Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-024-02472-6

Managing product-inherent constraints with artificial intelligence: production control for time constraints in semiconductor manufacturing

Marvin Carl May[1] · **Jan Oberst**[1] · **Gisela Lanza**[1]

Received: 17 May 2023 / Accepted: 19 July 2024

Abstract
Continuous product individualization and customization led to the advent of lot size one in production and ultimately to product-inherent uniqueness. As complexities in individualization and processes grow, production systems need to adapt to unique, product-inherent constraints by advancing production control beyond predictive, rigid schedules. While complex processes, production systems and production constraints are not a novelty per se, modern production control approaches fall short of simultaneously regarding the flexibility of complex job shops and product unique constraints imposed on production control. To close this gap, this paper develops a novel, data driven, artificial intelligence based production control approach for complex job shops. For this purpose, product-inherent constraints are resolved by restricting the solution space of the production control according to a prediction based decision model. The approach validation is performed in a real semiconductor fab as a job shop that includes transitional time constraints as product-inherent constraints. Not violating these time constraints is essential to avoid scrap and similarly increase quality-based yield. To that end, transition times are forecasted and the adherence to these product-inherent constraints is evaluated based on one-sided prediction intervals and point estimators. The inclusion of product-inherent constraints leads to significant adherence improvements in the production system as indicated in the real-world semiconductor manufacturing case study and, hence, contributes a novel, data driven approach for production control. As a conclusion, the ability to avoid a large majority of violations of time constraints shows the approaches effectiveness and the future requirement to more accurately integrate such product-inherent constraints into production control.

Keywords Complex job shop · Time constraint · Semiconductor manufacturing · Production control

Introduction

The trend towards individualization is advancing at breakneck speed and enforces ever increasing requirements on production. Customization and individualization lead to manufacturing systems with unique products. Each unique product exhibits traits that differentiate it from similar, yet not identical, unique products. Examples can be due to personalization or in a re-manufacturing setting where each core that is returned from the usage phase exhibits unique characteristics (Wurster et al., 2022). This uniqueness increases the complexity in manufacturing and feeds the trend towards complex job shops, in which each job has unique characteristics and thus, needs unique processing, transportation or setup times (Waschneck et al., 2016). These require more flexible production networks (Yin et al., 2021) and a much more flexible manufacturing system with an in turn fast, real-time production control system that is capable of handling unique products and all their individual constraints (Yuan et al., 2021). The manufacturing of computer chips, in other words semiconductors, takes place in complex job shops and presents a great example for such product-inherent constraints: Time constraints that limit the waiting and transportation time between two or more processing steps.

Semiconductor products play a key role in the fourth industrial revolution. With digitization taking place in almost

✉ Marvin Carl May
marvin.may@kit.edu

[1] wbk Institute of Production Science, Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, 76131 Karlsruhe, Germany

Published online: 03 August 2024

every industry, the demand for semiconductors continues to grow, especially driven by booming markets like artificial intelligence, 5 G and the Internet of Things (IoT) (May et al., 2024). This leads to strong competition within the industry and manufacturers having to produce cost-effectively. To realize necessary cost reductions, improvement of operational processes offers the best opportunities (Mönch et al., 2009).

Manufacturing of semiconductors takes place in so-called fabs – short for semiconductor fabrication plants. Re-entrant product flows, re-routing due to machine failures, arrival of urgent jobs and stringent quality requirements make it one of the most complex and dynamic manufacturing environments (Mönch et al., 2013). Several hundred process steps in different work areas are required to turn a wafer into a chip, associated with many process-related challenges.

During the time between one process step and another, natural phenomena such as oxidation, crystal formation and ion migration can cause wafers' surfaces to change their properties (Lima et al., 2021). There is a plethora of research on the physical and chemical property changes during theses processes. Kinetic models play an important role here (Markowich et al., 2012) yet they are focused solely on a product variant level and do not regard the interplay with fab operations. Thus, operations management in a semiconductor fab has to abstract these restrictions into product-inherent constraints for high quality. To ensure this quality is not negatively affected, time limits between various operations are installed. If a lot of wafers, signifying the product, exceeds the time limit, it either has to be scrapped or expensive rework is required (Klemmt & Monch, 2012). When a lot reaches the starting equipment of a time constraint, a decision whether a lot is released and a time constraint is started has to be taken. This time consuming and stressful task is usually performed manually based on heuristics and the experience of the operators (Lima et al., 2017b), giving rise to the need of an automated approach that improves time constraint adherence.

