-consuming and expensive task.

Recently, artificial intelligence (AI) in manufacturing has been propelled from hype to reality [18–20] by a convergence of algorithmic advances, data proliferation due to increased digitalisation, reduced data storage costs and a tremendous increase in computing power. It has now become possible for practitioners to better address the challenges of analysing throughput bottlenecks by using AI. Using production system data, the dynamics of a production system can be understood in detail, throughput bottlenecks automatically identified and diagnosed, and then possible actions prescribed. Over the last decade, academic researchers successfully used AI for throughput bottleneck analysis in production systems. They have thus advanced the field significantly from its earlier analytical and simulation-based approaches. However, there is no currently available comprehensive presentation of the state-of-the-art in AI for throughput bottleneck analysis, summarising what has been achieved and how AI has helped advance the field. Most importantly, there is no account of the challenges encountered, and the research opportunities this may provide. This information is also important for practitioners who are increasingly showing interest to implement AI solutions for throughput bottleneck analysis [21](p.7). But before the implementation, they need to identify different AI solutions and analyse their usefulness in the specific context of their factory. This study provides support and guidelines to make it easier for them to navigate through the different AI solutions and to select the appropriate AI solution.

Therefore, the purpose of this paper is to present the state-of-the-art of academic literature on the use of AI for throughput bottleneck analysis. Specifically, this paper makes three contributions: (1) it presents a classification structure (identifying, diagnosing, predicting and prescribing) for existing research efforts, with special emphasis on its impact in improving real-world, shop-floor throughput bottleneck management, (2) it describes the state of the art of AI (in terms of input data, modelling approach and output) for throughput bottlenecks analysis; and (3) it provides a wide range of future research directions, and a series of practical recommendations, to influence the future development and use of AI for throughput bottleneck analysis in practice.

The remainder of this paper is structured as follows. Section 2 presents the theoretical background on throughput bottlenecks and AI. Section 3 presents the methodology adopted to conduct our systematic literature review. Section 4 presents the state-of-the-art research in AI for throughput bottleneck analysis. Section 5 presents future research directions. This is followed by practical recommendations in Section 6. Section 7 provides the limitations of the study. Finally, Section 8 concludes the paper, summarising the main conclusions of this study.

2. Theoretical background

To understand state-of-the-art results in throughput bottleneck analysis, readers must be knowledgeable about the fundamentals of throughput bottlenecks and AI. Therefore, in this section, the theory of throughput bottlenecks will be presented first. The main challenges of throughput bottleneck elimination in real-world shop floor practices are then explained. These will help the readers to comprehensively understand the phenomenon of throughput bottlenecks in production systems. Also, it will make the results of the paper broadly accessible to a larger audience. Finally, there will be an overview of AI within the manufacturing domain.

2.1. Theory of throughput bottlenecks

According to [22], to explain any theory, it is important to understand four building blocks of that theory: (1) object of study, (2) concepts, (3) propositions and (4) domain.

This section explains the theory of throughput bottlenecks by identifying and describing its building blocks.

Fig. 1 shows a conceptual model of the building blocks of throughput bottleneck theory. In this figure, the outermost dashed block represents the domain, the solid inner block denotes the object of study, the three innermost blocks represent the construct, and the arrows represent the propositions. Each building block is then explained.

2.1.1. Object of study

As in this paper, the focus is on studying throughput bottlenecks in a production system. Thus, the object of study is the "production system".

2.1.2. Concepts

Understanding throughput bottlenecks within a production system requires three concepts: (1) variability, (2) throughput bottlenecks and (3) throughput. Variability is defined as "any deviation from absolute regularity" [23] (p.5). Variability focuses on the stochastic effects of machines in a production system [24,25,6]. Stochastic effects are caused by process time variabilities [26]. These are random events such as unplanned stops, variations in the processing times, setups, operator delays and so on. It is important to note that these variabilities may be different for different machines in a production system and may change with time. Throughput is defined as "the number of products per unit time interval from the production system" [2] (p. 3823). A unit time interval may be on any time scale, such as a shift, day, week or month. Throughput is a measure of the performance of a production system [2]. Throughput bottlenecks are defined as "the machines whose performance impedes the overall system performance most strongly" [13] (p.5019). For example, take a production system with ten machines. If two of them impede throughput from the production system more strongly than the other machines, then these two are called throughput bottlenecks.

2.1.3. Propositions

The relationships between variability, throughput and throughput bottlenecks are explained in the following sentences. *Arrow A* is the relationship between variability and throughput [27]. The greater the processing time and flow time variabilities, the greater the fluctuation in production system throughput. For example [28,29], showed that variable processing times in a machine produced significantly lower throughput. *Arrow B* is the relationship between variability and throughput bottlenecks. Variability in a set of machines affects throughput more than other machines [7]. These sets of machines are called throughput bottlenecks [12,13,30]. *Arrow* C is the relationship between throughput bottlenecks and production system throughput. Eliminating throughput bottlenecks leads to increased throughput [31, 32].

2.1.4. Domain

The concepts and propositions of throughput bottlenecks are intended to hold and be generalisable within the domain of discrete flow line production systems (or commonly called as flow shops) with or without buffers (such as automotive machining production systems or assembly production systems) [11,33,34].

To summarise, the theory of throughput bottlenecks explains how the variability of random events in machines brings about throughput bottlenecks and constrains production system throughput. Once throughput bottlenecks are eliminated, greater throughput is obtained. But, when throughput bottlenecks are eliminated, the system dynamics change, affecting processing time and flow time variabilities. A new set of machines will then emerge as throughput bottlenecks [35]. To

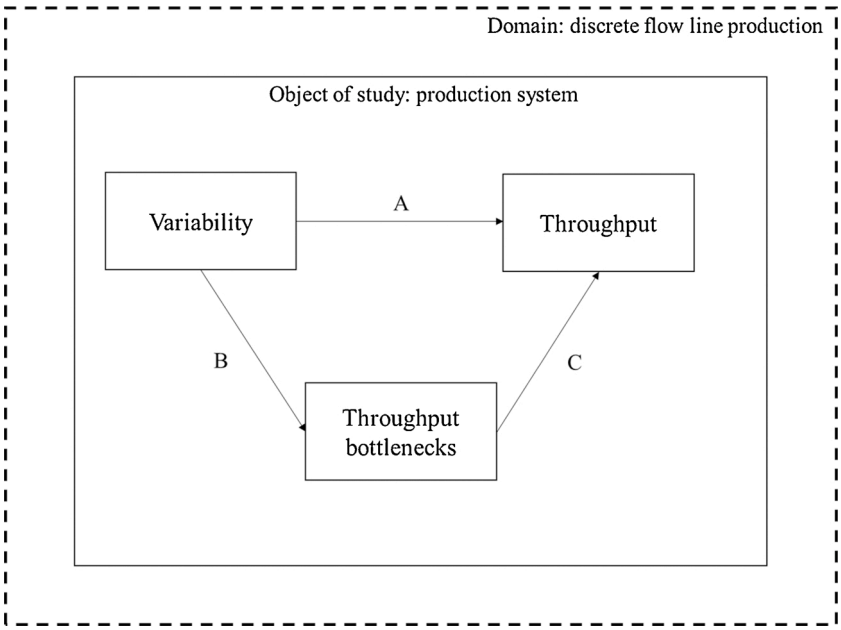

Fig. 1. Conceptual model of throughput bottlenecks in production systems.

achieve higher throughput, these new throughput bottlenecks need to be identified and eliminated. This cycle of identifying, analysing and eliminating throughput bottlenecks is a continuous process in real-world production systems until a desired level of throughput is reached.

2.2. Practical challenges of eliminating throughput bottlenecks

To successfully realise the cycle of identifying and eliminating throughput bottlenecks, practitioners need to repeatedly analyse throughput bottlenecks at different time intervals. However, on the real-world shop floor, there are four frequent challenges in identifying and eliminating throughput bottlenecks: (1) time frame (2) shiftiness (3) multiple root causes and (4) equivocality. Addressing these four challenges will have a maximum impact on the production system throughput. These challenges are explained below.

2.2.1. Time frame

Throughput bottlenecks need to be identified and analysed in different time frames to allow the planning and execution of throughput improvement actions [2]. For example, on a shop floor, practitioners need to analyse short-term throughput bottlenecks (such as machines behaving like bottlenecks within a production shift) and take rapid short-term action (running those bottlenecks during breaks, ensuring they are not starved and so on). This allows practitioners to reduce throughput fluctuations and achieve the target short-term throughput. At the same time, practitioners need to analyse long-term throughput bottlenecks (such as machines behaving like bottlenecks over multiple production runs). This allows them to take long-term actions (making suitable jigs and fixtures to simplify the workload or allocating buffer spaces before a bottleneck) and substantially increase long-term throughput.

2.2.2. Shiftiness

Throughput bottlenecks are dynamic on the shop-floor [11]. In other words, the throughput bottleneck location shifts from one machine to another in a production system. For example, on a machining production line, a practitioner might realise that a milling machine in the production system was the throughput bottleneck in a previous production shift, whereas in the current production shift a grinding machine has become the throughput bottleneck. Throughput bottleneck locations shift for three reasons: (1) variability in process times, (2) product mix and (3) practitioners' actions. Process time variabilities occur due to random processing times, unplanned stops and so on. This changes the system dynamics, leading to a shift in the throughput bottleneck location [6]. Similarly, different products may have different processing times on different machines in the production line, leading to shifting throughput bottleneck locations [36]. Finally, if practitioners take action to eliminate the current throughput bottlenecks, then this also changes the system dynamics and new throughput bottlenecks emerge [35]. Therefore, practitioners need to monitor the location of throughput bottlenecks continuously and act quickly.

2.2.3. Multiple root causes

Multiple root causes (such as random variations in cycle time, minor stops, setup times and different causes of unplanned stops) may cause throughput bottlenecks to emerge in a production system [26,10]. Often, there is no single root cause of throughput bottlenecks [37], with root causes appearing in combination. For example, a machine might behave like a throughput bottleneck because it has greater random variations in cycle times and longer unplanned stops. In such situations, practitioners need to prioritise the right root causes of the throughput bottleneck which, upon being eliminated, gives maximum improvement in throughput.

2.2.4. Equivocality

Among production and maintenance practitioners, there is ambiguity regarding the identification of throughput bottlenecks. This is because there are differing views on throughput bottlenecks (such as cycle time bottlenecks, downtime bottlenecks, setup time bottlenecks), making it difficult to reach a consensus on selecting the right set of bottleneck machines. For example, production practitioners might claim that only the machine with the highest cycle time (without considering other process time variabilities) constitutes a throughput bottleneck [38]. Maintenance practitioners, on the other hand, might claim that the machine with the highest downtime (without considering other variabilities) constitutes the bottleneck [39]. Therefore, there is no consensus on correctly identifying a bottleneck. The impact of this ambiguity is that practitioners might take a set of actions (based on their views) that are ineffective in eliminating the bottleneck.

2.3. Artificial intelligence in manufacturing

AI has no single accepted definition [40] (p.119). It is subject to multiple interpretations by researchers from different fields (computer science, mathematics and so on) which makes establishing a common understanding of AI challenging. The focus of this paper is not on elaborating on and arguing about these differing interpretations by researchers in various fields. Rather, this paper is interested in exploring the usefulness of AI in solving the problem of throughput bottleneck analysis within a manufacturing context. This means it is important to understand AI and its applications within the manufacturing field. This will promote appreciation and interpretation of the role of AI in throughput bottleneck analysis.

Within manufacturing, academic scholars and practitioners tend to view the goal of AI as making computers more intelligent. This is done by processing large amounts of data, uncovering hidden patterns and unknown correlations, learning from the data and finding the best possible solutions to real-world problems [41,42]. A variety of tools are used to facilitate this process; statistical tools, rule-based methods, machine learning (ML), deep learning (DL), reinforcement learning, probabilistic graphical models, soft computing, knowledge representation (such as knowledge graphs), game theory, and even traditional computer algorithms (such as planning and search algorithms). However, it is also important to note that the existing manufacturing literature is ambiguous regarding which set of tools belongs to AI. For example [43], and [44] consider only DL (which is based on neural networks) as a single type of AI whereas [45] (p.1596) shows a relationship describing DL as a subset of ML, which is a subset of AI. In real-world manufacturing practice, it needs to be acknowledged that ML and AI are often used interchangeably. In this paper, AI is considered to be a set of tools used to process data, extract patterns and learn from that data.

