Flexible Services and Manufacturing Journal
https://doi.org/10.1007/s10696-023-09496-9

A reinforcement learning/ad-hoc planning and scheduling mechanism for flexible and sustainable manufacturing systems

Panagiotis D. Paraschos[1] · **Georgios K. Koulinas**[1] · **Dimitrios E. Koulouriotis**[1]

Accepted: 4 April 2023

Abstract

The process scheduling is still considered a crucial subject for manufacturing industry, due to the ever-changing circumstances dictated by the nowadays product demand and customer trends. These conditions are often associated with increasing costs and energy consumption, considerably affecting the long-term sustainability of manufacturing plants. To mitigate that effect, one should create an effective strategy tailoring integrated operations and processes to the customer demand and trends faced by the nowadays industry. A well-known approach to this matter is the technologies introduced by manufacturing paradigms, e.g., Industry 4.0 and smart manufacturing. As suggested in literature, these technologies are capable of helping decision-makers by continuously gathering significant information about the state of machinery and manufactured goods. This information is thereafter utilized to identify weaknesses and strengths demonstrated within manufacturing plants. To this end, the present paper presents a process optimization framework implemented in a three-stage production line prone to systematic degradation faults. Aiming at strengthening profitability, the framework engages reinforcement learning with ad-hoc manufacturing/maintenance control in decision-making carried out in implemented machines. Simulation experiments showed improved process planning and inventory management enabling cost-effective green and sustainable manufacturing in manufacturing plants.

Keywords Nonconforming items · Waste management · Parametric control · Lean manufacturing

✉ Panagiotis D. Paraschos
pparasc@pme.duth.gr

Georgios K. Koulinas
gkoulina@pme.duth.gr

Dimitrios E. Koulouriotis
jimk@pme.duth.gr

[1] Department of Production and Management Engineering, Democritus University of Thrace, Xanthi, Greece

Published online: 20 April 2023

1 Introduction

Due to the persistent rising of energy consumption and operational costs, the long-term sustainability is more than ever a concerning and pressing matter for manufacturing plants. In this respect, it is heavily relied on the process plan and design defined by manufacturing experts involved in decision-making. Along with process planning, the same experts are engaged in scheduling inspections on the items generated during the manufacturing process in order to increase the output quality and minimize waste, which is the main focus and goal of green and lean manufacturing concepts. These concepts dictate the reuse of the already generated material contributing to the decrease of energy consumption (Lim et al. 2022; Jum'a et al. 2022). Thus, to create an energy-efficient and cost-effective manufacturing system, the design process followed by experts should involve the collection of a considerable amount of data on several aspects of manufacturing control, such as the energy consumption, or machine availability (Ahmad et al. 2022; Antons and Arlinghaus 2022). Traditionally, this collection is conducted using a set of sampling devices, e.g., meters and sensors. The procured data could be thereafter processed by machine learning techniques, such as support vector machines (Jeong 2022). However, utilizing merely sampling devices for knowledge extraction, it is probable that manufacturing facilities would be confronted with several challenges, e.g., data inconsistency, in their effort to extract the required data from their operations and processes (Corallo et al. 2022). Serving as a solution to the problem related to data collection, the digitization of integrated processes and operations has enabled manufacturing experts to simulate system behavior and design final products considering several factors, such as job execution time and tardiness (Kenett and Bortman 2022; Iqbal et al. 2022). Adopting such an approach, the aim is to design optimal, flexible and energy-wise operations for manufacturing plants acknowledging the requirements imposed by customers and emerging circumstances, e.g., carbon footprint (He et al. 2015). Although, this digitization substantially requires the integration of advanced and intelligent technologies, e.g., cyber-physical systems (CPS), introduced in manufacturing by initiatives, e.g., smart manufacturing and Industry 4.0 (Karnik et al. 2022).

In an effort to demonstrate the applicability of such technologies in manufacturing environments, an intelligent process planning framework is presented in the present paper. This application attempts to jointly schedule activities involving material management and system maintenance. To achieve that, it couples ad-hoc process planning with reinforcement learning (RL) decision-making. Following this type of decision-making, the framework communicates with a manufacturing environment by utilizing two decision-making agents. In this regard, these agents are implemented in a specified number of manufacturing machines. Using this setup, the intention is to improve the quality inspection carried out early on and avoid the authorization of redundant processing activities that may have a negative effect upon profitability. The latter is facilitated by the introduction of parametric manufacturing and maintenance policies, e.g., KANBAN and opportunistic maintenance. These policies are frequently employed in the process

control literature and the real-world manufacturing industry. Complementing reinforcement learning, they are capable of ensuring high-quality stock generation and high system maintenance (Paraschos et al. 2022, 2021).

For evaluation purposes, a multi-stage manufacturing/remanufacturing system serves as a manufacturing environment for the proposed framework. It is common in the real-world and complex manufacturing environments, e.g., supply chains, to process products over multiple stages before their delivery to customers. To better understand the multi-stage manufacturing process, let us provide an example. This example involves an assembly system that integrates three item processing stages. In the 1st stage, the system receives and process the raw materials to prepare them for the next stage of manufacturing. In the 2nd stage, the main item assembly is carried out. The assembled items are moved into the final stage. In this stage, these items receive the final refinements and delivered to the customers. Likewise, the studied one follows a similar architecture to that of the real-world ones by integrating multiple machines, processing products. Analytically, it comprises three separate machines generating and reusing goods. Completed items, either work-in-progress or final, are stored in storehouses until they are moved into subsequent stages of manufacturing process, or procured by customers. Similar to real-world production lines, the studied one operates under uncertain and fluctuating conditions, e.g., customer arrivals. In this regard, random events have an explicit effect on the integrated operations and functionality. Furthermore, due to their uninterrupted operation, the involved machines are prone to faults degrading their productivity and maximizing the likelihood of malfunction. Based on these, one can assume that the corresponding costs are likely to be increased and become unsustainable. Therefore, the proposed framework is implemented in an attempt to introduce flexible and intelligent planning within the described manufacturing context while simultaneously adopting and enabling concepts relevant to green and lean manufacturing.

The following key-points summarize the paper's contribution:

(a) A three-stage production line is studied. Its architecture and behavior bear resemblance to the ones assumed by the real-world stochastic manufacturing environments. In this respect, a discrete event approach is exploited to model the functionality of the system examined under real world-like events, such as frequent equipment failures.
(b) An optimization framework is proposed to enhance the productivity, sustainability and profitability of the system with green and sustainable practices. It integrates two decision-making agents interacting with the 1st and 3rd manufacturing machines involved in the studied production line. In addition to reinforcement learning, ad-hoc manufacturing/maintenance policies (for example, CONWIP and Opportunistic maintenance) are exploited as well. The intention behind this implementation is the generation of optimal, or near optimal joint control policies for integrated activities, e.g., remanufacturing, for the reduction of material waste and redundant activity authorizations.
(c) A thorough experimental study is conducted. In this study, a series of experiments study the performance of the presented framework integrated in the

context of the studied manufacturing environment, behaving under real world-simulated conditions.

The rest of the paper has the following structure. The related work pertinent to the present paper is acknowledged in Sect. 2. In Sect. 3, the production line examined is presented. Section 4 details the employed optimization framework. Section 5 analyzes the experiments conducted for the evaluation of the framework's functionality. Section 6 provides research implications and proposes future research involving the presented framework.