To that end, this study presents a novel approach for production control with product-inherent constraints on the example for time constraints in semiconductor manufacturing. The proposed model is based on uncertainty informed artificial intelligence and provides production control with material flow by inhibiting the onset of processing certain lots if the risk of violating the time constraint exceeds a determined threshold. The AI model is used to forecast expected transition times, that cannot exceed the time constraint, so that lots are withhold if their estimated probability of time constraint adherence suggests that potential scrapping or rework will become necessary. Thus, the model provides a novel approach to utilize the uncertainty in manufacturing related artificial intelligence models to distill knowledge on product individual level into an intelligent production control. The model is validated with real world industrial semiconductor manufacturing data from an entire fab over several months.

The paper is organized as follows. The literature review is presented in "Literature review" section and focuses on product-inherent constraints in manufacturing and the state-of-the-art of manufacturing under such time constraints in semiconductor fabs. Furthermore, the main contribution of this work is highlighted. "Decision model" section gives an overview of the manufacturing setting at hand and defines the problem treated in this paper. Additionally, the specifications of the constructed models are presented. "Numerical experiments" section presents the results of the numerical study carried out in the real-world manufacturing setting. In "Experimental validation of constrained actions" section, the results are discussed and the benefits of the modeling approaches are presented. Finally, "Conclusions and further developments" section concludes and summarizes the paper and discourses possible relevant further developments.

Literature review

To that end, this paper presents a holistic literature review on the state-of-the-art for adequately considering product-inherent constraints for production control in general in "Product-inherent production control constraints" section. In the concrete context of semiconductor manufacturing, the state-of-the-art techniques are reviewed in "State-of-the-art time constraint control in semiconductor manufacturing" section. The contribution of this paper is outlined in "Contribution" section.

Product-inherent production control constraints

Generally, production control consists of scheduling at a higher level and dispatching at a lower level. It can therefore be distinguished from production planning which consists of making strategic decisions on a quarterly base and releasing orders at a monthly scale (Mönch et al., 2013).

Product-inherent constraints on the production control can be of different type and are found across various industries. Production processes can be constrained for yield, safety, environmental or quality reasons. Another example can be found in the chemical industry in polymerization reactors. Several process variables like reactor temperature and feed flow rate have to be constrained in order to comply with quality and safety requirements (Abel et al., 2000). Highly energy intensive industries like steel manufacturing are forced to constrain their processes and energy consumption in order to reduce their environmental impact (Somboonwiwat et al., 2018). In order to accomplish quality goals and increase throughput, manufacturers may also

impose time constraints within production processes. Time constraints can be required when scheduling because one process step may depend on the output of a previous step forcing it to wait. For example in oncology clinics, treatments may sometimes only start after a specific start time because the patient may only be treated after an oncologist appointment (Liang et al., 2015). Analogically, a process step may require an intermediate product forcing it to be scheduled after the process step of which the output is the intermediate product (Sundaramoorthy & Karimi, 2004). Time constraints can also be imposed to limit the time between multiple process steps in order to prevent the product from degrading. For example, when using cold cure bonding, the time between the anodizing and the bond lay-up process step must not exceed a certain time limit (Higgins, 2000). When handling perishable products in the food industry (e.g. yogurt), each process step in the production has to be performed within a certain time limit for the product not to perish and lose its value (Amorim et al., 2013). Also, holding times of intermediate products can be limited due to sterility reasons in pharmaceutical manufacturing (Eberle et al., 2016). This phenomenon can also be observed in semiconductor manufacturing. Time constraints are frequently imposed to mitigate negative effects on the wafers surface caused by oxidation and contamination (Wang et al., 2018). Given the increasingly available data and availability of high performance algorithms Artificial Intelligence and Machine Learning techniques are perfectly fit to counter such product-inherent constraints. As the field is large and evolving fast the reader shall be referred to the respective state-of-the-art literature.

Due the frequent occurrence of time constraints in semiconductor manufacturing and a high availability of quality data, semiconductor manufacturing is investigated in this paper in order to verify the proposed production control model. A state-of-the-art literature review of time constraint control in semiconductor manufacturing is conducted which is presented in the following.