It is also important to acknowledge that, in manufacturing academic literature, AI also appears under multiple names. These include big data analytics (cf. [46]), data mining (cf [47].), predictive modelling (cf [48].), data science (cf [49].), pattern recognition (cf [50].) and data-driven (cf [51].), where the goal is also to process huge volumes of data, learn from it and make computers more intelligent. A general glossary for AI may be found in [52].

An AI solution to any given problem has three important aspects: (1) desired output, (2) input data and (3) modelling approach [53,54]. AI solutions always begin by defining their desired output. To successfully achieve that output, input data must then be defined. Understanding the type of input data is critical to understanding how AI is used in solving the problem. Finally, it is important to understand the steps within a given modelling approach (such as feature engineering, classification) used to obtain a desired output from the input data.

3. Review methodology

Understanding the current state of the art of AI of throughput bottleneck analysis means collecting and analysing relevant research work from the existing literature. To do this, the guidelines published in [55] and [56] are adapted (examples of this adaptation are also described in [57,58]). The three steps, defined by [55] and [56], for systematically collecting and analysing research efforts are: (1) material collection, (2) descriptive analysis and (3) category selection. These are outlined below.

3.1. Material selection

To identify and select relevant literature, we searched Scopus, the largest scholarly database [59]. This database has also been used by other researchers within manufacturing to conduct systematic literature reviews (cf [60–62]). To facilitate the search process in Scopus, search terms need to be identified and selected. We adapted the procedure for structuring search terms into different levels [58] (p.4806) to construct a funnelling structure and identify relevant literature. It is common to use a three-level search structure, with one level broadly describing the research field, the second describing the core problem and the third defining the solutions (methods, approaches and so on) used to solve that problem, as shown in Table 1. Each of these levels is described below.

Level 1 defines the field for the literature search. As the focus of this paper is to summarise the state-of-the-art of throughput bottleneck analysis in the manufacturing field, the search terms include the word “manufacturing”. However, “manufacturing” is also referred to in research papers as “production”, hence the term “production” is also included.

Level 2 focuses on the core problem addressed in this paper which is throughput bottlenecks. However, there is a possibility that the term “bottlenecks” may be used implicitly in research papers to mean throughput bottlenecks. Hence to capture this, the term “bottleneck” is used as a selection criterion.

Level 3 identifies the type of approach used to analyse throughput bottlenecks. As the focus of this paper is AI for throughput bottlenecks, the term artificial intelligence is used. However, a myriad of terminologies is used in academic research to describe the term AI. These include data analytics, data science, machine learning, predictive modelling, pattern recognition, learning and so on (as described in Section 2.3). In order not to miss any source in the initial search, a broad range of terms (as shown in Table 1) were used as selection criteria.

The search terms shown in Table 1 were searched for under “article title, abstract, keywords” in the Scopus database to retrieve the relevant studies. We decided to start our search in 2009 when [2] (p. 3843) in their review of analytical approaches to throughput analysis indicated that, with the availability of real-time production system data, a new paradigm for analysing production systems was emerging. The search was conducted on 23rd October 2020.

The initial search yielded a total of 622 documents. To review these documents, a systematic review methodology was adapted from [63] (p.168), as shown in Fig. 2. The initial search results had a range of document types such as journal articles, conference proceedings, book chapters, trade journals and editorials. As it is common for researchers in the manufacturing engineering and AI fields to report their research efforts in the form of either journal articles or conference proceedings, only these two document types were considered in Scopus. Irrelevant subject areas such as biology, earth and planetary sciences, veterinary sciences and so on were then excluded. The subject areas kept for analysis were engineering, computer science, business management and accounting, decision sciences and mathematics. These refinements yielded a final set of 405 documents. The titles and abstracts of these documents were then reviewed. This included reading the titles and abstracts and excluding documents that did not address the problem of throughput bottlenecks in manufacturing. After this filtering process, a set of 57 documents remained. Each of these 57 documents was then read thoroughly to filter out documents that did not focus on using AI tools. For example, analytical approaches and discrete event simulation-based approaches for throughput bottleneck analysis were excluded as they did not rely on the data to analyse throughput bottlenecks but instead used models to understand and experiment with

Table 1
Three-level structure of keywords.

Level	Terms
1	“Manufacturing” OR “production”
	AND
2	“bottleneck”
	AND
3	“Artificial intelligence” OR “data-driven” OR “data mining” OR “machine learning” OR “pattern recognition” OR “statistics” OR “prediction” OR “big data analytics” OR “data science” OR “learning”

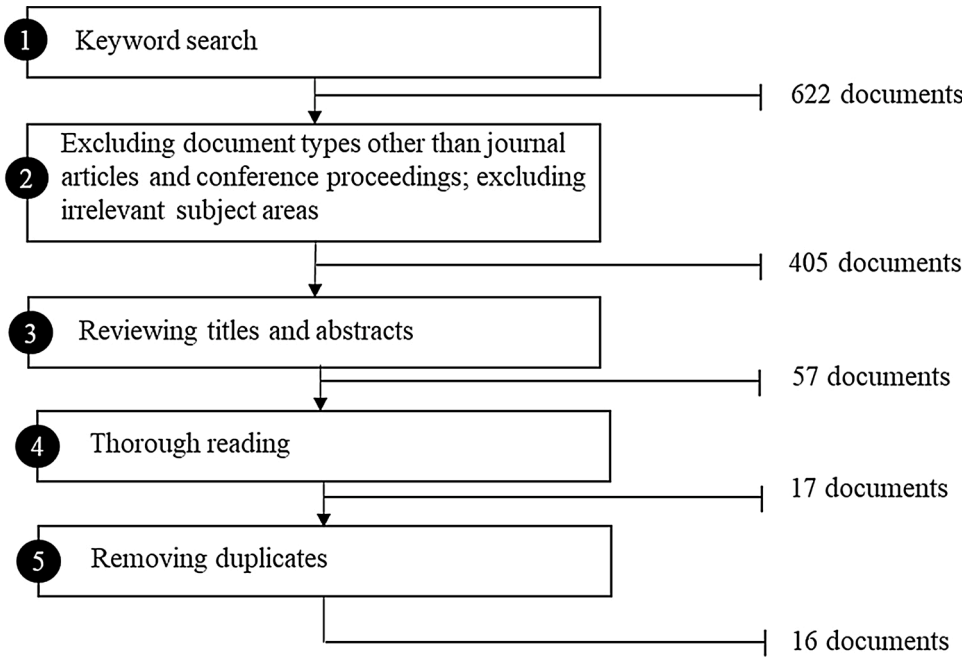

Fig. 2. Systematic review methodology.

throughput bottlenecks in the production system. This process resulted in a set of 17 documents. Finally, the duplicate documents (those documents published as conference papers and later in an extended version as a journal paper keeping the same core idea) were identified and removed. After this process, a final set of 16 documents remained. These 16 were then reviewed, to understand how they helped advance the field beyond previous analytical and simulation-based solutions and to identify future research directions.

3.2. Descriptive analysis

To understand the chronological order of research efforts, their dissemination venue and their authors, it is common to discuss the details of the documents such as their title, year and place of publication and authors' country of affiliation [63] (p.169). This data is provided in Table 2.

From the Venue column in Table 2, it can be seen that 14 documents were published in journals and two in conference proceedings. The most frequent publications used to disseminate research results are leading industrial engineering journals, such as the International Journal of Production Research (three publications), Computers and Industrial Engineering (three publications) and the Journal of Manufacturing Systems (two publications). This highlights the importance of throughput bottlenecks.

From the Country column (specifying the authors' affiliation country) and Year of publication column in Table 2, it may be inferred that research into AI for throughput bottleneck analysis in manufacturing systems was first explored by researchers in the US, as reported in publications [13,33,35]. Overall, the US and Sweden account for two-thirds of the literature, whereas other countries (like Italy, New Zealand, Switzerland and China) jointly account for one-third.

Also, based on the Title of publication column, different terms may be understood to represent AI for throughput bottleneck analysis. The term "data-driven" was used in the titles of eight publications. The term "prediction" was used in four publications and the rest of the publications use the name of the core AI tool used for throughput bottleneck analysis, such as "adaptive network-based fuzzy inference system (ANFIS)" [64] and "hierarchical clustering" [65]. This confirms that different terminologies are used to express the idea of processing the production system data, learning the patterns from that data and using it to analyse throughput bottlenecks.

3.3. Category selection

Various criteria may be adopted to classify and analyse the retrieved documents, for example based on the nature of the data, the AI tools used for throughput bottleneck analysis and so on. To maintain a close link to practice this study adopts the Gartner data analytics framework [66] and categorizes studies according to their goal as: (1) describe, (2) diagnose, (3) predict and (4) prescribe. Many fields adapt the Gartner framework based on its suitability for literature classification purposes (for example, in finance [67] and maintenance [53]). Note that the original term "describe" (meaning describing the past) in the Gartner framework is replaced with "identify" because it better suits the context of this study. The four different categories can be described as follows:

- *Identify* focusses on identifying the historical throughput bottlenecks in a production system. It answers the question, "which machines were throughput bottlenecks in the production system?" AI tools used to identify throughput bottlenecks examine the characteristics of the machines by learning from the historical data in the production system to identify throughput bottlenecks.
- *Diagnose* focusses on identifying possible root causes of historical throughput bottlenecks in a production system. It answers the question, "what were the root causes of historical throughput bottleneck machines?" AI tools are used to explore drivers of historical throughput bottlenecks.
- *Predict* focusses on identifying future throughput bottlenecks in a production system. It answers the question, "what will be the throughput bottlenecks in the production system?" AI tools use historical data sets to learn the patterns, forecast future patterns and use these forecasts to predict throughput bottlenecks.
- *Prescribe* focusses on identifying and recommending actions on future throughput bottlenecks. It answers the question, "what actions need to be taken on future throughput bottlenecks?" Once the future throughput bottlenecks are known, possible actions on them are prescribed to reduce their effect on overall production system throughput.

Table 2
List of publications on AI for throughput bottleneck analysis.

Title of publication	Year of publication	Venue	Reference	Country
Data-driven bottleneck detection of manufacturing systems	2009	International Journal of Production Research	[13]	USA
Bottleneck detection of complex manufacturing systems using a data-driven method	2009	International Journal of Production Research	[33]	USA
Throughput bottleneck prediction of manufacturing systems using time series analysis	2011	Journal of Manufacturing Science and Engineering, Transactions of the ASME	[35]	USA
Bottleneck prediction method based on improved adaptive network-based fuzzy inference system (ANFIS) in semiconductor manufacturing system	2012	Chinese Journal of Chemical Engineering	[64]	China
Real-time data-driven average active period method for bottleneck detection	2016	International Journal of Design and Nature and Eco dynamics	[72]	Sweden
A statistical framework of data-driven bottleneck identification in manufacturing systems	2016	International Journal of Production Research	[71]	Italy, China
An algorithm for data-driven shifting bottleneck detection	2016	Cogent Engineering	[34]	Sweden
A two-layer long short-term memory network for bottleneck prediction in multi-job manufacturing systems	2018	ASME 2018 13th International Manufacturing Science and Engineering Conference, MSEC 2018	[80]	USA
Data-driven detection of moving bottlenecks in multi-variant production lines	2018	IFAC – Papers Online	[70]	Switzerland
Data-driven algorithm for throughput bottleneck analysis of production systems	2018	Production and Manufacturing Research	[37]	Sweden
A data-driven algorithm to predict throughput bottlenecks in a production system based on active periods of the machines	2018	Computers and Industrial Engineering	[79]	Sweden
A proactive task dispatching method based on future bottleneck prediction for the smart factory	2019	International Journal of Computer Integrated Manufacturing	[81]	China, New Zealand
A prognostic algorithm to prescribe improvement measures on throughput bottlenecks	2019	Journal of Manufacturing Systems	[82]	Sweden
A parallel-gated recurrent units (P-GRUs) network for the shifting lateness bottleneck prediction in make-to-order production system	2020	Computers and Industrial Engineering	[36]	China
A generic hierarchical clustering approach for detecting bottlenecks in manufacturing	2020	Journal of Manufacturing Systems	[65]	Sweden, Germany
A data-driven approach to diagnose throughput bottlenecks from a maintenance perspective	2020	Computers and Industrial Engineering	[75]	Sweden, Germany

The above categorization is important for throughput bottleneck management in real-world, shop-floor practice. On the shop floor, the cycle of throughput bottleneck elimination (as explained in Section 2.1) involves identifying, diagnosing, predicting and prescribing steps towards determining elimination actions. For example, identifying and diagnosing historical throughput bottlenecks are important steps in retrospective investigation or analysis of the causes that led to the occurrence of throughput bottlenecks. This will help shop-floor practitioners take appropriate reactive actions to eliminate these bottlenecks. Knowing future bottlenecks will help practitioners plan proactive actions. A detailed explanation of the real-world significance of these categories is provided in the next section (Section 4).