2 Literature review

Frequently, the relevant research on process control evaluates production systems processing either one (Adeinat et al. 2022), or multiple items (Beraudy et al. 2022). The processing of these items is carried out in machines set either parallelly (Zhang and Chen 2022), or in a serial manner (Tu and Zhang 2022). As these machines are prone to failures, it is likely that they could generate non-conforming products, along with standard ones (Ye et al. 2021).

Given the circumstances above, publications aim to decrease operational costs employing approaches, ranging from control charts to genetic algorithms. Formulating a mixed integer programming-based control model and scheduling production activities according to the shortest processing time (SPT) rule, Bhosale and Pawar (2019) endeavored to improve material management by means of genetic algorithm. Aiming at optimizing the productivity and profitability of manufacturing plants, Hoseinpour et al. (2021, 2020) formulated a mathematical model, considering production planning with resource constraints in an effort to decide on outsourcing manufacturing activities by means of optimization algorithms, e.g., particle swarm optimization. Gharbi et al. (2022) considered the life of perishable product inventory in their formulated stochastic production control model in the context of failure-prone manufacturing systems. Metzker et al. (2023) addressed the lot-sizing problem, proposing a dynamic programming-based production optimization model for carrying out multi-period decisions under uncertainty. For multi-stage manufacturing systems, Manafzadeh Dizbin and Tan (2019) adopted a Base Stock-based production model for creating product stock, determining the frequency of manufacturing jobs according to inter-event data generated by a Markov model. Kim and Kim (2022) presented a production model to control the make-to-stock/make-to-order production rate in a two-stage manufacturing system. Similar to production control, maintenance activities are performed on the basis of ad-hoc models in order to preempt system degradation. For example, Li et al. (2022) considered an imperfect maintenance control model to decrease maintenance cost under constraints, e.g., the maintenance frequency, authorizing periodic maintenance activities through Monte Carlo. Attempting to jointly optimize activities in single-stage manufacturing systems under fluctuating customer demand and recurrent failures, Polotski et al. (2019) considered a set of Hamilton-Jacobi-Bellman (HJB) equations solved by means of an algorithmic

approach. Towards the same goal, Xanthopoulos et al. (2018) developed a mechanism integrating reinforcement learning decision-making to optimize the manufactured product stock and backorder levels, while ensuring the high operability and robustness of manufacturing machines.

Moreover, there has been a research interest in optimizing the quality of output product stock within the context of manufacturing facilities using control charts, or reinforcement learning. For example, to enhance the profitability of a supply chain, Wu (2020) endeavored to improve the output product quality by assuming a make-to-stock manufacturing model optimized through dynamic programming. In addition to optimizing processes, e.g., production, Tasias (2022) focused on detecting deviations in the quality of manufactured items using a Bayesian control chart. Similarly, Hajej et al. (2021) considered a quality control model in order to identify low-quality items applying a dynamic sampling strategy to the completed product stock. Adopting a more dynamic approach, Paraschos et al. (2020) utilized a reinforcement learning framework to create joint strategies contributing to the improvement of the output quality, given the observed condition of manufactured items and manufacturing system. In addition to quality control, publications also endeavored to implement green manufacturing strategies in order to reuse manufactured material in manufacturing/remanufacturing systems. For example, Sarkar and Bhuniya (2022) formulated an inventory model for remanufacturing returned products aiming at achieving high profitability under green manufacturing constraints. Liu and Papier (2022) employed a heuristic approach for scheduling remanufacturing activities while assuming that both new and remanufactured items are identical.

Concluding this section, as the waste management remains a challenging area in process control, the authorization of activities and inspection of product quality are mostly ad-hoc-based in the pertinent literature. That is, publications devise control charts focused and implemented in specific cases set by designers. Clearly, one can consider that this approach is not an efficient or long-term solution, due to the growing complexity of products and the fluctuating customer demand. Therefore, a more general framework providing stock awareness to decision-makers should be devised. In this context, the present paper aims at providing such a solution that tackles issues mentioned above, e.g., material management. To this end, this solution is a process planning framework devising joint control policies related to product quality, system condition, and product stock generation. It pairs multi-agent reinforcement learning-based decision-making with ad-hoc manufacturing/maintenance control enabling sustainable, green, and lean manufacturing through material and operation management.

For clarity reasons, Table 1 compares the contribution of the present paper with the cited publications. Note, CA refers to computational algorithms (e.g., particle swarm optimization, genetic algorithm, etc.); RL and DP denote reinforcement learning and dynamic programming, respectively; MM, PC, and QC represent material management, process control, and quality control; RL/Ad-hoc control denotes the reinforcement learning combined with ad-hoc control policies (e.g., Base Stock, condition-based maintenance etc.).

Table 1 A table of comparison

Publication	Problem	Model	Optimization	Control policy
Bhosale and Pawar (2019)	PC	x	CA	Production
Gharbi et al. (2022)	PC	x	CA	Production
Hajej et al. (2021)	PC, QC	x	CA	Production, maintenance, quality sampling
Hoseinpour et al. (2021, 2020)	PC	x	CA	Production
Kim and Kim (2022)	PC	x	CA	Production
Li et al. (2022)	PC	x	CA	Maintenance
Liu and Papier (2022)	MM	x	CA	Remanufacturing
Manafzadeh Dizbin and Tan (2019)	PC	x	CA	Production
Metzker et al. (2023)	PC	x	DP	Production
Paraschos et al. (2020)	PC, QC		RL/Ad-hoc control	Production, maintenance, recycle
Polotski et al. (2019)	PC	x	CA	Production, maintenance
Sarkar and Bhuniya (2022)	MM	x	CA	Manufacturing, remanufacturing
Tasias (2022)	PC, QC	x	CA	Production, maintenance, inspections
Wu (2020)	QC	x	DP	Remanufacturing
Xanthopoulos et al. (2018)	PC		RL	Production, maintenance
Present paper	PC, QC, MM		RL/Ad-hoc control	Production, maintenance, remanufacturing, recycling

3 Three-stage manufacturing/remanufacturing system

The system implements a multi-stage serial manufacturing line. This line involves three machines and storehouses, producing and stockpiling a single type of items. The architecture of the described line is graphically presented in Fig. 1. According to this figure, the machines of the first two stages generate incomplete versions of the final goods stored in their corresponding storage. These items are received and processed by the 3rd machine to create a stock of final goods. Every authorized production activity is associated with a cost L^{π}.

However, the quality of goods and the effectiveness of the manufacturing line are substantially degraded by recurring failures correlated with the persistent system operation. Let $\omega = [0, 1, \ldots, d]$ and d denote the manufacturing/remanufacturing system condition and the deterioration stages, respectively. After the occurrence of a failure, the system is transitioned from condition ω into the next one denoted as $\omega + 1$. Due to this transition, the system quickly becomes substantially deteriorated. To avoid any malfunction, the production machines are frequently maintained with a cost L^{σ_ω}. One can assume that $L^{\sigma_1} < L^{\sigma_2} < \cdots < L^{\sigma_d}$. Furthermore, when the machines are deemed inoperable, their operability is restored by authorizing repair operations associated with a corresponding cost L^{ϵ}.

In terms of output product quality, items generated by the machines are classified into three categories relevant to their present quality: top, second-rate, and flawed goods. Due to the high standards set by customers, only the top and second-rate goods are procured by customers with R^t and R^m, respectively. The unsold second-rate goods are recycled and a profit R^b is obtained by the system. Contrariwise, as the flawed products remain unsold, the system authorizes remanufacturing operations in order to make them salable. The cost of remanufacturing operations is L^{ρ}. The remanufactured items are sold for R^r. Lastly, if an item did not satisfy the requirements of a customer, it is returned to the manufacturing/remanufacturing system. The returning fee is equivalent to A^p.