State-of-the-art time constraint control in semiconductor manufacturing

In semiconductor manufacturing, equipment, denoting the manufacturing equipment which is used interchangeably to the commonly used machines, is the biggest cost driver (Hong et al., 2023). Thus, using capacities with opportunistic behavior and ensuring zero defect manufacturing is central (Valet et al., 2022) and accordingly, fabs are run for 24 h on every single day. In order to minimize setups and improve coordination, wafers, with several ICs each, are stacked to lots, each containing up to 50 or typically 25 wafers (Ziarnetzky et al., 2017). This technologically intense manufacturing, with complicated chip design, is seen in the presence of abrasive processes, chemical and electro-physical processes, forming and cutting processes. As ICs are manufactured layer by layer metrology and advanced surface engineering that are required need complicated machines or equipment as well as complex control and organization of production systems (Mönch et al., 2013). Aside from technical and technological complexity,the greatest challenge in producing semiconductors is to coordinate this complex job shop (Mönch et al., 2011). Each wafer requires 1000 or more processing steps, that need several minutes up to many hours each (Valet et al., 2022). Recurrent material flow gradually builds these integrated chips, layer by layer. Factories in semiconductor manufacturing are operated on the verge of the physical and technological boundaries resulting in only a part of the ICs from manufacturing being able to reach the highest usability level (Mönch et al., 2009). Yield denotes this share of functional chips over total manufactured and should be kept as high as possible. Besides errors, yield loss can also be made up of wafers contamination. The contamination often stems from ion migration, crystal formation, native oxidation or the deposition of dust (Lima et al., 2021). Such a contamionation, also called impurity, alters the surface of the chips and inhibits the electrical flow from following the designed patters. To minimize contamination equipment in manufacturing of semiconductors is, hence, located in a so called clean room (Klemmt & Monch, 2012). Still, each wafer can only remain in the clean room for several hours, because otherwise the wafers often have to be scrapped due to the contamination which cannot always be cleaned (Altenmüller et al., 2020). Not violating these time constraints is, thus, crucial to the success of semiconductor manufacturers (Arima et al., 2015).

Production and business activities of companies are crucially impacted by production planning and control (PPC) (Wang & Liu, 2013). The main goals of PPC in semiconductor manufacturing are the minimization of costs and an increase in productivity, while continuously improving quality and due date performance (Uzsoy et al., 1992). In order to identify relevant literature treating time constraint management in semiconductor manufacturing, a systematic literature review using grounded theory is conducted using the approach proposed by Wolfswinkel et al. (2013). The review was performed by querying Scopus about publications dealing with time constraints or time coupling and relevant alternative descriptions in semiconductor manufacturing. The literature is filtered by title, abstract and keywords to preserve production system based approaches leading to 28 publications. These are extended with a forward and backward search. The results are clustered into capacity planning, scheduling and dispatching according to the PPC hierarchy as outlined in the following.

Time constraints have to be taken into account at all the different stages of PPC. According to Mönch et al. (2013), PPC can be divided into three different levels: capacity planning, scheduling and dispatching. The literature treating time

constraint control at a capacity planning and scheduling level is summarized in Table 1.

Overall, only few papers regard time constraints at a capacity planning level. The majority of publications rely on queuing systems, assuming a general independent distribution of interarrival times (GI) and a general distribution for service times (G) with a number of m machines in order to form a GI/G/m queuing network, like Tu and Liou (2006) and Kitamura et al. (2006). Furthermore, Pappert et al. (2016) perform production simulations and conclude that time constraints significantly reduced production capacity. Additionally, Kuo et al. (2011) introduce a modeling approach based on neural networks with the goal of reducing the cycle time of a wafer fab. Lastly, Mastrangelo et al. (2024) build a policy to modulate a two-staged manufacturing system's capacity with time constraints. They apply a markovian system representation and validate the model in real environment in diffusion and cleaning stages to show a significantly pareto-improvement on throughput and quality yield.

The biggest share of the identified literature, however, considers time constraints at a scheduling level. As displayed in Table 1, the scheduling problem is predominantly modeled with mixed integer programming (MIP), mixed integer linear programming (MILP) or mixed integer non-linear programming (MINLP). Few exceptions are a disjunctive graph, a Kanban system, an analytical model or constraint programming (CP), also in combination with MIP.