Finally, a three-step categorisation is adopted to present the AI solution for each bottleneck category. These are: (1) input data, (2) modelling approach (including the feature engineering steps) and (3) outputs, as clarified in Section 2.3 above.

4. State-of-the-art research in AI for throughput bottleneck analysis

This section presents the state-of-the-art AI solutions for each category defined in Section 3.3. For each category, an introduction that explains the real-world significance of the category is presented first.

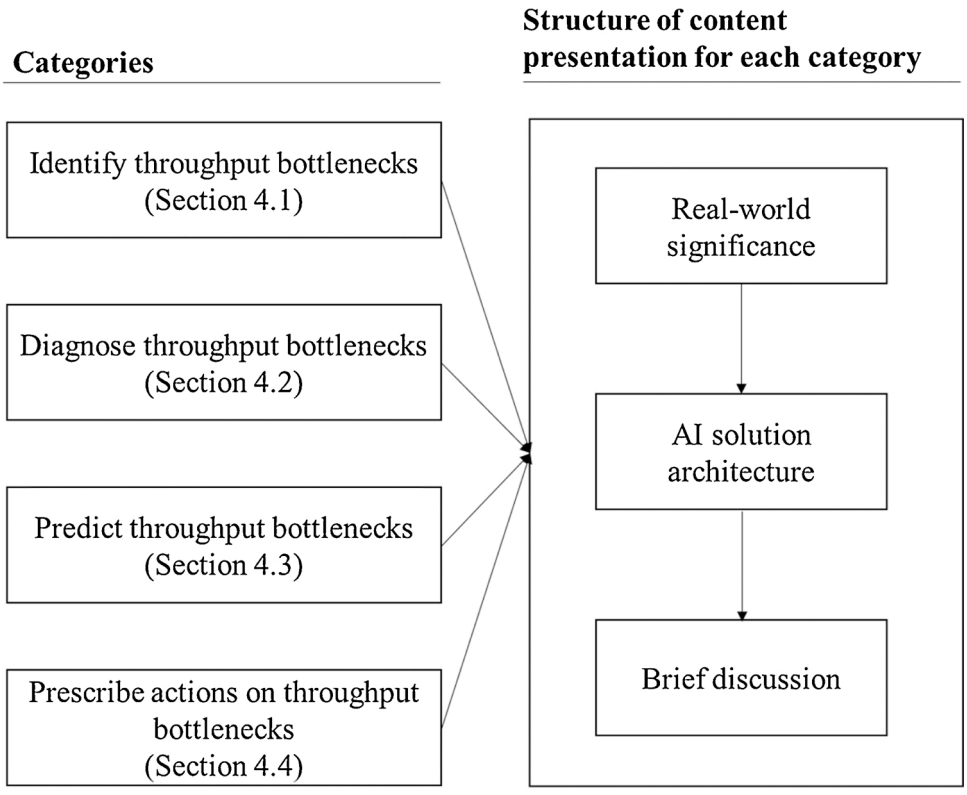

Fig. 3. Structure of information presented for each category.

The documents presenting an AI solution are then identified from Table 2 and the AI solution's architecture is summarised in tabular format with information on input data, modelling approach and output. It is important to note that different categories may have different tabular formats, especially concerning modelling approaches. This is because the steps within a modelling approach are unique across categories. Finally, a brief discussion on the AI solution's architecture and the challenges associated with using that solution in real-world practice is presented. To facilitate the reading process for each category, a reading guide for the readers of the paper appears in Fig. 3.

4.1. Identifying historical throughput bottlenecks

In practice, throughput bottlenecks can be analysed in two different time frames: long-term and short-term (as explained by the time frame challenge in Section 2.2). [2] (p.3838) indicates that short-term throughput bottlenecks are responsible for impeding short-term throughput and that eliminating it should ultimately help remove long-term bottlenecks. However, the definitions of short-term and long-term periods remain still unclear in the literature. [2] (p.3838) pointed out that "long-term" and "short-term" throughput bottlenecks are not clearly defined and emphasised the need for a proper definition. [68] (p.196) argues that "long-term" and "short-term" are relative definitions in the context of throughput bottleneck management and states that it is difficult to find a precise definition.

Reflecting on the existing arguments about "long-term", "short-term" and "real-world practice" in throughput bottlenecks, the authors propose the following definitions. "Long-term" is considered to be a time with a specific number of production runs (one production run is equivalent to one production cycle, shift or day). Consequently, machines that act as throughput bottlenecks over a certain number of defined production runs are called "long-term bottlenecks". "Short-term" is considered to be a specific time instant within a production run. Machines acting as throughput bottlenecks within a production run at different instants of interest are called "short-term bottlenecks".

4.1.1. Identify long-term historical throughput bottlenecks

In a real-world shop floor scenario, if practitioners want to significantly increase the current throughput, (by, say, 40 %–50 % [69]), then long-term throughput bottlenecks should be identified, plus the necessary actions to eliminate them. These actions are time-consuming and demand significant capital investment. Some examples include cycle time reduction involving changing or redesigning machine components (increasing the ram speed in a pressing machine for example), semi-automation (automated unloading), smart fixtures, changing the layout to minimise wasted motion (such as picking and stacking parts or placing buffers), optimising shop-floor maintenance practices to improve throughput, upgrading the throughput bottleneck machine or even deciding to buy an additional machine to eliminate throughput bottlenecks. AI solutions are needed which can analyse historical data and inform practitioners where long-term throughput bottlenecks were in a production system.

Six AI solutions are presented in the literature which support practitioners in identifying long-term throughput bottlenecks. The architecture of AI solutions is presented in Fig. 4. The technical aspects of the AI solutions are summarised in Table 3. The output of all these solutions is a set of long-term throughput bottlenecks in the production system.

Every solution presented in Table 3 treats the task of identifying long-term throughput bottlenecks as a classification problem. In other words, each machine in a production system needs to be classified as either a long-term throughput bottleneck or a non-bottleneck. Different researchers have used different input data sets in this process and constructed different modelling approaches.

These modelling approaches may be broadly divided into two modules: extracting features and classification. Most AI solutions use statistical tools to extract statistical features such as averages and confidence intervals from the input data. One exception is [65] (p.153), which uses unsupervised machine learning-based clustering tools to group machines with similar and dissimilar behaviour based on their time-series profiles. These groups are then used as features to identify long-term throughput bottlenecks in the production system. Once the

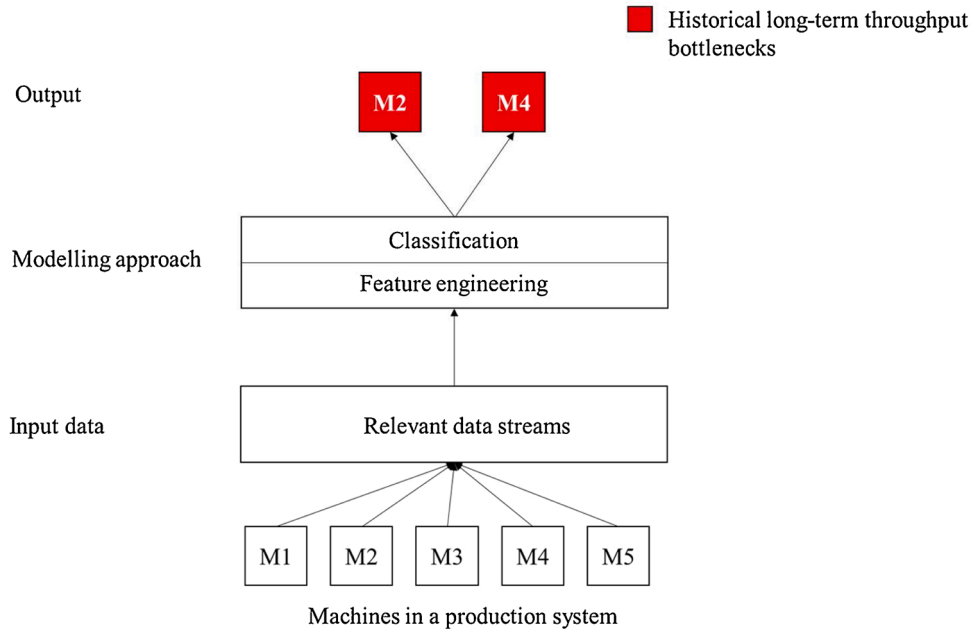

Fig. 4. Illustration of architecture of AI solutions for identifying historical long-term throughput bottlenecks.

Table 3
Architecture of AI solutions for identifying long-term historical throughput bottlenecks.

Reference	Input data	Modelling approach		Output
		Feature engineering	Classification	
[13,33]	Blockage and starvation durations	Average total duration of blockage and starvation times	Rule-based classification	Set of long-term throughput bottlenecks
[72]	Active durations	Average, confidence intervals of active durations	Rule-based classification	
[71]	Blockage and starvation durations (or) active durations (or) inter-departure durations	Average of selected input data generated by batch means technique	Rule-based classification	
[37]	Active durations	Average, confidence intervals of the active duration	Rule-based classification	
[70]	Set of machines' cycle times for every product and Takt time for every product	Cumulative probability distribution function	Rule-based classification	
[65]	Active duration time series	Groups of similar and dissimilar machines, representative time series for each group	Rule-based classification	

features have been extracted, classification techniques are employed to classify the machines as throughput bottlenecks or non-throughput bottlenecks. Predefined rules are used for this classification. [13] (p.5024) and [33] (p.6934) use a rule which compares each machine's blockage and starvation durations to identify the turning machines (the bottlenecks) for which blockage and starvation patterns change. Similarly, [70] (p.162) uses a rule which compares cycle times and takt times to classify throughput bottlenecks, [65] (p.153) classifies a group of machines as a throughput bottleneck if that group has the highest active durations. [71] (p.6321), and [72] (p.433) and [37] (p.234) use rule-based hypothesis testing techniques to identify throughput bottlenecks.

Challenges to implementing the solutions shown in Table 3 include the need for sufficient historical data that needs to cover enough past production cycles. Moreover, as can be seen from Table 3, the existing research is limited to using machines' activities-based input data such as active times, blockage and starvation times and so on. This input data needs to be interpreted cautiously. Finally, there may also be other contextual factors such as worker availability, supply logistics to different machines and product mix. These may contribute to machines acting as throughput bottlenecks but not be factored in by existing AI solutions.

4.1.2. *Identify short-term historical throughput bottlenecks*

Practitioners need to prioritise short-term throughput bottlenecks to reduce fluctuations and achieve the target production run throughput (the shift throughput). For example, [73] (p.213) indicated that the throughput of a semiconductor production line varies between 475 and 800 wafers per day. These fluctuations occur due to the existence of short-term throughput bottlenecks. These bottlenecks need to be monitored continuously and require immediate attention if they should undergo any undesirable random processing events (including real-time monitoring of production lines from back offices). For example, if there is a breakdown event at a short-term throughput bottleneck then maintenance practitioners need to prioritise this machine. Information on short-term throughput bottlenecks is also necessary to dynamically rebalance a production line, including shifting workers to throughput bottleneck machines from upstream or downstream machines, running the throughput bottlenecks during shift breaks and dynamically changing throughput bottleneck-orientated releases [74]. This requires effective AI solutions to help practitioners quickly identify short-term throughput bottlenecks.

The overall architecture of the AI solution is similar to that used in

identifying long-term throughput bottlenecks illustrated in Fig. 4 above. However, the technical aspects are different. The AI solution and its technical aspects (developed to identify short-term throughput bottlenecks) are presented in Table 4.

[34] developed a solution for identifying short-term throughput bottlenecks. This solution identifies throughput bottlenecks continuously during a production run. It treats the problem as a binary classification problem in which machines are classified as bottleneck or non-bottleneck. For this the total active duration for the active state of every machine at every instant is extracted. Rules taken from the active period method (as described in [11] (p.1081)) are then used to classify the machines as either a throughput bottleneck or non-throughput bottleneck.

There are two main challenges in using the solution presented in [34]. Firstly, input data should be available and updated in real-time and without delay. Secondly, the production system needs to have reached a steady state before this solution can be used. If more than one machine has the same longest uninterrupted active durations at a time instant, the AI solution identifies all those machines as short-term throughput bottlenecks. Hence, in real-world practice, once a production run starts, short-term throughput bottlenecks cannot immediately be identified. This is because all machines might be active when a production run starts due to the left-over jobs in the machines from the previous production run.