4 Multi-agent reinforcement learning optimization framework

4.1 Overview

This section presents the functionality of the framework proposed for process planning in degrading multi-stage manufacturing/remanufacturing system, endeavoring to reduce operational costs correlated with the frequency of activities, e.g., manufacturing,

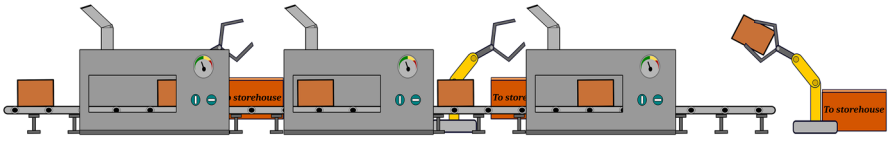

Fig. 1 Three-stage production line

and the size of final product stock. Figure 2 depicts the architecture of the proposed implementation. The mechanism communicates with manufacturing machines involved in the examined manufacturing line and decides on an effective course of action, contributing to the profit maximization. This communication is occurred at specific time intervals. Let τ^e denote these intervals. The decision-making process is conducted by two agents integrated into the 1st and 3rd stage of the production process. At every τ^e, prior to decision-making, the agents obtain details regarding the state of their corresponding manufacturing machines. The state of each machine is considered the input of the proposed implementation and defined by parameters related to degradation, status, top-quality item stock, and faulty item stock. Given these parameters, the agents aim to additively improve the total profitability of the system during their interaction with the machines, utilizing reinforcement learning decision-making complemented with ad-hoc manufacturing/maintenance control. To this end, they formulate process control strategies that authorize a variety of activities, e.g., manufacturing and maintenance. This strategy represents the output of the proposed optimization framework. A detailed presentation of the described functionality is given in the following subsections.

4.2 Formulating state

During their communication with the machines, the agents receive a representation of the current machine state defined mathematically as a vector. This vector contains four variables describing the behavior of the machines, that is, degradation, status, top product stock, and flawed item stock. In this respect, the current machine state vector is:

$$M = (\theta, \kappa, s_a, s_f) \tag{1}$$

where θ is the machine degradation, κ represents the machine status, s_a and s_f are the top product stock and flawed item stock. Note, $\theta \in [0, \dots, \omega]$, $s_a, s_f \in [0, \dots, P_{max}]$, where P_{max} denotes the maximum capacity of the machine's product storage, and

$$\kappa = \begin{cases} 0, \text{ unavailable} \\ 1, \text{ in service} \\ 2, \text{ do nothing} \\ 3, \text{ maintenance mode} \end{cases}.$$

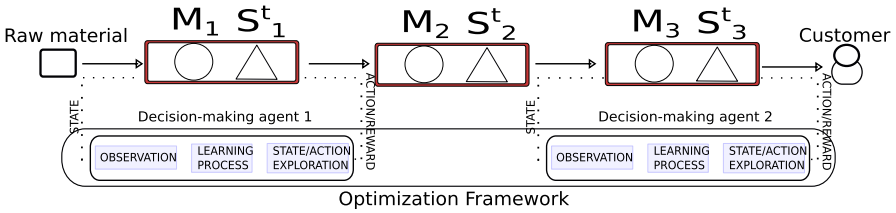

Fig. 2 The proposed implementation

4.3 Defining action-set

As described in Sect. 3, the integrated machines are frequently involved in manufacturing, maintenance, remanufacturing, and recycling activities. These activities are initiated by both agents according to the present state of the machines. The action selection process carried out by the agents is depicted in Fig. 3.

According to Fig. 3, the functionality of the decision-making agents can be described as follows. To ensure the high output product quality, the production process occurs in the manufacturing machines when the latter are in a relative perfect condition ($\theta = 0$). Reaching the maximum capacity of storage with the stockpiling of either top or flawed products ($s_a = P_{max} \parallel s_f = P_{max}$), the machines enter into maintenance mode in order to be restored in a prior condition. However, in the case of having only flawed products in storage ($s_f = P_{max}$), the agents decide on initiating remanufacturing or recycling operations to minimize the stock of low-rate/flawed products, and thus create an additional revenue stream for the system in question.

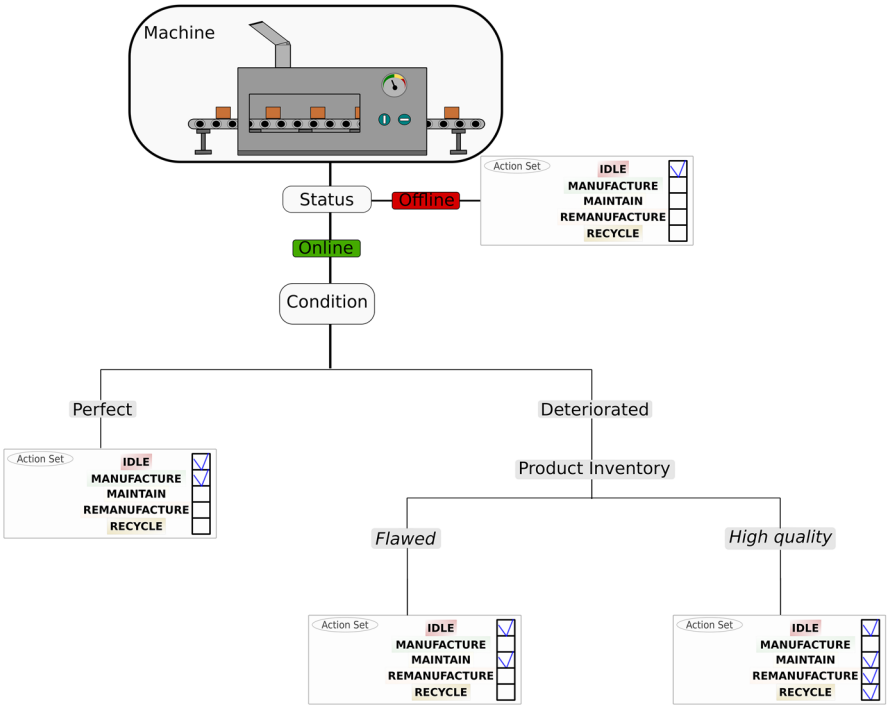

Fig. 3 The action selection process

4.4 Integrating control policies

In the context of the examined system, the integrated parametric control concerns the planning of manufacturing and maintenance operations. The aim of this integration is the control over the occurrence of the aforesaid operations in an attempt to decrease the associated costs. In this respect, six policies are utilized, namely, condition-based maintenance, periodic maintenance, opportunistic maintenance, constant work-in-progress, base stock, and extended KANBAN. Throughout this section, these policies are described in length.

4.4.1 CONWIP

Constant work-in-progress (Xanthopoulos and Koulouriotis 2014), often called CONWIP, is a manufacturing control policy that aims to restrict the amount of work-in-progress goods throughout the production line. The policy implements a similar functionality to the one of the KANBAN policy when applied to a single-machine system. When a final item leaves the storehouse, the manufacturing process is initiated in the 1st stage. All the subsequent machines continually produce new material. Thus, one can define a control parameter K^c, which equals the sum of the maximum capacity of work-in-progress and final items.

4.4.2 Condition-based maintenance

According to condition-based maintenance control policy, a machine is engaged in a maintenance activity when it finds itself in a deteriorated condition. To this end, a threshold c_{thr} is defined referring to the preferred system condition, in which the system is being maintained. Formally, after its condition reaches c_{thr}, the machine receives maintenance. This functionality is illustrated in Fig. 4.

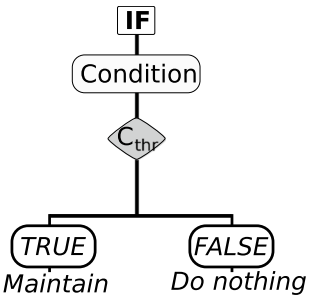

Fig. 4 Condition-based maintenance

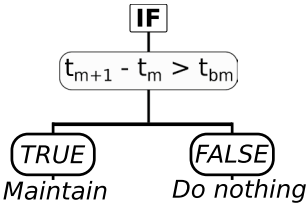

Fig. 5 Periodic maintenance

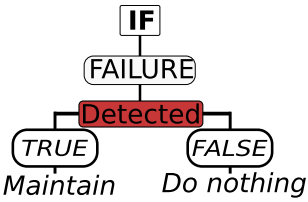

Fig. 6 Opportunistic maintenance

4.4.3 Extended KANBAN

Extended KANBAN is a pull mechanism combining both base stock and KANBAN concepts (Paraschos et al. 2022). In this respect, the mechanism utilizes two control parameters related to the number of KANBAN cards and customer orders. The received orders are transferred to every machine integrated in the production line. The production process in the system is initiated only when a final item is acquired by a customer and the respective KANBAN card is transmitted to machines implemented in prior stages.

4.4.4 Periodic maintenance

Periodic maintenance is periodically conducted in the system at specified time intervals. Let $t_m = [t_m^1, t_m^2, \dots]$ represent a series of time intervals, where the machine is maintained, and t_{bm} denote the time between occurring maintenance activities. Following the paradigm of the periodic maintenance policy, the system is maintained according to Fig. 5.

4.4.5 Base stock

Base stock is a manufacturing control policy generating new items in respect to the present demand for items (Duri et al. 2000). That is, after the placement of an order, the production process begins in the integrated machine. Once an item is finished, it is obtained by the customer, and thus fulfilling the placement order.

4.4.6 Opportunistic maintenance

To extend the lifespan of a machine, opportunistic maintenance repairs deteriorated equipment, no matter if it is malfunctioning or not. This strategy aims to extend the

reliability of the machine by minimizing the likelihood of failure. Figure 6 depicts the logic of conducting opportunistic maintenance in a manufacturing system.

4.5 Defining objective

After opting for an action at τ^e, the 1st and 3rd manufacturing machines receive a reward in the next time interval, that is, $\tau^e + 1$. This reward represents the total revenue of the machines. It is equal to the profits obtained by offering the completed goods to the customers minus the costs related to integrated activities, such as repair and product returns. Formally, the received reward is formulated as follows:

$$u = R^t + R^m + R^r + R^b - L^\pi - L^{\sigma_\omega} - L^\epsilon - L^\rho - A^p \tag{2}$$

Given the expression above, it is sensible that the reward of the 1rd machine would involve only costs while the one obtained the 3rd machine would involve both profits and costs, as the customers obtain the completed items from that stage. In regard to the notation of expression 2, R^t and R^m denote the revenues obtained by selling top-quality and second-rate products; R^r and R^b represent the recycled product and remanufactured item revenues, respectively; L^π and L^{σ_ω} correspond to the production and maintenance costs; L^ϵ and L^ρ are the costs relevant to repair and remanufacturing activities; A^p is the fee of returning products to the examined manufacturing system.

The objective of the agents is to increase the profitability of the manufacturing line by deriving an optimal joint control policy π_c. It is assumed by the agents using the following expression:

$$\overline{u} = \lim_{x \to \infty} \frac{1}{i} \sum_{\omega=1}^{i} E[u] \tag{3}$$

where $E[u]$ refers to the expected value of the total revenues, estimated by expression 2.

4.6 Learning the defined objective

In order to fulfill the objective, the agents use reinforcement learning algorithms to determine joint control policies. In this paper, two model-free average reward algorithms are integrated into the mechanism: R-Learning (Schwartz 1993) and R-Smart (Gosavi 2004). Implementing both algorithms, the agents endeavor to incrementally maximize the profitability through the approximation of q-values and average rewards. Mathematically, these approximations are conducted according to the following expressions:

$$V_q(M, \Delta) = V_q(M, \Delta) + \gamma[u - \overline{u} + V_q(M', \Delta') - V_q(M, \Delta)] \tag{4}$$

$$V_r(M,\Delta) = (1-\delta)V_r + \epsilon\overline{u} \tag{5}$$

$$V_r(M,\Delta) = V_r + \epsilon[u - \overline{u} + V_q(M',\Delta') - V_q(M,\Delta)] \tag{6}$$

In the expressions 4–6, the employed notation can be defined as follows. M and Δ refer to the machine state and the opted activity at present decision-making time-interval (τ^e); M' and Δ' represent the machine state and the authorized activity at the next decision-making time-interval ($\tau^e + 1$); u and $\overline{u}$ are the reward and the expect reward estimated using expressions 2 and 3, respectively; δ, ϵ, and γ are real-valued hyper-parameters, i.e., parameters that control the learning process, as defined by expressions 4–6.

Concerning the implemented algorithms, expressions 4 and 6 are used under R-Learning, while R-Smart-based decision-making utilizes expressions 4 and 5. Analytically, expressions 4 estimates the q-values and stores them in a table. The values in their corresponding table are updated when the agents decide on the action associated with the highest q-value (V_q). Contrariwise, expressions 5 and 6 calculate the average reward received by each agent. It is calculated in all decision-making time intervals. Finally, the exploration of the action-state space is conducted according to the e-greedy paradigm.

5 Experiments

5.1 Setup

In this section, the multi-agent optimization framework is evaluated in 36 experiments. These experiments simulate real-world situations commonly met in manufacturing industry. To elaborate on, the operation of a production line is frequently affected by fluctuations in the recurrence of activities and machine degradation. To

Table 2 A sample of mean activity rates utilized in simulation experiments

Mean rates					
Arrival	Manufacturing	Failure	Maintenance	Repair	Remanufacturing
1.25	5.82	8.23	9.24	27.93	2.50
1.25	3.98	7.47	9.24	27.93	4.41
4.53	3.98	7.47	9.24	40.17	4.41
4.53	5.82	8.23	9.24	40.17	6.32
4.53	5.82	8.23	9.24	40.17	2.50
4.53	2.14	6.86	12.80	34.05	2.50
4.53	3.98	7.47	12.80	34.05	4.41
4.53	5.82	8.23	12.80	34.05	6.32
7.34	5.82	8.23	16.35	34.05	2.50

this end, the conducted simulations are defined by mean rates associated with order arrival, machine failures, maintenance, repair, manufacturing and remanufacturing. Table 2 provides a sample of these rates. To better illustrate the concept of the conducted experiments, let us describe the experimental scenario listed in the 3rd row of Table 2. According to this scenario, the three-stage manufacturing system would receive a moderate number of customer orders and authorize scarcely manufacturing/remanufacturing activities. Its condition and produced goods would be subject to constant degradation due to the high frequency of failure activities and the scarcity of maintenance activities. However, repair activities would be frequently authorized to restore the system in order to start the production of new items for the customers.