4.2. Diagnose historical throughput bottlenecks

Identifying historical throughput bottlenecks is the first important step. The next step is to diagnose their root causes and plan the right elimination actions, as described in Section 2.2 (challenges of multiple root causes). Root causes of throughput bottlenecks may be determined by investigating and analysing different sources of process time variability (see Section 2.1). However, there are different categories of process time variabilities, such as random processing times, unplanned stops and so on [37]. Each category may have numerous subcategories which exponentially increase the complexity of identifying the right root causes and planning the right actions to eliminate them. For example, in real-world production systems, there might be several categories of unplanned stops such as tool error, component breakage, fixture setting errors and stops due to a reduction in oil pressure. Moreover, these categories might be related to product types. [75] (p.9) indicated that there were 615 unique unplanned stops on a throughput bottleneck machine in a real-world production line. In such scenarios, it is challenging to examine the details of each process time variability and plan the right actions manually. AI solutions are required to diagnose the various root causes of historical throughput bottlenecks. This may then help practitioners to take the right actions.

Table 5 shows the AI solutions and their technical details focusing on diagnosing throughput bottlenecks. Meanwhile, the architecture of the AI solution is presented in Fig. 5. [75] (p.4) proposes a solution that may diagnose the various unplanned maintenance stops on throughput bottlenecks. In real-world production systems, unplanned stops are highlighted as of the major process time variabilities causing occurrences of throughput bottlenecks by [76](p.1), [77](p.5831) and [75] (p.7).

Table 4
AI solution architecture for identifying short-term historical throughput bottlenecks.

Reference	Input data	Modelling approach		Output
		Feature engineering	Classification	
[34]	Sampled binary-coded active states (sampling rate once per second)	Total duration of uninterrupted active state at every instant	Rule-based classification	Short-term throughput bottlenecks

Table 5
AI solution architecture for diagnosing throughput bottlenecks.

Reference	Input data	Modelling approach	Output
[75]	Total duration, cumulative frequency, co-efficient of variation, mean stop time, product types for each type of unplanned stop	K-means clustering	Visual plots representing each cluster information

[75] (p.5) considers the diagnosis of unplanned stops to be a clustering problem and designs a solution to group unplanned stops into distinct clusters (based on the input data) using unsupervised machine learning-based k-means clustering techniques. Thereafter [75](p.10) uses visual plots to represent the clustering results. Practitioners (especially maintenance practitioners) need to interpret these plots manually and prioritise the various unplanned stops for action. It is also important to note that the distinct group of unplanned stops created by the application of AI solutions is based entirely on numerical calculations. It is therefore imperative to interpret the cluster results whilst maintaining relevance to real-world practice.

There are two main challenges in using the above solution to diagnose throughput bottlenecks. Firstly, the number of unique unplanned stops should be high enough for the AI solution to work (at least the number of intended clusters). Secondly, existing AI solutions may only be used in environments for which unplanned stops are the major root cause of throughput bottlenecks. There may be many different root causes, such as random processing times, setup times and so on (see Section 2.1). In these situations, the existing AI solution offers limited benefits to practitioners.

4.3. Predict throughput bottlenecks

In real-world practice, practitioners often have regular shop-floor meetings before a production run starts, for example convened as morning meetings, pulse meetings or continuous improvement meetings [78] (p.85). In these meetings, it is common to discuss the performance of the previous production run. This is done by comparing the actual throughput to the target throughput for the previous production run. The meeting will analyse the gaps, identify which machines were acting as throughput bottlenecks and then plan appropriate actions during the upcoming shift. However, due to the stochastic nature of production systems, there is no guarantee that the historical throughput bottlenecks will continue to act as throughput bottlenecks (the shiftiness challenge in Section 2.2). Practitioners may make better-informed decisions and plan proactive action if they know the upcoming throughput bottlenecks in the system before the production run starts. Thus, AI solutions need to support practitioners in making informed decisions on the shop floor.

Six AI solutions were found in the literature that focus on predicting throughput bottlenecks. The technical details of these AI solutions are summarized in Table 6, whilst the architecture of the AI solution is illustrated in Fig. 6.

The solutions presented in Table 6 treat the throughput bottleneck prediction problems as classification problems. In other words, each machine in the production systems needs to be classified as a probable throughput bottleneck or non-throughput bottleneck for the future production run. To accomplish this, different researchers have developed different AI solutions using different input data sets.

The solutions presented in Table 6 may be divided into two categories based on the input data: (1) solutions that use only machine data and (2) solutions that use machine data in conjunction with other contextual data. [35] (p.4) and [79] (p.538) use only machine data

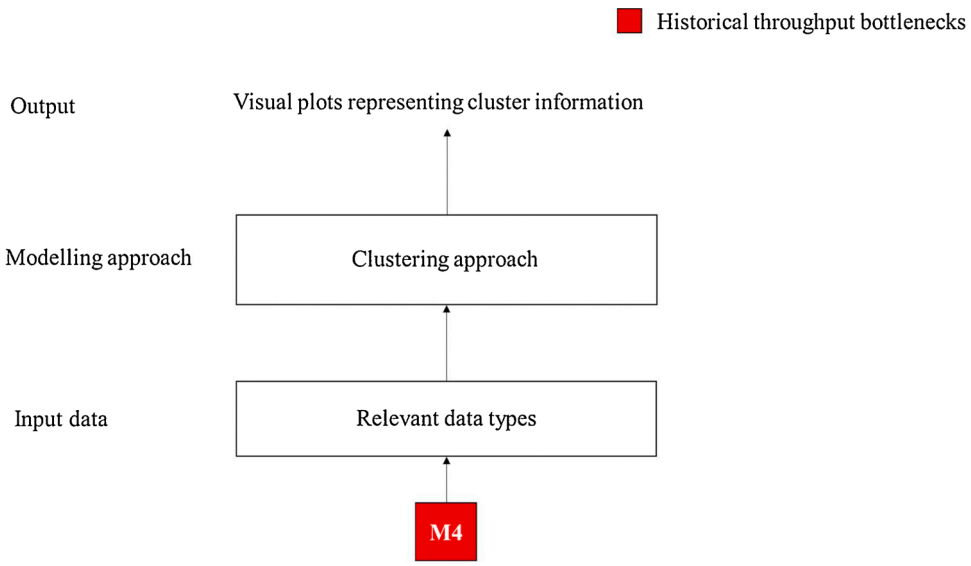

Fig. 5. Illustration of architecture of AI solution for diagnosing historical throughput bottlenecks.

(such as blockage and starvation times and active times) to predict future throughput bottlenecks. However, using only a machine's data may be ineffective when it comes to making accurate predictions in a stochastic production system. This is because other contextual factors (such as product mix and buffers) also affect the locations of throughput bottlenecks in a production system. Therefore, [64] (p.1084), [80] (p.5), [81] (p.282) and [36](p.4) used machine data in conjunction with other contextual data (such as buffer amount and product types) to predict throughput bottlenecks.

The modelling approach consists of two modules: forecasting and classification. In the forecasting step, the aim is to learn historical patterns from the input data, use this information and then forecast future patterns. Various AI tools are used for this purpose. These tools may be broadly classified into two categories: statistical and ML tools. Their purpose is to learn what has happened in the past while uncovering (a) unseen patterns in the input data and (b) interactions between the machines and their relationships and then use this information to forecast future values. [35] (p.2) and [79] (p.538) use statistical time series forecasting tools called auto-regressive integrated moving average (ARIMA). These tools provide a good means to learn the linear relationships in the input data but they fail to effectively learn non-linear relationships [36] (p.2). Non-linear relationships do exist in a production system, as several machines are interacting with each other, and learning this information can increase accuracy. Therefore [64,80,81, 36], use ML tools to capture these non-linear relationships. Specifically, they employ techniques such as ANFIS, LSTM, DNN and P-GRU, as can be observed from Table 6.

In the classification step (as with the identification of historical throughput bottlenecks), rule-based methods are used to classify the machines as throughput bottlenecks or non-throughput bottlenecks. [35] (p.1) and [80] (p.6) use a rule which compares each machine's forecast blockage and starvation durations and finds the turning machines (the predicted bottlenecks) based on changes in blockage and starvation patterns. Meanwhile [79] (p.538) uses the rule that the machine with the highest forecast active duration is the throughput bottleneck, [64] (p.1082) and [81] (p.283) uses the rule that the machine with the highest production load is classified as a throughput bottleneck, and [36] (p.3) uses the rule that machines with the highest relative lateness are throughput bottlenecks.

There are two main challenges in using AI solutions to predict throughput bottlenecks. Firstly, there needs to be sufficient historical data. In the research efforts to date (cf. Table 6), the amount of historical production system data that can be used to train AI is set (for example, [35] (p.4) uses data from 85 previous shifts and [80] (p.6) uses six months' worth of historical data). However, [79] (p.542) argued that not all historical data is useful in predicting throughput bottlenecks. This is because major improvements might be made in the production flow (such as installing a new parallel machine) and the data from before such improvements will not represent current production system dynamics and thus lead to inaccurate prediction of throughput bottlenecks. Secondly, it is a challenge to generalise a forecasting methodology across all production systems (for example, it cannot be said with certainty that LSTM as proposed by [80](p.6) can be used to predict throughput bottlenecks for all production systems). The selection of a particular forecasting methodology depends on the production system dynamics and available type of data. Sometimes, getting better performance means running multiple forecasting methodologies in an ensemble.

4.4. *Prescribe actions on throughput bottlenecks*

Predicting throughput bottlenecks informs practitioners which machines in a production system are likely to behave as throughput bottlenecks. However, practitioners may be interested in knowing the answer to a more pragmatic question: what concrete actions they must be prepared to take to mitigate upcoming throughput bottlenecks? This question is not directly answered by predictive insights. Practitioners must manually assess different possible actions and determine which actions are to be implemented. Practitioners face two main challenges in such a process. Firstly, the predicted type of throughput bottlenecks (such as cycle time bottlenecks or downtime bottlenecks) is not known when planning specific actions (the consensus challenge in Section 2.2). Secondly, in a real-world production system setting, there are too many variables, constraints and system-level trade-offs which need to be considered when deciding which actions are best for eliminating future throughput bottlenecks. AI solutions may help to support practitioners in better addressing these challenges and prescribe the right set of actions on throughput bottlenecks.

[82] proposed a partial solution aimed at prescribing actions on predicted throughput bottlenecks. The technical aspects of the solution are summarized in Table 7, whilst its architecture is illustrated in Fig. 7.

[82] (p.274) developed a two-stage solution. The first stage involves forecasting the various machine states' duration of the predicted throughput bottlenecks. This forecasting is based on the input machine states data (expressed as time series) and by using suitable time series forecasting techniques. The forecast values are then used to predict the

Table 6
AI solutions architecture for predicting throughput bottlenecks.

Reference	Input data	Modelling approach		Outputs
		Forecasting methodology	Classification	
[35]	Blockage duration time series and starvation duration time series	Auto-regressive moving average (ARMA) for forecasting blockage and starvation times	Rule-based classification	A set of predicted throughput bottlenecks
[64]	Processing times, utilisation rate, buffer length, mean time between failure (MTBF), mean time to repair (MTTR), work in progress, product types and releasing strategies	Adaptive neuro-fuzzy inference systems (ANFIS) for forecasting production load	Rule-based classification	
[79]	Active duration time series	Suitable time series forecasting technique	Rule-based classification	
[80]	Product mix, operator shift, cycle time, blockage duration time series and starvation duration time series	Long short-term memory (LSTM) for forecasting blockage and starvation	Rule-based classification	
[81]	Processing times, utilisation rate, buffer length, mean time between failure (MTBF), mean time to repair (MTTR), work in progress, product types and releasing strategies	Deep neural networks (DNN) for forecasting production load	Rule-based classification	
[36]	Work in progress, waiting time, utilisation rate, failure time and starvation time	Parallel gated recurrent units (P-GRU) for forecasting lateness index	Rule-based classification	

type of throughput bottlenecks, which is used as input for the second stage. In the second stage, prescriptive rules are used to provide a list of recommended actions. Although this solution does not prescribe an optimal set of actions, it is a step towards prescribing actions.

Similar to other AI solutions, a first main challenge in implementing this solution is that detailed work records on throughput bottlenecks should be available in digital format. A further challenge is that, for the prescriptive solutions to work effectively, these detailed work records should be complete, i.e. there should be no missing or partial information.

4.5. Summary

In the previous sections, the different AI solutions for throughput bottleneck analysis were classified into four types: (1) identity, (2) diagnose, (3) predict and (4) prescribe. Based on the number of publications across these four categories, identify and predict have received the maximum research attention.