In the prior section, it was stated that two reinforcement learning algorithms and six control policies combined are integrated in the mechanism. The main goal is to test and evaluate the concept of controlling the recurrence of activities and the quality of output stock in the context of a three-stage production line. The implemented versions of mechanism are listed in Table 3. Every mechanism was evaluated under every experimental scenario and completed 5.50 million items in the 3rd storage facility. Each simulation experiment was replicated 10 times. Concerning the hyper-pameters of both reinforcement learning algorithms, γ, δ, and ϵ were equal to 0.005. As for the control parameters of ConMain and PerMain, c_{thr} and t_{bm} are equivalent to 1.0 and 10.0, respectively. Finally, the proposed mechanism, along with the manufacturing system simulator, were coded in C++.

5.2 Profitability

Figure 7 depicts the performance of proposed optimization framework's iterations integrated in the production line. It is illustrated that the most profitable versions are the ones integrating R-Smart algorithm. This observation illustrates the efficiency of R-Smart against R-Learning in terms of learning capability under fluctuating conditions, such as recurrent order arrival. Concerning the production control, the production line achieves an equilibrium between gains and losses implementing extended KANBAN and constant work-in-process control policies. Evidently, this shows that

Table 3 Utilized framework versions

Mechanisms	Abbreviation
Extended Kanban-Condition-based Maintenance	ExKan-ConMain
Extended Kanban-Periodic Maintenance	ExKan-PerMain
Extended Kanban-Opportunistic Maintenance	ExKan-OppMain
CONWIP-Condition-based Maintenance	ConW-ConMain
CONWIP-Periodic Maintenance	ConW-PerMain
CONWIP-Opportunistic Maintenance	ConW-OppMain
Base Stock-Condition-based Maintenance	BasStoc-ConMain
Base Stock-Periodic Maintenance	BasStoc-PerMain
Base Stock-Opportunistic Maintenance	BasStoc-OppMain

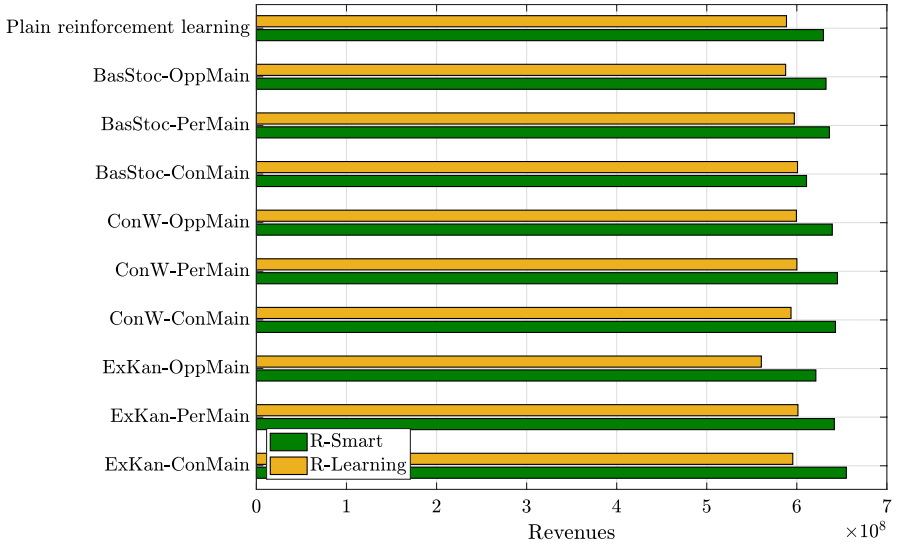

Fig. 7 The revenue stream of the examined three-stage production line

the implementation of KANBAN-like systems improves the awareness of the production system by reducing the uncontrollable authorization of production activities and providing an enhanced control over the final product stock. In addition, the majority of the policy-integrated mechanisms outperform the ones implementing plain R-Smart and R-Learning. This likely suggests that the plain reinforcement

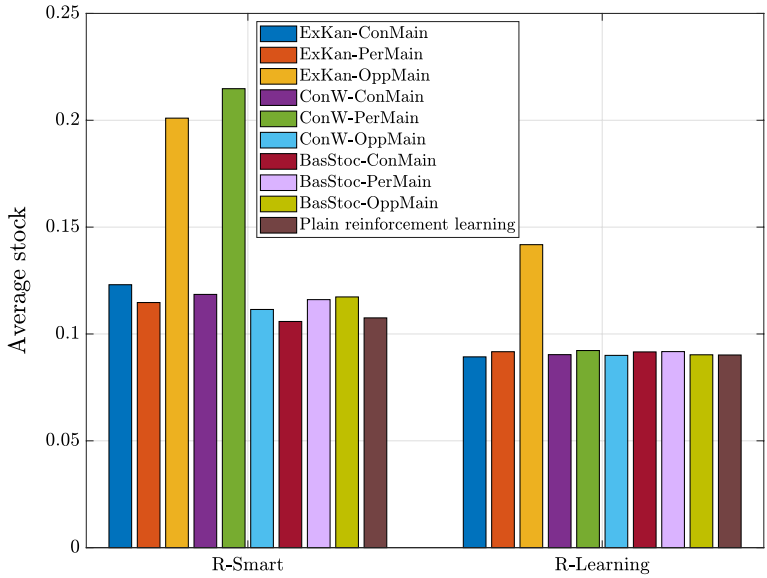

Fig. 8 Average top product stock

learning versions authorize a high number of activities being unable to cover the increasing operational costs by procuring only revenues. However, an interesting note can be made for ExKan-OppMain (R-Learning), which procures lesser revenues than its counterparts. This observation can be explained as follows. Though extended KANBAN is cost-efficient, the authorization of opportunistic maintenance minimizes the revenues. This suggests that the opportunistic maintenance activities are recurrent in the system, as they repair equipment when their deterioration is detected. This suggestion is supported by the stockpiling of top-quality products in Fig. 8. Therefore, the implementation of such activities cannot be considered as a cost-effective solution for the production line in question.

5.3 Product quality

Figure 8 shows the dissemination of the top-quality product stock across the evaluated iterations of the presented approach. In terms of employed learning algorithm, the R-Smart mechanisms ensure higher product quality than the ones manufactured by the R-Learning versions. This suggests that the learning process of the R-Smart mechanisms is substantially enhanced and supported by the usage of a different reward equation, that is, Eq. 5. Regarding the production process, the ExKan-OppMain and ConW-PerMain generate a high amount of products. It is likely that the combination of the KANBAN-like mechanisms with periodic and opportunistic

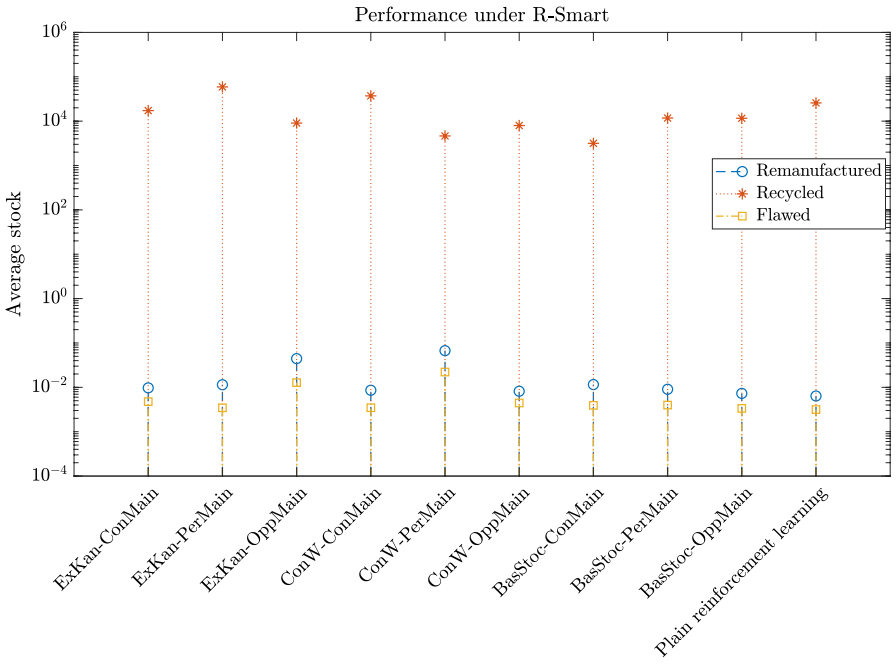

Fig. 9 Dissemination of average reformed and defective product stock under R-Smart

maintenance facilitates the versatility of the production line in circumstances associated with high product demand and persistent degradation in quality. That is, by the authorization of both periodic and opportunistic maintenance activities, equipment is maintained either periodically or at the moment the failure is identified. Clearly, this maintenance strategy minimizes the likelihood of system downtime and therefore improves the revenue stream of the production system.

Figure 9 presents the stock consisted of recycled, remanufactured and flawed products when the proposed mechanisms integrate R-Smart-based learning process. In this respect, this illustration compares the performance of the implemented mechanisms in terms of green and sustainable manufacturing. Given the implemented remanufacturing and recycling activities, the figure illustrates that the recycled products are more than the remanufactured ones. This observation suggests that the agents prefer recycling activities over remanufacturing ones. It can be explained as follows. Due to the customer preference over top-quality good, the second-rate items remain unsold and stockpiled in the 3rd storage facility. To minimize their number, the agents opt to authorize frequent recycling activities in order to create space for the generation of salable products, such as the top-quality ones. In terms of remanufactured material, a significant remark can be made for ConW-PerMain and ExKan-OppMain. As illustrated in Fig. 8, both iterations achieved at ensuring the top-quality of manufactured goods, while turning faulty material into salable products through remanufacturing activities. Though their performance is inferior to that displayed by ExKan-ConMain, one can deduce that the combination of KANBAN-like

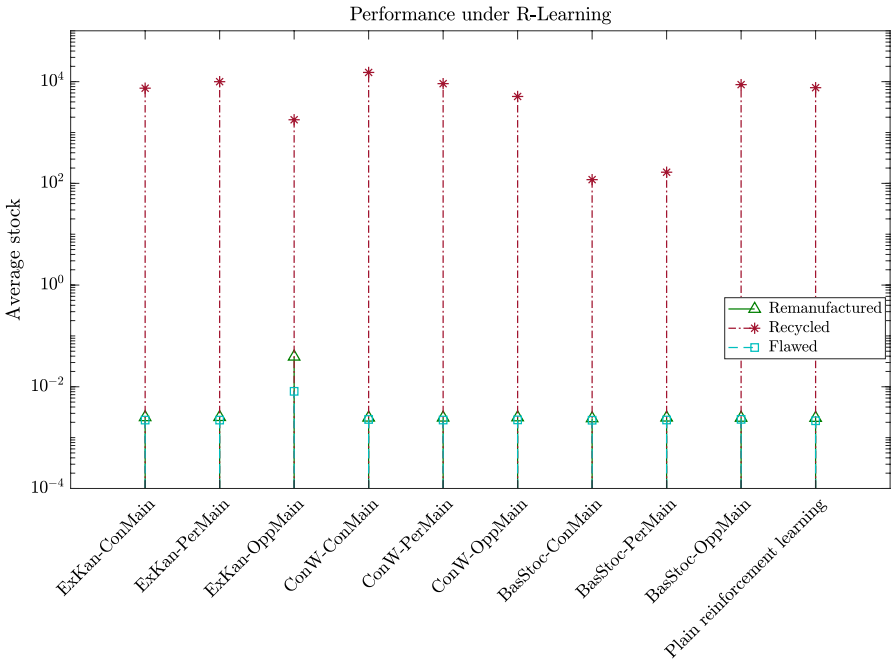

Fig. 10 Dissemination of reformed and defective product under R-Learning

policies with periodic or opportunistic maintenance control efficiently contributes to the stock-awareness of the system by improving the quality of both the production process and manufactured items.

Figure 10 demonstrates the fluctuation in the available recycled, remanufactured and flawed product stock when the agents employ R-Learning. Similar to Fig. 9, this figure shows that the agents authorize more recycling activities than remanufacturing ones. It is likely that the mechanisms under R-Learning completed a high number of second-rate items. This is partially supported by Fig. 8, which shows that the R-Learning-based mechanisms stockpiled a few top-quality products. It can be attributed to the constant degradation of the manufacturing mechanism due to occurred failures. This results in degrading the quality of the output inventory as well. Therefore, the mechanisms decided to remove the completed second-rate products from the output inventory by recycled them in an effort to obtain additional revenue. In regard to the performance of the mechanisms, it is illustrated that ConW-ConMain is efficient in minimizing the second-rate items adopting a "green" activity, that is, recycling. Concerning the amount of remanufactured items, the best performing mechanism is ExKan-OppMain. That is, the mechanism remanufactures a high amount of flawed products and thus creates mid-quality ones. Along with top-quality products, these can be sold to customer fulfilling pending orders placed in the manufacturing system. In this direction, extended KANBAN manages to rapidly respond to the fluctuating demand authorizing new production activities throughout the production line. As illustrated in Fig. 11, the agents are keen to maintain the system recurrently

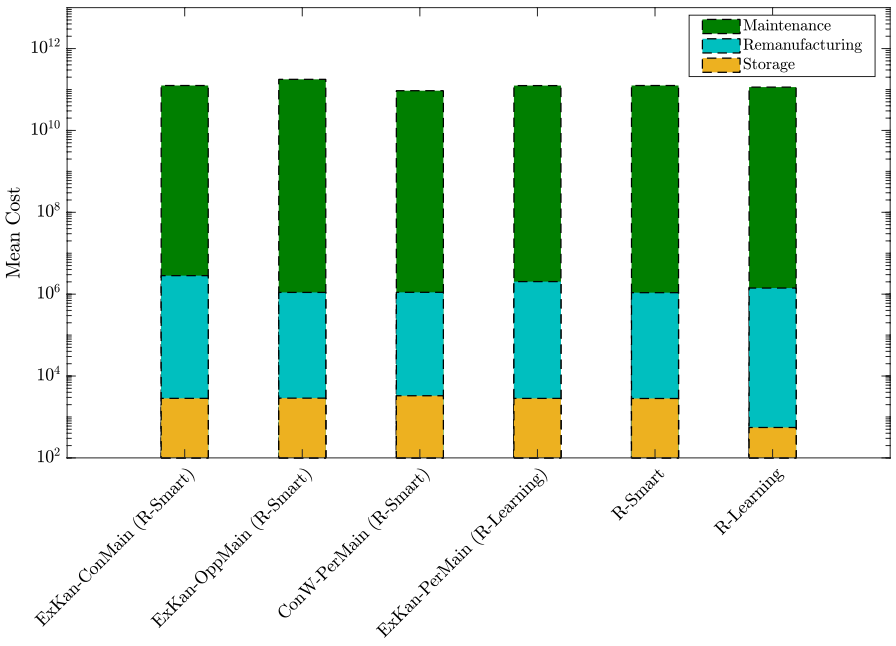

Fig. 11 Dissemination of costs

by replacing, or repairing the malfunctioning equipment through opportunistic maintenance. Furthermore, it is worth noting that the stock of R-Learning-based frameworks, except ExKan-OppMain, is equally divided between remanufactured and flawed products. One possible explanation for this observation is the top-quality stock. In this regard, the inventory contains mostly top products preventing the stockpiling of low-quality items manufactured by the machines. This behavior shows the benefit of integrating a reinforcement learning-based decision-making process in manufacturing/remanufacturing systems contributing to sustainability and material waste minimization.