In the identify category, the various AI solutions use only machine data such as active periods, blockage, and starvation times and so on, to identify historical throughput bottlenecks. In the predict category, some AI solutions use only machine data, while the rest uses machine data in conjunction with other contextual data. Moreover, in these two categories, the problem was formulated as a classification problem, classifying the machines as throughput bottlenecks or non-throughput bottlenecks. A variety of pre-defined rules were used to facilitate this process. For the diagnosis category, existing research work was limited to diagnosing unplanned stops based on the different process time variabilities (see Section 2.1). This problem was formulated as a clustering problem, aiming to expose the underlying patterns in the occurrence of unplanned stops. For the prescribe category, the problem was formulated as a rule-based problem, in which pre-defined prescriptive rules are used to prescribe actions on throughput bottlenecks. All the existing research efforts in all categories were based on sample data in which a batch of data was used to construct, test and verify the performance of the AI solutions. It should be noted that none of the AI solutions was reported to be actually implemented in the real-world.

The following sections present and discuss promising future research directions and practical recommendations for throughput bottleneck analysis. These directions and recommendations are based on the theory of throughput bottlenecks (Section 2.1), AI literature, a careful review of existing AI solutions (Sections 4.1, 4.2, 4.3 and 4.4), the authors' practical experience of working with throughput bottlenecks in practice for several years, as well as reflections on the gaps between theory and practice.

5. Future research directions

The existing AI solutions may be further enriched to provide deeper analysis of throughput bottlenecks and give a list of actions for eliminating throughput bottlenecks. Below, the authors list a couple of key research directions which will help advance the field of AI for throughput bottlenecks.

5.1. Factory floor data fusion

Most of the existing AI solutions use machine-level data to analyse throughput bottlenecks. Machine-level data characterises the machine activities (such as MTTR, MTBF, active duration, inactive durations and so on) [13,37]. Although this is valuable information, it does not fully explain throughput bottlenecks in a way that might allow practitioners to take concrete actions. This may be further improved by integrating machine-level data with other data sources. Data fusion from multiple sources will help to further reduce ambiguity when identifying throughput bottleneck machines and actions. Moreover, data fusion can also be seen as a step towards the vision of achieving self-aware production systems [83]. In other words, it will provide enough ability to capture, characterise and forecast anticipated production system dynamics, as well as to predict throughput bottlenecks and take elimination actions automatically, without human intervention or action. This integration needs to be of two types: (1) machine-component and (2) contextual production data.

(1) Machine-component data fusion. This refers to integrating machine-level data with component-level data. Machines have many different components and manufacturing companies are increasingly installing sensors to monitor them. Integrating sensor and machine data will help with deep diagnosis of throughput bottlenecks, understanding

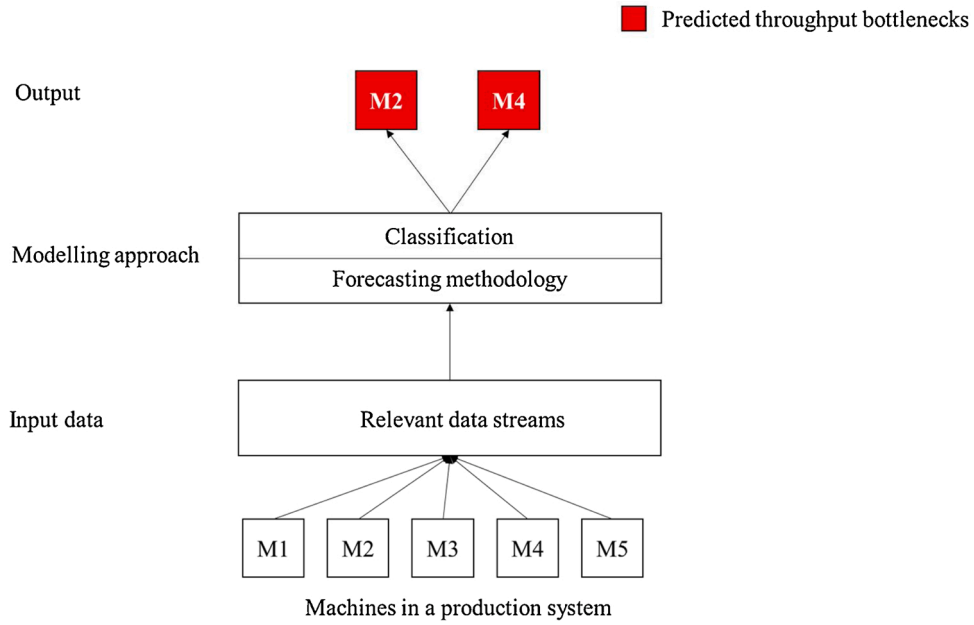

Fig. 6. Illustration of architecture of AI solutions for predicting throughput bottlenecks.

Table 7
AI solution architecture for prescribing actions on throughput bottlenecks.

Reference	Stage	Input data	Modelling approach	Output
[82]	1	Individual bottleneck machine states (such as producing, down, etc.) time-series data of predicted throughput bottlenecks	Forecasting methodology: suitable time series forecasting technique to forecast values of each machine state	Forecast duration of each machine state
	2	Forecast duration of each machine state, historical actions list	Prescriptive rules	List of recommended elimination actions

machine health and gaining knowledge of exact root causes. This type of integration may also help with prescription as it opens the way to prescribing concrete actions for eliminating throughput bottlenecks. For example, [37] (p.239) points to the downtime type of throughput bottlenecks in a production system, based on machine-level data. However, if machine-level data is fused with component-level data, the root causes of why a downtime throughput bottleneck occurred may be deduced. However, fusing these two different types of data poses an open challenge as they reveal information with two different granularities. Extensive future research in this direction will be needed, to develop AI solutions that provide even more detailed diagnostics of throughput bottlenecks.

(2) Contextual production data. Fusion with other contextual production system data may drive improvements in the accuracy of throughput bottleneck analysis and lead to a better understanding of bottlenecks in a given context. For example, a combination of machine data, buffer data, product types and releasing strategies as shown in [64,80,81,36]. Also other types of data may be combined, such as logistics data, product quality data (incorporating yield aspects into bottleneck analysis), maintenance work order data and production planning data. AI solutions are needed which explore how such data might be combined systematically and used in the effective realisation of prescriptive throughput bottleneck management.

5.2. Ensuring data quality

Although it has become possible to collect different types of data, ensuring data quality is a must if AI is to deliver meaningful value to practitioners. The importance of data quality has been widely discussed for many years in the AI literature [84] and manufacturing literature (e.g [2,54].). However, it has received less attention in the context of throughput bottleneck analysis. Future research is therefore needed to explore ways of ensuring the right data quality.

Addressing data quality has several challenges. Firstly, on the shop floor, data collection technologies may be unreliable, or their performance may deteriorate with time, leading to the recording of incomplete information [2] (p.3839). This needs to be compensated for systematically to increase the accuracy of throughput bottleneck analysis. For example, all data from the factory machines should be collected and be available to allow production flow to be traced and AI to identify throughput bottlenecks. Having incomplete information from some machines may lead AI to identify throughput bottlenecks incorrectly, which will certainly impact factory performance. Secondly, actions aimed at eliminating throughput bottlenecks (such as reducing downtime) may be traced to machine-level actions (as argued in Section 4.2) and eventually down to component-level actions. In such scenarios, any changes to the component level will impact production system throughput. Therefore, ensuring data quality at all levels (system, machine and component) is the key to successfully eliminating throughput bottlenecks in a production system. Thirdly, issues such as noisy data and the handling of outliers and inliers need to be addressed systematically to avoid the introduction of biased AI solutions. Lastly, production processes also change over time. This has various causes, such as the ageing of machines, implementation of lean practices and general production management practices in factories. In such scenarios, changes occur in the underlying pattern of input data and consequently, AI may also start to drift over time and suffer accuracy losses. To address this, AI needs to be quickly adaptable to such changes, if it is to reliably analyse throughput bottlenecks. This aspect needs to be given special attention in throughput bottleneck analysis. Initial guidance towards achieving this is provided in [79] (p.537).

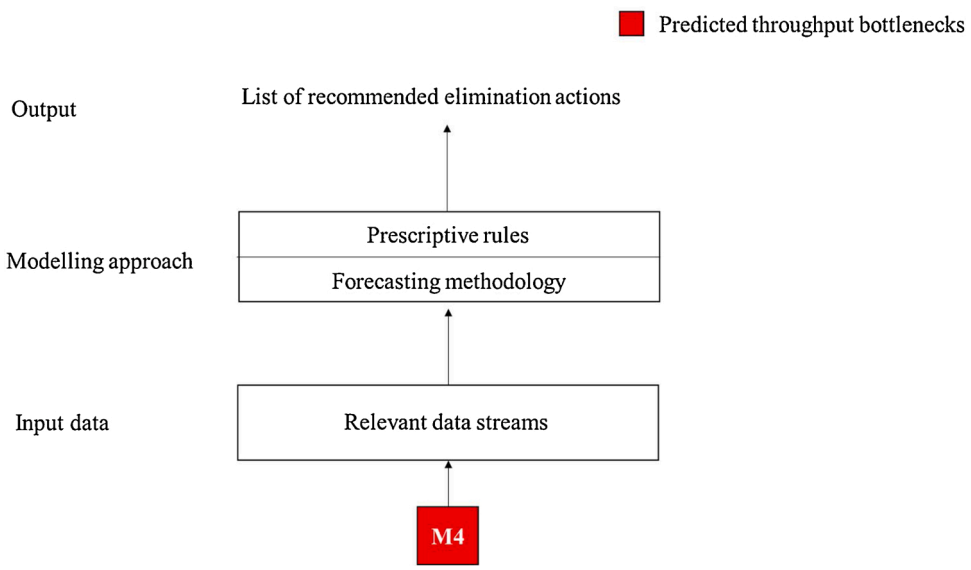

Fig. 7. Illustration of architecture of AI solution for prescribing actions on predicted throughput bottlenecks.

5.3. Benchmarking AI performance with practice

Practitioners commonly want the performance of AI solutions to be 100 %. As a result, academic researchers also focus on trying to develop highly accurate AI solutions. The literature concentrates on comparing the performances of various competing AI solutions focussing entirely on accuracy. For example, [36] (p.9) compared five different AI solutions to predict throughput bottlenecks in production systems, with the argument that the best AI solution was the one that gave the smallest error. Although this is a valuable comparison from an academic research perspective, it must be acknowledged that the performance of AI needs to be benchmarked against current practice. [79] (p.542) indicated that it is common for practitioners to have a naïve approach to their work, assuming for example that bottlenecks in a previous production run will remain as such in the next one. Although this is a simple method, the performance of any sophisticated AI should surpass the naïve approach and show its benefit in real-world practice. Further research efforts are required to explore relevant benchmarks and prove the value of AI in improving practice.

5.4. Scaling from line to factory to supply chain

Most of the existing AI solutions summarized in Table 2 were demonstrated by analysing the throughput bottlenecks in a specific production line, such as a door assembly line [33], or machining line [65]. These solutions have proved their value in helping optimise line-specific throughput bottleneck management and increased individual line throughput in factories with multiple production lines. However, increasing throughput from individual production lines may result in excess inventory if the downstream production lines are working at a slower pace. Hence, to gain maximum benefits, there is a need to scale AI solutions to the entire factory. Such scaling requires access to wider data sets from every machine in a factory. These data sets must then be analysed together to identify throughput bottleneck locations in the factory. All this poses challenges given the increased complexity due to machines, processes and systems. For example, in an automotive plant, how can the data from thousands of machines performing different types of operations (such as blanking, machining, assembly, welding, paint shops, heat treatment and washing operations), all with different failure modes, processing times, and so on, be fused without losing information? Such scaling would significantly change a factory's throughput bottleneck management practices. Instead of localised management of individual production lines, a factory perspective would be followed, focusing on improving the factory's throughput. Further development in this direction will certainly benefit companies by ensuring a smooth, swift flow of material to end products.

Similarly, [2] (p.3843) argues that to gain maximum benefit and avoid the accumulation of excess inventory, throughput bottleneck information must be combined with dynamics from outside the factory, such as material supplies and shipping-due dates from customers. AI can be used to combine information from customers (such as demand information) and suppliers (such as ordering raw materials) with throughput bottleneck information from the factory floor, thus optimising the entire flow of products across the supply chain. This will help practitioners to prioritise actions on throughput bottlenecks according to the desired customer shipment level (such as eliminating unnecessary demand from reworks and adding capacity). Further development in this direction will help unleash factories' true productivity potential.