5.4 Received costs

Figure 11 depicts costs procured by proposed frameworks. For this depiction, ExKan-ConMain (R-Smart), ExKan-OppMain (R-Smart), ConW-PerMain (R-Learning), plain R-Smart and R-Learning frameworks were selected based on their performance in terms of attained product quality. It is illustrated that the opportunistic maintenance-integrated framework procures the highest maintenance cost. This can be attributed to the nature of opportunistic maintenance. In this context, the agents make decisions on maintaining recurrently the machines on the basis of observed failures or malfunctioning equipment. This approach allows the system to be preemptively maintain machines avoiding redundant down-times and degradation in product quality. On the other hand, ConW-PerMain (R-Smart) obtains lesser maintenance cost due to the periodic nature of the authorized maintenance activities. In this respect, it endeavored to maintain the machines with minimum set of activities. Furthermore, regarding the remanufacturing activities, it is indicated that ExKan-ConMain (R-Smart) initiated a high number of such activities compared to its counterparts. This evidently suggests that the agents under the described variation effectively manage the waste produced by the machines due to their degradation. Lastly, according to the storage cost illustrated in Fig. 11, it is suggested that plain R-Learning mechanism stores a decreased amount of products when compared against to the other implemented frameworks. It is apparent that the machines under this iteration is not involved in recurrent manufacturing activities. This supports the reason why the R-Learning-integrated iterations are not as effective as the R-Smart-implemented ones.

6 Conclusion

This paper examines failure-prone machines integrated in a multi-stage production line, processing one type of products. To optimize the behavior of such systems, a reinforcement learning-based framework is proposed. For the decision-making process, this framework implements two agents in the specific stages of production process for planning several activities, e.g., production and remanufacturing. Furthermore, to complement the reinforcement learning-based decision-making process, ad-hoc control policies related to production and maintenance are employed. The

goal of this integration was to improve the waste management of the system and minimize the number of redundant activity authorizations. Finally, the functionality of the proposed approach was investigated and analyzed within the context of simulation experiments. These experiments suggested that the revenue stream of the manufacturing/remanufacturing system was mostly based on recycled and remanufactured products, despite selling top-quality goods. In this respect, one can deduce that the presented reinforcement learning/ad-hoc control mechanism devises a successful green manufacturing strategy reusing low-quality material and minimizing the production of new material.

The illustrated results present implications regarding the applicability of the presented optimization framework in real-world applications. First, the framework could be efficient into the context of complex manufacturing environments. This argument could be supported by its performance in the examined multi-stage manufacturing system. It is shown that the framework is efficient in detecting and reforming low-quality material in the system. This is achieved by the green strategies, namely recycling and remanufacturing, which attempt to improve the cost-effectiveness of the system reusing already generated material. Clearly, an environment-friendly manufacturing environment could be formulated given the framework's efficiency in material management. In terms of implementation, the proposed framework could be easily implemented to a variety of manufacturing systems with slight modifications due to the nature of the reinforcement learning/ad-hoc policy-based decision-making. In this regard, reinforcement learning does not require an explicit and detailed manufacturing system model compared to other optimization methods, such as dynamic programming (Sutton and Barto 2018). Furthermore, ad-hoc control policies, e.g., Base Stock and opportunistic maintenance, are frequently applied in the real-world manufacturing industry, since they can be easily configured due to the parametric nature. Given the above, the proposed optimization framework is an intelligent system that could improve the productivity and sustainability of the real-world manufacturing environments by efficiently managing material with green practices.

In the future, the optimization framework could be implemented in complex and large supply chains involving multiple products. In this context, a detailed inventory model should be formulated as well. This would be utilized for the optimization of inventory control in supply chains. For the optimization process, the decision-making process carried out by the presented framework's agents could be augmented through metaheuristic techniques, e.g., simulated annealing. Furthermore, the present paper studied a production line consisted of three serial machines. In this respect, a future work could tackle the parallel manufacturing control problem implementing a modified iteration of the methodology presented in this paper. Such an iteration would produce interesting results and provide substantial implications concerning the implementation of flexible scheduling in parallel machines.

Funding Open access funding provided by HEAL-Link Greece.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Adeinat H, Pazhani S, Mendoza A et al (2022) Coordination of pricing and inventory replenishment decisions in a supply chain with multiple geographically dispersed retailers. Int J Prod Econ 248(108):461. https://doi.org/10.1016/J.IJPE.2022.108461

Ahmad T, Madonski R, Zhang D et al (2022) Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: key developments, challenges, and future research opportunities in the context of smart grid paradigm. Renew Sustain Energy Rev 160(112):128. https://doi.org/10.1016/j.rser.2022.112128

Antons O, Arlinghaus JC (2022) Data-driven and autonomous manufacturing control in cyber-physical production systems. Comput Ind 141(103):711. https://doi.org/10.1016/J.COMPIND.2022.103711

Beraudy S, Absi N, Dauzère-Pérès S (2022) Timed route approaches for large multi-product multi-step capacitated production planning problems. Eur J Oper Res 300(2):602–614. https://doi.org/10.1016/J.EJOR.2021.08.011

Bhosale KC, Pawar PJ (2019) Material flow optimisation of production planning and scheduling problem in flexible manufacturing system by real coded genetic algorithm (RCGA). Flex Serv Manuf J 31(2):381–423. https://doi.org/10.1007/s10696-018-9310-5

Corallo A, Crespino AM, Lazoi M et al (2022) Model-based Big Data Analytics-as-a-Service framework in smart manufacturing: a case study. Robot Comput-Integr Manuf 76(102):331. https://doi.org/10.1016/j.rcim.2022.102331

Duri C, Frein Y, Di Mascolo M (2000) Comparison among three pull control policies: Kanban, base stock, and generalized Kanban. Ann Oper Res 93(1–4):41–69. https://doi.org/10.1023/a:1018919806139

Gharbi A, Kenné JP, Kaddachi R (2022) Dynamic optimal control and simulation for unreliable manufacturing systems under perishable product and shelf life variability. Int J Prod Econ 247(108):417. https://doi.org/10.1016/j.ijpe.2022.108417

Gosavi A (2004) A reinforcement learning algorithm based on policy iteration for average reward: empirical results with yield management and convergence analysis. Mach Learn 55(1):5–29. https://doi.org/10.1023/B:MACH.0000019802.64038.6c

Hajej Z, Rezg N, Gharbi A (2021) Joint production preventive maintenance and dynamic inspection for a degrading manufacturing system. Int J Adv Manuf Technol 112(1–2):221–239. https://doi.org/10.1007/s00170-020-06325-3

He Y, Li Y, Wu T et al (2015) An energy-responsive optimization method for machine tool selection and operation sequence in flexible machining job shops. J Clean Prod 87(C):245–254. https://doi.org/10.1016/J.JCLEPRO.2014.10.006

Hoseinpour Z, Kheirkhah AS, Fattahi P et al (2020) The problem solving of bi-objective hybrid production with the possibility of production outsourcing through meta- heuristic algorithms. Management 4:1–17. https://doi.org/10.31058/j.mana.2021.42001

Hoseinpour Z, Taghipour M, Beigi JH et al (2021) The problem solving of bi-objective hybrid production with the possibility of production outsourcing through imperialist algorithm, NSGA-II, GAPSO hybrid algorithms. Turk J Comput Math Educ TURCOMAT 12(13):8090–8111

Iqbal N, Khan AN, Rizwan A et al (2022) Enhanced time-constraint aware tasks scheduling mechanism based on predictive optimization for efficient load balancing in smart manufacturing. J Manuf Syst 64:19–39. https://doi.org/10.1016/J.JMSY.2022.05.015

Jeong YS (2022) Secure IIoT information reinforcement model based on IIoT information platform using blockchain. Sensors 22(12):4645. https://doi.org/10.3390/s22124645

Jum'a L, Zimon D, Ikram M et al (2022) Towards a sustainability paradigm; the nexus between lean green practices, sustainability-oriented innovation and Triple Bottom Line. Int J Prod Econ 245(108):393. https://doi.org/10.1016/J.IJPE.2021.108393

Karnik N, Bora U, Bhadri K et al (2022) A comprehensive study on current and future trends towards the characteristics and enablers of industry 4.0. J Ind Inf Integr 27(100):294. https://doi.org/10.1016/J.JII.2021.100294

Kenett RS, Bortman J (2022) The digital twin in Industry 4.0: a wide-angle perspective. Qual Reliab Eng Int 38(3):1357–1366. https://doi.org/10.1002/qre.2948

Kim H, Kim E (2022) A hybrid manufacturing system with demand for intermediate goods and controllable make-to-stock production rate. Eur J Oper Res 303(3):1244–1257. https://doi.org/10.1016/j.ejor.2022.03.039

Li X, Ran Y, Wan F et al (2022) Condition-based maintenance strategy optimization of meta-action unit considering imperfect preventive maintenance based on Wiener process. Flex Serv Manuf J 34(1):204–233. https://doi.org/10.1007/s10696-021-09407-w

Lim MK, Lai M, Wang C et al (2022) Circular economy to ensure production operational sustainability: a green-lean approach. Sustain Prod Consum 30:130–144. https://doi.org/10.1016/J.SPC.2021.12.001

Liu B, Papier F (2022) Remanufacturing of multi-component systems with product substitution. Eur J Oper Res 301(3):896–911. https://doi.org/10.1016/j.ejor.2021.11.029

Manafzadeh Dizbin N, Tan B (2019) Modelling and analysis of the impact of correlated inter-event data on production control using Markovian arrival processes. Flex Serv Manuf J 31(4):1042–1076. https://doi.org/10.1007/S10696-018-9329-7/TABLES/16

Metzker P, Thevenin S, Adulyasak Y et al (2023) Robust optimization for lot-sizing problems under yield uncertainty. Comput Oper Res 149(106):025. https://doi.org/10.1016/j.cor.2022.106025

Paraschos PD, Koulinas GK, Koulouriotis DE (2020) Reinforcement learning for combined production-maintenance and quality control of a manufacturing system with deterioration failures. J Manuf Syst 56:470–483. https://doi.org/10.1016/j.jmsy.2020.07.004

Paraschos PD, Koulinas GK, Koulouriotis DE (2021) Parametric and reinforcement learning control for degrading multi-stage systems. Procedia Manuf 55:401–408. https://doi.org/10.1016/j.promfg.2021.10.055

Paraschos PD, Xanthopoulos AS, Koulinas GK et al (2022) Machine learning integrated design and operation management for resilient circular manufacturing systems. Comput Ind Eng 167(107):971. https://doi.org/10.1016/j.cie.2022.107971

Polotski V, Kenne JP, Gharbi A (2019) Optimal production and corrective maintenance in a failure-prone manufacturing system under variable demand. Flex Serv Manuf J 31(4):894–925. https://doi.org/10.1007/s10696-019-09337-8

Sarkar B, Bhuniya S (2022) A sustainable flexible manufacturing–remanufacturing model with improved service and green investment under variable demand. Expert Syst Appl 202(117):154. https://doi.org/10.1016/j.eswa.2022.117154

Schwartz A (1993) A Reinforcement Learning Method for Maximizing Undiscounted Rewards. In: Proc Tenth Int Conf Mach Learn ICML93, pp 298–305, https://doi.org/10.1016/B978-1-55860-307-3.50045-9

Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. The MIT Press, Cambridge, MA

Tasias KA (2022) Integrated quality, maintenance and production model for multivariate processes: a Bayesian approach. J Manuf Syst 63:35–51. https://doi.org/10.1016/J.JMSY.2022.02.008

Tu J, Zhang L (2022) Performance analysis and optimisation of Bernoulli serial production lines with dynamic real-time bottleneck identification and mitigation. Int J Prod Res. https://doi.org/10.1080/00207543.2021.2019343

Wu CH (2020) Production-quality policy for a make-from-stock remanufacturing system. Flex Serv Manuf J 33(2):425–456. https://doi.org/10.1007/S10696-020-09379-3

Xanthopoulos AS, Koulouriotis DE (2014) Multi-objective optimization of production control mechanisms for multi-stage serial manufacturing-inventory systems. Int J Adv Manuf Technol 74(9):1507–1519. https://doi.org/10.1007/S00170-014-6052-8

Xanthopoulos AS, Kiatipis A, Koulouriotis DE et al (2018) Reinforcement learning-based and parametric production-maintenance control policies for a deteriorating manufacturing system. IEEE Access 6:576–588. https://doi.org/10.1109/ACCESS.2017.2771827

Ye Z, Yang H, Cai Z et al (2021) Performance evaluation of serial-parallel manufacturing systems based on the impact of heterogeneous feedstocks on machine degradation. Reliab Eng Syst Saf 207(107):319. https://doi.org/10.1016/J.RESS.2020.107319

Zhang X, Chen L (2022) A general variable neighborhood search algorithm for a parallel-machine scheduling problem considering machine health conditions and preventive maintenance. Comput Oper Res 143(105):738. https://doi.org/10.1016/J.COR.2022.105738

Panagiotis D. Paraschos received the Diploma degree in Production and Management Engineering from Democritus University of Thrace, Xanthi, Greece. He is currently working towards a Ph.D. degree in Production and Management Engineering at Democritus University of Thrace, Xanthi, Greece. His research interests include computational intelligence, game artificial intelligence, machine learning, deep learning, dynamic difficulty adjustment, affective computing procedural content generation, and intelligent control systems.

Georgios K. Koulinas received the Ph.D. degree in Production and Management Engineering from the Democritus University of Thrace, Greece. He is currently an Assistant Professor with the Department of Production and Management Engineering, Democritus University of Thrace, Greece and an Adjunct Lecturer with the Hellenic Open University, Greece. He has authored several scientific articles. His research interests include hyper-heuristic algorithms, metaheuristics, fuzzy systems, machine learning, risk assessment, project scheduling, and safety systems.

Dimitrios E. Koulouriotis received the diploma (combined B. Sc. And M.Sc.) degree in Electrical and Computer Engineering from the Democritus University of Thrace, Greece, the M.Sc. degree in Electronic and Computer Engineering and the Ph.D. in Production and Management Engineering from the Technical University of Crete, Greece. Currently, he is a Full Professor and Chairman in the Department of Production and Management Engineering and Director of the Industrial Production Laboratory and Ergonomics and Safety Laboratory at Democritus University of Thrace, Greece. He is the author of three books and numerous articles. His research interests include intelligent systems, machine learning, knowledge management, industrial and management engineering, machine vision and signal processing, system safety and business intelligence.