5.5. Humans-in-the-loop (HITL)

HITL is a paradigm within AI that involves incorporating human feedback to improve the AI's performance. For example, it was shown that HITL can link practitioners' intelligence and AI to create a collective superintelligence for medical radiographic diagnosis [85] (p.1). However, HITL has not received much attention in the throughput bottleneck analysis literature. Thus, future research into integrating practitioner feedback may help improve the accuracy of AI for throughput bottleneck analysis.

HITL may be used in different ways for throughput bottleneck analysis. For example, within existing AI solutions, a time window of historical data for identifying long-term historical throughput bottlenecks is a parameter that needs to be set by practitioners [65] (p.146). They do so by using their deep domain knowledge of the production system; a practice that may be combined with AI. AI may select an automatic optimal window for detecting long-term throughput bottlenecks. Such a solution should include aspects relating to the changing dynamics of production systems brought about by continuous improvement efforts. This output may be verified by factory practitioners (and AI-learning based on the practitioner's interpretation) to provide better estimates in the future. The impact of having such a solution is that it will reduce the risk of having too much historical data which is unrepresentative of recent production system dynamics. Further research is required to design an appropriate AI solution that can

automatically detect the window period.

Similarly, the performance of existing AI solutions in predicting throughput bottlenecks is evaluated based on a solution's average performance. For example, [80] (p.7) reports an average error rate of 2.468, while [79] (p.542) reports an accuracy of 86.13 % in predicting throughput bottlenecks. This means that existing AI solutions have been shown to provide results, sometimes correctly and sometimes not. In other words, AI sometimes has a high level of confidence in predicting throughput bottlenecks but that at other times, this confidence is low. These lower confidence outputs may be verified and augmented by practitioners using their deep knowledge of factory dynamics. The practical significance of this is that AI may provide rapid insights into high-confidence throughput bottlenecks, with practitioners able to act immediately without further analysing the outputs, thus saving time. For less confident outputs, practitioners may further evaluate them and make informed decisions. In such scenarios, active learning might be provided to an AI, through practitioner feedback in the form of training data not originally provided as part of a confident diagnosis. To realize this further research is needed to determine the right level of confidence in the AI's predictions and the exact procedure for operationalising it in practice.

5.6. Explainable AI

Many of the AI solutions, especially those using deep learning, are often called "black boxes" as they do not explain their predictions in a way that is comprehensible to humans [86]. This problem has prompted much discussion and research in the AI field about explainable AI (commonly referred to as XAI) [87]. In XAI, a model is created on top of the existing AI solution to explain its mechanisms (such as how much weight AI solutions gave to the features derived from input data). This is accomplished by the use of techniques such as partial dependence plots and SHAP (SHapley Additive exPlanations). XAI has started to receive attention within manufacturing and researchers have started developing XAI models [88].

Exploring the possible uses of XAI is also a potential future research direction within throughput bottleneck analysis.

Firstly, there is the issue of identifying which types of throughput bottleneck analysis require XAI. The authors would argue that not all types require it. For example, having an XAI solution is more beneficial when identifying long-term throughput bottlenecks because elimination actions on these require significant time and money. Instead of merely identifying throughput bottlenecks, a more human-type explanation of throughput bottlenecks may be offered, referring to contextual factors such as buffers, the dynamics of upstream and downstream machines relative to throughput bottlenecks, and specific conditions related to production lines. Practitioners may then take time to interpret these explanations, augment them with their domain knowledge and take confident action. Similarly, when predicting throughput bottlenecks, for example for the next eight-hour shift, instead of merely giving information on a probable set of throughput bottlenecks, XAI can provide explanations on such aspects as how much a particular contextual factor contributed towards making a machine act like a throughput bottleneck. But when identifying throughput bottlenecks in real-time, then having less XAI might be useful (for example, highlighting the throughput bottleneck and its status rather than explaining contextual information), as practitioners might not have sufficient time to interpret the results. Hence, more rigorous research activity is needed to identify concrete use cases within throughput bottleneck analysis (in which XAI may be useful) and designing appropriate XAI solutions.

Secondly, in the AI literature XAI is described as creating a model for interpreting final AI insights [87]. In the real world, these models have an impact whenever an AI solution is used in a decision situation. Models which explain the insights obtained from AI may increase trust in that AI solution. On the other hand, there may be other black boxes in the process of developing AI solutions (from the practitioner's perspective) that need to be explained. For example, explanations need to be given on the data types that were used, automatic AI feature generation process, AI data pre-processing procedures (including treatment procedures for outliers and inliers, assumptions made by AI and so on). Practitioners need full transparency on how the AI processed the data to better understand the insights it provides. Future research is needed to create such transparency since it will increase practitioners' trust in AI, thereby facilitating institutionalisation of the AI in practice.

5.7. Closing the gap between prototype AI and its implementation in practice

The existing AI solutions for throughput bottleneck analysis (developed in academic research) may be considered a prototype. Existing studies typically start with a batch of data from a real-world production system, process it, develop and employ different AI solutions, compare their performances, select the best-performing solution and then further optimise it to improve its performance. The AI solution is then assumed to be working perfectly when implemented in real-world practice. However, there are two problems with such an approach: (1) AI drifting, and (2) usefulness.

(1) AI drifting: developing the AI solution based on limited data and further optimising it may cause the AI solution to overfit. The result of overfitting is that, while AI solutions may look better for a given batch of data, their accuracy risks deteriorating when exposed to new data sets during practical implementation. This may create a gap between the prototype AI solution's accuracy and the accuracy obtained after implementation.

(2) Usefulness: existing research efforts try to make AI solutions achieve better accuracy by processing the data [80,36]. However, when the AI is implemented in the real world, how can practitioners be sure, for example, that the throughput bottlenecks identified by the AI solution are the true throughput bottlenecks in the real-world?

Both questions can be answered if the AI solutions are implemented in the real-world and the effects studied rigorously. However, no current study reports a real-world implementation and its effects. This indicates an implied tendency for researchers to stop once an adequate AI prototype solution has been demonstrated. Researchers then assume that their AI solutions work in real-world practice.

Implementation and studying the effects of AI are therefore considered the most important future research direction for addressing the problems of drifting and usefulness. For example, AI drifting problems may be studied by verifying whether the accuracy obtained in the prototype stage matches that after implementation. This may lead to the development of solutions that make AI solutions more resilient. Similarly, the usefulness problem may only be studied by taking elimination action on throughput bottlenecks highlighted by the AI solution, and observing whether the actual throughput increases in the real-world (see Section 2.1). This may help refine the existing solutions, making them more useful in real-world applications. Specifically, implementation may also help researchers study how to integrate AI solutions into shop-floor practice and how they should be presented to practitioners (for example, as a measure of probabilities calculated from the data, in visual form, or as direct recommendations). To achieve this, researchers need to work with practitioners to build a shared understanding of priorities and limitations, thereby improving their ability to create AI solutions with real-world impact. Such efforts may help advance the field of throughput bottleneck research whilst creating an impact within the industry.

5.8. Digital twin for throughput bottleneck analysis

The existing AI solutions process the historical production system data, analyse them, and give insights on throughput bottlenecks. These insights can then be used by practitioners to plan for appropriate elimination actions. However, there are limited possibilities to verify if the

planned actions might effectively eliminate throughput bottlenecks. This process can be further improved by integrating AI with discrete event simulation models of the production system, commonly called a digital twin [89]. Exploring the possibilities of using AI with discrete-event simulation models is also a potential research direction within the throughput bottleneck analysis research field.

One example of a potential digital twin setup for throughput bottleneck analysis is described in the following sentences. A digital twin can consist of three elements: (1) real-time production system data (including all possible types of data such as product types, machine data, list of historical actions taken in the production system, etc.) (2) a discrete event simulation model of the production system, and (3) an AI agent. The real-time production system data can be continuously fed to the simulation model using automated input data management techniques [90]. The model can then produce the simulated data of the production system and the simulated data can, in turn, be used to train the AI agent. Once trained, the AI agents can automatically analyse the throughput bottlenecks, identify a set of elimination actions, test the impact of each of the actions using the simulation model and prescribe a concrete set of actions that need to be taken in the real-world. Practitioners can then implement those actions to eliminate the throughput bottlenecks. The feedback from taking those actions in the real-world can be given as input to the AI agent so that the AI agent can learn from the feedback (e.g., using reinforcement learning), become more accurate, and find novel ways in prescribing actions over time. Determining the exact setup procedure of such a digital twin (including the components of the digital twin) and its operationalization procedure needs rigorous research.

6. Practical recommendations

As shown throughout Section 4, academic research provides practitioners a set of AI solutions for throughput bottleneck analysis. However, taking these academic research efforts into real-world practice requires more than science. This is where practitioners play an important role. The authors offer seven practical recommendations to aid successful adoption of the various AI solutions from the literature. These are: (1) start small, (2) perfect data myth, (3) augmentation, (4) avoid incomplete data, (5) standardise data collection (6) teamwork and (7) reporting structure.

6.1. Start small

To unleash and realise the full potential of AI in throughput bottleneck analysis, practitioners need to start small, so they can be quick in implementing AI solutions and demonstrate successful AI in factories. For example, practitioners may start their journey of implementing AI in throughput bottleneck analysis by deploying data collection technologies to collect machine data from their production systems. Once this data collection is started, practitioners may immediately begin tracking real-time throughput bottlenecks using the AI tools (as summarized in Section 4.1.2) and take real-time action. This simple way of tracking throughput bottlenecks may help reduce variations in throughput. Once sufficient data has been collected over time, practitioners may revisit it and identify long-term throughput bottlenecks using AI tools (as summarized in Section 4.1.1). They may also diagnose throughput bottlenecks using AI tools (see Section 4.2), which can help to significantly increase throughput. By accumulating data over time, patterns might start to emerge. This allows for predicting the likelihood of machines acting as throughput bottlenecks in the future (see Section 4.3). The next step is to eliminate throughput bottlenecks by establishing correlational rules between throughput bottlenecks, non-throughput bottlenecks, historical actions taken on such scenarios and desired future throughput from the system. By establishing these correlational patterns, prescriptive AI tools (as summarized in Section 4.4) may provide a concrete set of actions.

6.2. Perfect data myth

In real-world practice, practitioners commonly initiate AI solutions once they have perfect data. This might be disadvantageous. Having perfect data is always a moving target. Practitioners may start with the data they have and try out different AI solutions and iterate from there. For instance, the AI solutions summarized in Table 3 may be implemented using event-log-type data (data that stores machines' activities with corresponding time stamps). [79] (p.536) has shown (through a real-world test study) that ANDON light information from machines may also be used to predict throughput bottlenecks. Such research efforts demonstrate that such simple data sets are good enough for an AI solution to have an impact in practice. Furthermore, [79] (p.542) argues that, even when the accuracy of such simple data sets is not very high, it still positively impacts improvements in throughput bottleneck management. More impact may be obtained in the future if a simple AI solution is already implemented and running on the shop floor, upon which iterations may occur.

6.3. Augmentation

While AI solutions may deliver powerful insights, practitioners need to understand that such insights require augmentation. Practitioners commonly set unrealistic expectations for AI solutions, for example to always provide 100 % accuracy. However, in the literature, the demonstrated accuracy (even with the most sophisticated AI tools) in predicting throughput bottlenecks is close to 90 % [36]. Closing this accuracy gap needs augmentation, using practitioners' deep domain knowledge of factory dynamics. Practitioners' feedback may also be used to train AI and improve its performance over time. This also contributes to XAI, as discussed in Section 5.6. Buying into the hype that AI may work perfectly, instantaneously and without supervision is often detrimental to the effective use of AI solutions.

6.4. Avoid incomplete data

Practitioners must be extra careful whenever there are manual data entry procedures. For example, it is still a common practice for maintenance practitioners to record a problem and the actions taken in free-form sentences. Although there are AI tools to interpret these sentences, practitioners need to write effectively and ensure no information is missing. This is critical, for example, when designing prescriptive AI tools (as shown in Section 4.4) which use historical data sets to recommend future actions.

6.5. Standardise data collection

Practitioners may consider standardising their data collection systems. The current practice is often to collect different types of operational data from different systems. For example, maintenance-related data is stored in computerised maintenance management systems (CMMS), whilst event log data describing machine activities are stored in a manufacturing execution system (MES). Storing data in different systems might pose a challenge when building AI solutions. For example, the solutions described in [80,36,81] use various types of operational data, such as maintenance data, cycle time data and so on. Storing this data in a common database is more advantageous for effective implementation than having to retrieve the data from different sources. Moreover, having a common database may also help when tracking products with operational machine information from a specific time window as well as revealing correlational patterns.

6.6. Teamwork

In real-world practice, data scientists take existing academic AI solutions, adapt them to the real-world environment and implement them.

However, this process is not the entirety of practitioners' involvement. Their input is needed at every stage if AI solutions are to be implemented in real-world practice. During adaption and implementation, data scientists must fully consider the dynamics and other idiosyncrasies of the particular factory. This information is best provided by practitioners since they will have developed deep domain knowledge over time. Practitioners and data scientists should therefore work as a team during the implementation process, enriching academic solutions with relevant practical information. An example of this adaptation procedure (with relevant input from data scientists and practitioners) is demonstrated in [75] (p.2). Such involvement may also help practitioners understand the mechanisms underlying AI solutions, leading to increased transparency of AI solutions, thereby contributing to XAI (see Section 5.6).

6.7. Reporting structure

If AI solutions are implemented in real-world settings, they should be reported in detail. Reports should include an unambiguous description of the entire AI solution, with details of data collection techniques, data pre-processing procedures (such as missing data management, inlier and outlier management), modelling approaches (including the reasons for choosing a particular model), software used (including library packages) and evaluation techniques. For example, [80] (p.6) demonstrates the AI solution for throughput bottleneck prediction on a real-world assembly line. However, it presents limited information and examples covering the data pre-processing procedure, it presents no explanation of the fusion of different types of input data, the software used to develop and test the AI solution and so on. Similarly, [81] (p.287) provides only limited information explaining the data collection, fusion procedure, and data pre-processing techniques used. Such gaps raise various challenges (such as management of missing data, downtime data due to external reasons such as power failures, thought process on hyper-parameter tuning of AI solutions) when researchers and practitioners adopt the AI solutions. Also, if possible full data sets that are used to create and test the AI solutions can be publicly provided. This can then be used by researchers and practitioners to conduct more experiments and further improve the AI solutions. Reporting the full information will help the academic and practitioner community to understand AI solutions, enhance their reproducibility and aid the successful transition of AI solutions, from their development in academic research to implementation in real-world practice.

7. Limitations

Although this study used an established and meticulous review methodology, it has some limitations. Firstly, the study is limited to collecting and analysing all the relevant publications retrieved from Scopus. However, Scopus is one of the standard databases used frequently by researchers to collect publications in the manufacturing field. Still, when the throughput bottlenecks research field grows, future research can also use other databases (e.g., Web of Science, Google Scholar, and IEEE Xplore) to collect relevant publications. Secondly, this study only identifies 16 papers which can be considered a small sample delimited to a small set of researchers. However, owing to the systematic literature review process, the final set of 16 papers was in fact representative for the research field of AI driven throughput bottleneck analysis. In line with more research efforts and a growing number of publications, it is relevant to replicate our review in the future and compare the results. Thirdly, this paper uses the Gartner data analytics framework to classify the papers, and it needs to be acknowledged that also other frameworks could be used. Although the Gartner analytics framework is easily understandable and relevant for practitioners, a potential future direction is to develop different classification frameworks to effectively communicate AI solutions to researchers and practitioners. Lastly, this paper considers AI to include a large variety of tools, ranging from statistical to deep and reinforcement learning (see Section 2.3). This is justified by the lack of a commonly accepted definition of AI and the fact that all AI-based tools can provide value from processing and analysing data in real-world practice. Still, the majority of the articles in our review consist of rule-based classification (see "Modelling approach" in Tables 3–7). Therefore, as the field of AI for throughput bottleneck analysis develops and matures over time, we expect to see both a broader range as well as more sophisticated modelling approaches. If and when a commonly accepted definition of AI is established in the future, it is relevant to replicate this study to further align the range of applications with the scope of AI solutions.

8. Conclusions

If higher levels of productivity are to be achieved, throughput bottlenecks in production systems must be eliminated. However, throughput bottlenecks first need to be analysed. Over the last decade, various research efforts have focused on developing AI solutions to help analyse throughput bottlenecks. In this paper, a systematic literature review was conducted to map the field and provide a state-of-the-art of AI solutions for throughput bottleneck analysis. A final literature set of 16 publications were retrieved as a result of the systematic review methodology. Using the Gartner data analytics framework, the literature was categorized into four categories: identify, diagnose, predict, and prescribe. For each category, the AI solution architecture was synthesized and summarised in terms of input data, modeling approach, and output data. From the categorisation, it has been identified that maximum research efforts were devoted to developing AI solutions to identify and predict categories, and fewer efforts were devoted to diagnose and prescribe categories. Although knowing the throughput bottleneck locations is the first step to eliminate them, future research efforts need to be more focused on the diagnosis and prescription of specific elimination actions to eliminate throughput bottlenecks. Additional promising future research directions (e.g. combing the emerging trends in AI such as XAI, HITL and digital twins, with the throughput bottleneck analysis problems) have also been identified and proposed based on real-world practice. Furthermore, practical recommendations (e.g. starting small, augmentation and teamwork) were also provided, which will help practitioners to implement the existing AI solutions for throughput bottleneck analysis. These recommendations will further advance the field of throughput bottlenecks analysis in real-world industrial practice.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

The authors would like to thank the FFI programme (funded by VINNOVA, the Swedish Energy Agency and the Swedish Transport Administration) for its funding of the Data Analytics in Maintenance Planning research project (DAiMP) [Grant number: 2015-06887], under which this research was conducted. The authors would also like to thank all the industrial partners in the DAiMP research project (AB Volvo, Volvo Cars, Scania AB and SKF AB), for sharing their views on the importance of AI to throughput bottleneck analysis in manufacturing. This work was conducted under the Sustainable Production Initiative and Production Area of Advance at Chalmers.

References

[1] Schmenner RW. The pursuit of productivity. Prod Oper Manage 2015;24:341–50. https://doi.org/10.1111/poms.12230.
[2] Li J, Blumenfeld DE, Huang N, Alden JM. Throughput analysis of production systems : recent advances and future topics. Int J Prod Res 2009;47:3823–51. https://doi.org/10.1080/00207540701829752.

[3] Alavian P, Eun Y, Meerkov SM, Zhang L. Smart production systems : automating decision- making in manufacturing environment. Int J Prod Res 2019:1–18. https://doi.org/10.1080/00207543.2019.1600765.

[4] Alavian P, Denno P, Meerkov SM. Multi-job production systems: definition, problems, and product-mix performance portrait of serial lines. Int J Prod Res 2017;55:7276–301. https://doi.org/10.1080/00207543.2017.1338779.

[5] Wu K, Zheng M, Shen Y. A generalization of the Theory of Constraints: choosing the optimal improvement option with consideration of variability and costs. IISE Trans 2020;52:276–87. https://doi.org/10.1080/24725854.2019.1632503.

[6] Wu K, Zhou Y, Zhao N. Variability and the fundamental properties of production lines. Comput Ind Eng 2016;99:364–71. https://doi.org/10.1016/j.cie.2016.04.014.

[7] Goldrat E, Cox J. The Goal: A Process of Ongoing Improvement. Third Revi. Great Barrington, MA: North River Press; 1990.

[8] Schmenner RW, Swink ML. On theory in operations management. J Oper Manage 1998;17:97–113. https://doi.org/10.1016/S0272-6963(98)00028-X.

[9] Zhang M, Matta A. Models and algorithms for throughput improvement problem of serial production lines via downtime reduction. IISE Trans 2020;0:1–15. https://doi.org/10.1080/24725854.2019.1700431.

[10] Zou J, Chang Q, Arinez J, Xiao G. Production performance prognostics through model-based analytical method and recency-weighted stochastic approximation method. J Manuf Syst 2018;47:107–14. https://doi.org/10.1016/j.jmsy.2018.04.017.

[11] Roser C, Nakano M, Tanaka M. Shifting bottleneck detection. In: Yucesan E, Chen C-H, Snowdon J, Charnes J, editors. Proc. 2002 Winter Simul. Conf, 2; 2002. https://doi.org/10.1109/WSC.2002.1166360.

[12] Roser C, Nakano M, Tanaka M. A practical bottleneck detection method. In: Peters B, Smith J, Medeiros D, Rohrer M, editors. Proc. 2001 Winter Simul. Conf; 2001. p. 949–53. https://doi.org/10.1109/WSC.2001.977398.

[13] Li L, Chang Q, Ni J. Data driven bottleneck detection of manufacturing systems. Int J Prod Res 2009;47:5019–36. https://doi.org/10.1080/00207540701881860.

[14] Pehrsson L, Ng AHC, Bernedixen J. Automatic identification of constraints and improvement actions in production systems using multi-objective optimization and post-optimality analysis. J Manuf Syst 2016;39:24–37. https://doi.org/10.1016/j.jmsy.2016.02.001.

[15] Li L. A systematic-theoretic analysis of data-driven throughput bottleneck detection of production systems. J Manuf Syst 2018;47:43–52. https://doi.org/10.1016/j.jmsy.2018.03.001.

[16] Thürer M, Ma L, Stevenson M, Roser C. Bottleneck detection in high-variety make-to-Order shops with complex routings: an assessment by simulation. Prod Plan Control 2021;0:1–12. https://doi.org/10.1080/09537287.2021.1885795.

[17] Colledani M, Ekvall M, Lundholm T, Moriggi P, Polato A, Tolio T. Analytical methods to support continuous improvements at Scania. Int J Prod Res 2010;48:1913–45. https://doi.org/10.1080/00207540802538039.

[18] Kuo YH, Kusiak A. From data to big data in production research: the past and future trends. Int J Prod Res 2019;57:4828–53. https://doi.org/10.1080/00207543.2018.1443230.

[19] Bukkapatnam STS, Afrin K, Dave D, Kumara SRT. Machine learning and AI for long-term fault prognosis in complex manufacturing systems. CIRP Ann 2019;68:459–62. https://doi.org/10.1016/j.cirp.2019.04.104.

[20] Arinez JF, Chang Q, Gao RX, Xu C, Zhang J. Artificial intelligence in advanced manufacturing: current status and future outlook. J Manuf Sci Eng 2020;142:1–16. https://doi.org/10.1115/1.4047855.

[21] Lee J, Ni J, Singh J, Jiang B, Azamfar M, Feng J. Intelligent maintenance systems and predictive manufacturing. J Manuf Sci Eng 2020:142. https://doi.org/10.1115/1.4047856.

[22] Whetten DA. What constitutes a theoretical contribution? Acad Manage Rev Manage Rev 1989;14:490–5.

[23] Hopp WJ. The lenses of lean : visioning the science and practice of efficiency. J Oper Manage 2020:1–17. https://doi.org/10.1002/joom.1115.

[24] Hopp WJ, Spearman ML. Factory physics: foundations of manufacturing management. 2nd ed. Long Grove, IL: Waveland Press; 2008.

[25] Schmitz JPM, Van Beek DA, Rooda JE. Chaos in discrete production systems? J Manuf Syst 2002;21:236–46. https://doi.org/10.1016/s0278-6125(02)80164-9.

[26] Wu K. An examination of variability and its basic properties for a factory. IEEE Trans Semicond Manuf 2005;18:214–21.

[27] Romero-Silva R, Marsillac E, Shaaban S, Hurtado-Hernández M. Serial production line performance under random variation: dealing with the 'Law of Variability. J Manuf Syst 2019;50:278–89. https://doi.org/10.1016/j.jmsy.2019.01.005.

[28] Hillier FS, Boling R. The effect of some design factors on the efficiency of production lines with variable element times. J Ind Eng 1966:651–8.

[29] Taylor GD, Heragu SS. A comparison of mean reduction versus variance reduction in processing times in flow shops. Int J Prod Res 1999;37:1919–34. https://doi.org/10.1080/002075499190833.

[30] Betterton CE, Silver SJ. Detecting bottlenecks in serial production lines – a focus on interdeparture time variance. Int J Prod Res 2012;50:4158–74. https://doi.org/10.1080/00207543.2011.596847.

[31] Li L, Chang Q, Ni J, Biller S. Real time production improvement through bottleneck control. Int J Prod Res 2009;47:6145–58. https://doi.org/10.1080/00207540802244240.

[32] Gopalakrishnan M, Skoogh A, Christoph L. Simulation based planning of maintenance activities by a shifting priority method. In: Tolk A, Diallo S, Ryzhov I, Yilmaz L, Buckley S, Miller JA, editors. Proceeding 2014 Winter Simul. Conf; 2014. p. 2600–8.

[33] Li L. Bottleneck detection of complex manufacturing systems using a data-driven method. Int J Prod Res 2009;47:6929–40. https://doi.org/10.1080/00207540802427894.

[34] Subramaniyan M, Skoogh A, Gopalakrishnan M, Salomonsson H, Hanna A, Lämkull D. An algorithm for data-driven shifting bottleneck detection. Cogent Eng 2016;3:1–19. https://doi.org/10.1080/23311916.2016.1239516.

[35] Li L, Qing C, Xiao G, Ambani S. Throughput bottleneck prediction of manufacturing systems using time series analysis. J Manuf Sci Eng 2011;133:1–8. https://doi.org/10.1115/1.4003786.

[36] Fang W, Guo Y, Liao W, Huang S, Yang N, Liu J. A Parallel Gated Recurrent Units (P-GRUs) network for the shifting lateness bottleneck prediction in make-to-order production system. Comput Ind Eng 2020;140:106246. https://doi.org/10.1016/j.cie.2019.106246.

[37] Subramaniyan M, Skoogh A, Salomonsson H, Bangalore P, Gopalakrishnan M, Sheikh Muhammad A. Data-driven algorithm for throughput bottleneck analysis of production systems. Prod Manuf Res 2018:6. https://doi.org/10.1080/21693277.2018.1496491.

[38] Chiang S-Y, Kuo C-T, Meerkov S. C-bottlenecks in serial production lines: identification and application. Math Probl Eng 2001;7:543–78.

[39] Chiang S-Y, Kuo C-T, Meerkov SM. DT-bottlenecks in serial production lines: theory and application. IEEE Trans Robot Autom 2000;16:567–80. https://doi.org/10.1109/70.880806.

[40] Olsen TL, Tomlin B. Industry 4.0: opportunities and challenges for operations management. Manuf Serv Oper Manage 2020;22:113–22. https://doi.org/10.1287/msom.2019.0796.

[41] Wuest T, Weimer D, Irgens C, Thoben K. Machine learning in manufacturing : advantages, challenges, and applications. Prod Manuf Res 2016;4:1–23. https://doi.org/10.1080/21693277.2016.1192517.

[42] Lee J, Davari H, Singh J, Pandhare V. Industrial Artificial Intelligence for industry 4.0-based manufacturing systems. Manuf Lett 2018;18:20–3. https://doi.org/10.1016/j.mfglet.2018.09.002.

[43] Wang J, Ma Y, Zhang L, Gao RX, Wu D. Deep learning for smart manufacturing: methods and applications. J Manuf Syst 2018;48:144–56. https://doi.org/10.1016/j.jmsy.2018.01.003.

[44] Skordilis E, Moghaddass R. A deep reinforcement learning approach for real-time sensor-driven decision making and predictive analytics. Comput Ind Eng 2020;147:106600. https://doi.org/10.1016/j.cie.2020.106600.

[45] Kusiak A. Convolutional and generative adversarial neural networks in manufacturing. Int J Prod Res 2020;58:1594–604. https://doi.org/10.1080/00207543.2019.1662133.

[46] Tewari S, Dwivedi UD. Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs. Comput Ind Eng 2019;128:937–47. https://doi.org/10.1016/j.cie.2018.08.018.

[47] Harding Ja, Shahbaz M, Srinivas Kusiaka. Data mining in manufacturing: a review. J Manuf Sci Eng 2006;128:969. https://doi.org/10.1115/1.2194554.

[48] Baptista M, Sankararaman S, De Medeiros IP, Nascimento C, Prendinger H, et al. Forecasting fault events for predictive maintenance using data-driven techniques and ARMA modeling. Comput Ind Eng 2018;115:41–53. https://doi.org/10.1016/j.cie.2017.10.033.

[49] Flath CM, Stein N. Towards a data science toolbox for industrial analytics applications. Comput Ind 2018;94:16–25. https://doi.org/10.1016/j.compind.2017.09.003.

[50] Yu J. Pattern recognition of manufacturing process signals using Gaussian mixture models-based recognition systems. Comput Ind Eng 2011;61:881–90. https://doi.org/10.1016/j.cie.2011.05.022.

[51] Rousseaux F. BIG DATA and data-driven intelligent predictive algorithms to support creativity in industrial engineering. Comput Ind Eng 2017;112:459–65. https://doi.org/10.1016/j.cie.2016.11.005.

[52] Hutson M. AI Glossary: artificial intelligence, in so many words. Science 2017;357:19. https://doi.org/10.1126/science.357.6346.19.

[53] Diez-olivan A, Del J, Galar D, Sierra B. Data fusion and machine learning for industrial prognosis : trends and perspectives towards Industry 4. 0. Inf Fusion 2019;50:92–111. https://doi.org/10.1016/j.inffus.2018.10.005.

[54] Gao RX, Wang L, Helu M, Teti R. Big data analytics for smart factories of the future. CIRP Ann 2020:1–25. https://doi.org/10.1016/j.cirp.2020.05.002.

[55] Mayring P. Qualitative content analysis. A companion to qualitative research. FORUM Qual Soc Res Sozialforsch 2000;1:159–76.

[56] Krippendorff K. Content analysis: an introduction to its methodology. Sage Publications; 2018.

[57] Govindan K, Soleimani H, Kannan D. Reverse logistics and closed-loop supply chain: a comprehensive review to explore the future. Eur J Oper Res 2015;240:603–26. https://doi.org/10.1016/j.ejor.2014.07.012.

[58] Özceylan E, Kalayci CB, Güngör A, Gupta SM. Disassembly line balancing problem: a review of the state of the art and future directions. Int J Prod Res 2019;57:4805–27. https://doi.org/10.1080/00207543.2018.1428775.

[59] Montoya FG, Alcayde A, Baños R, Manzano-Agugliaro F. A fast method for identifying worldwide scientific collaborations using the Scopus database. Telemat Informatics 2018;35:168–85. https://doi.org/10.1016/j.tele.2017.10.010.

[60] Abedinnia H, Glock CH, Grosse EH, Schneider M. Machine scheduling problems in production: a tertiary study. Comput Ind Eng 2017;111:403–16. https://doi.org/10.1016/j.cie.2017.06.026.

[61] Trigueiro de Sousa Junior W, Barra Montevechi JA, de Carvalho Miranda R, Teberga Campos A. Discrete simulation-based optimization methods for industrial engineering problems: a systematic literature review. Comput Ind Eng 2019;128:526–40. https://doi.org/10.1016/j.cie.2018.12.073.

[62] Egger J, Masood T. Augmented reality in support of intelligent manufacturing – a systematic literature review. Comput Ind Eng 2020;140:106195. https://doi.org/10.1016/j.cie.2019.106195.
[63] Alrabghi A, Tiwari A. State of the art in simulation-based optimisation for maintenance systems. Comput Ind Eng 2015;82:167–82. https://doi.org/10.1016/j.cie.2014.12.022.
[64] Cao Z, Deng J, Liu M, Wang Y. Bottleneck prediction method based on improved adaptive network-based fuzzy inference system (ANFIS) in semiconductor manufacturing system. Chin J Chem Eng 2012;20:1081–8. https://doi.org/10.1016/S1004-9541(12)60590-4.
[65] Subramaniyan M, Skoogh A, Muhammad AS, Bokrantz J, Johansson B, Roser C. A generic hierarchical clustering approach for detecting bottlenecks in manufacturing. J Manuf Syst 2020:55. https://doi.org/10.1016/j.jmsy.2020.02.011.
[66] Chandler N, Hostmann B, Rayner N, Herschel G. Gartner' s business analytics framework. Gartner 2011:1–12.
[67] Heo W, Lee JM, Park N, Grable JE. Using artificial neural network techniques to improve the description and prediction of household financial ratios. J Behav Exp Financ 2020;25:100273. https://doi.org/10.1016/j.jbef.2020.100273.
[68] Li L, Ni J. Short-term decision support system for maintenance task prioritization. Int J Prod Econ 2009;121:195–202. https://doi.org/10.1016/j.ijpe.2009.05.006.
[69] Krishnan S, Dev AS, Suresh R, Sumesh A, Rameshkumar K. Bottleneck identification in a tyre manufacturing plant using simulation analysis and productivity improvement. Mater Today Proc 2018;5:24720–30. https://doi.org/10.1016/j.matpr.2018.10.270.
[70] Roh P, Kunz A, Netland T. Data-driven detection of moving bottlenecks in multi-variant production lines. IFAC-PapersOnLine 2018;51:158–63. https://doi.org/10.1016/j.ifacol.2018.08.251.
[71] Yu C, Matta A. Data-driven bottleneck detection in manufacturing systems: a statistical approach. Int J Prod Res 2016;54:6317–22. https://doi.org/10.1080/00207543.2015.1126681.
[72] Subramaniyan M, Skoogh A, Gopalakrishnan M, Hanna A. Real-time data-driven average active period method for bottleneck detection. Int J Des Nat Ecodyn 2016:11. https://doi.org/10.2495/DNE-V11-N3-428-437.
[73] Wang LC, Chu PC, Lin SY. Impact of capacity fluctuation on throughput performance for semiconductor wafer fabrication. Robot Comput Integr Manuf 2019;55:208–16. https://doi.org/10.1016/j.rcim.2018.03.005.
[74] Thürer M, Stevenson M. Bottleneck-oriented order release with shifting bottlenecks: an assessment by simulation. Int J Prod Econ 2018;197:275–82. https://doi.org/10.1016/j.ijpe.2018.01.010.
[75] Subramaniyan M, Skoogh A, Muhammad AS, Bokrantz J, Johansson B, Roser C. A data-driven approach to diagnosing throughput bottlenecks from a maintenance perspective. Comput Ind Eng 2020:150. https://doi.org/10.1016/j.cie.2020.106851.
[76] Chang Q, Biller S, Xiao G. Transient analysis of downtimes and bottleneck dynamics in serial manufacturing systems. J Manuf Sci Eng 2010:132. https://doi.org/10.1115/1.4002562.
[77] Patti AL, Watson KJ. Downtime variability: the impact of duration-frequency on the performance of serial production systems. Int J Prod Res 2010;48:5831–41. https://doi.org/10.1080/00207540903280572.
[78] Hopp WJ, Iravani SMR, Shou B. A diagnostic tree for improving production line performance. Prod Oper Manage 2007;16:77–92. https://doi.org/10.1111/j.1937-5956.2007.tb00167.x.
[79] Subramaniyan M, Skoogh A, Salomonsson H, Bangalore P, Bokrantz J. A data-driven algorithm to predict throughput bottlenecks in a production system based on active periods of the machines. Comput Ind Eng 2018:125. https://doi.org/10.1016/j.cie.2018.04.024.
[80] Lai X, Shui H, Ni J. A two-layer long short-term memory network for bottleneck prediction in multi-job manufacturing systems. ASME 2018 13th Int. Manuf. Sci. Eng. Conf. MSEC 2018, 3; 2018. p. 1–9. https://doi.org/10.1115/MSEC2018-6678.
[81] Huang B, Wang W, Ren S, Zhong RY, Jiang J. A proactive task dispatching method based on future bottleneck prediction for the smart factory. Int J Comput Integr Manuf 2019;32:278–93. https://doi.org/10.1080/0951192X.2019.1571241.
[82] Subramaniyan M, Skoogh A, Sheikh Muhammad A, Bokrantz J, Turanoğlu Bekar E. A prognostic algorithm to prescribe improvement measures on throughput bottlenecks. J Manuf Syst 2019:53. https://doi.org/10.1016/j.jmsy.2019.07.004.
[83] Bagheri B, Yang S, Kao HA, Lee J. Cyber-physical systems architecture for self-aware machines in industry 4.0 environment. IFAC-PapersOnLine 2015;28:1622–7. https://doi.org/10.1016/j.ifacol.2015.06.318.
[84] Sheng VS, Provost F. Get another label? Improving data quality and data mining using multiple, noisy labelers categories and subject descriptors. Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min 2008:614–22.
[85] Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. Npj Digit Med 2019:2. https://doi.org/10.1038/s41746-019-0189-7.
[86] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15. https://doi.org/10.1038/s42256-019-0048-x.
[87] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012.
[88] Hrnjica B, Softic S. Explainable AI in manufacturing: a predictive maintenance case study. In: IFIP Int. Conf. Adv. Prod. Manag. Syst; 2020. https://doi.org/10.1007/978-3-030-57997-5_8.
[89] Lugaresi G, Matta A. Automated manufacturing system discovery and digital twin generation. J Manuf Syst 2021;59:51–66. https://doi.org/10.1016/j.jmsy.2021.01.005.
[90] Skoogh A, Johansson B, Stahre J. Automated input data management: evaluation of a concept for reduced time consumption in discrete event simulation. Simulation 2012;88:1279–93. https://doi.org/10.1177/0037549712443404.