To minimize the waiting time variation and thus reduce time constraint violations, Yu et al. (2013) develop a MIP-based scheduler for up to 25 jobs that can exactly solve this limited problem set. As the MIP model is not able to solve problems with more than 25 jobs, an additional approximate solution is provided, showcasing the disadvantage of the state-of-the-art approaches. While An et al. (2016) are able to find optimal solutions for problems with up to 30 jobs within a reasonable computation time, Kim and Lee (2017) was only able to solve 20 jobs with branch and bound techniques, due to the consideration of complex time constraints. The idea behind decomposition-based approaches is to divide the problem into multiple sub-problems inspired by the divide and conquer paradigm. For example, Sun et al. (2005) and Maleck et al. (2019) divide the scheduling problem into three levels and propose a solution for each problem individually. Klemmt and Monch (2012) and Jung et al. (2014) break down the problems recursively to find near-optimal solutions and optimize KPIs such as total tardiness and time constraint violation rates. Genetic algorithms (GA), on the contrary, are inspired by natural selection and belong to the class of evolutionary algorithms. They are commonly used to find solutions to complex problems that are impossible to solve exactly within a reasonable time frame. Klemmt et al. (2008) find that exact approaches are limited in terms of problem dimension and therefore apply a GA to find near-optimal solutions to four representative oven batching problems. Finally, Han and Lee (2023) consider a three-machine flow shop with missing operations and provide various types of heuristic algorithms to solve the scheduling problem heuristically. Two metaheuristic algorithm—iterated greedy and simulated annealing—have shown to be effective and efficient.

Only few studies deviate from a MIP-based modeling approach. A disjunctive graph representation is used by Yugma et al. (2012) to group lots in batches and apply a simulated annealing algorithm to optimize cycle time and machine capacity. A decision support system is developed by Perraudat et al. (2019) using a kanban system model. Wu et al. (2016b) build an analytical model to quantify the trade-off between a higher capacity and a lower rework rate requiring a higher and a lower WIP-level respectively.

At a dispatching level, individual lots have to be managed. Due to the high uncertainty of production control in semiconductor manufacturing, the non-linear material flow, lots re-entering the production flow and sudden machine breakdowns, managing time constraints ultimately comes down to dispatching the right lots at the right time. The respective literature is shown in table 2.

Lima et al. (2017a, 2019) and Sadeghi et al. (2015) apply a method based on sampling for predicting if a lot will adhere to time constraints upon entry. Building on this, Lima et al. (2021) present an improved algorithm compared to the original by Lima et al. (2017a), along with a problem modeling approach that aligns more closely with real industrial conditions. Additionally, Lima et al. (2017b) develop a decision support system to identify tool interruptions by grouping machines based on shared recipes and aggregating time constraint transitions by their ending equipment. In contrast, Altenmüller et al. (2020) use reinforcement learning (RL) and train a RL agent that is rewarded for reducing time constraint violations, successfully outperforming dispatching rules such as the First In, First Out (FIFO) heuristic.

Several studies focus on implementing rule-based systems, often validated through simulations. For instance, Tu et al. (2010) introduce dynamic job control in the furnace area, allowing managers to dynamically manage work in progress (WIP) and ensure adherence to time constraints. Arima et al. (2015) maximize throughput for lots adhering to time constraints by combining dispatching and loading rules. Similarly, Kobayashi et al. (2013) aim to optimize dispatching rules for a re-entrant flow shop. Kopp et al. (2020) apply a rule-based approach that considers lot priority, setup time, and a time constraint criticality factor. Ciccullo et al. (2014) and Pirovano et al. (2020) examine the batching process before a time constraint between cleaning and diffusion processes, proposing heuristic algorithms to avoid scrapped lots. Furthermore, Zhang et al. (2016) develop a plan-based control system to handle the dynamic and stochastic envi-

Table 1 Classification of the relevant literature based on modeling, approach and objective at a capacity planning and scheduling level

Level	Modeling	Approach	Source
Capacity planning	Experiments	Simulations	Pappert et al. (2016), Huang et al. (2011)
	Queueing system	Queueing theory	Tu and Chen (2009a, 2009b, 2010, 2011), Tu and Liou (2006), Kitamura et al. (2006), Ono et al. (2006)
	Neural network	Regression	Kuo et al. (2011)
Scheduling	MIP/MILP/ MINLP	Branch and bound	An et al. (2016), Kim and Lee (2017)
		Cuckoo search algorithm	Zhou et al. (2019)
		Decomposition	Jung et al. (2014), Klemmt and Monch (2012)
		Estimation of distribution algorithm	Wang et al. (2014)
		Exact solution	Maleck et al. (2017), Yu et al. (2013), Cho et al. (2014), Kao et al. (2011), Klemmt et al. (2008)
		Genetic algorithm	Lee (2020), Wang et al. (2015), Chien and Chen (2007), Klemmt et al. (2008)
		Simulated annealing	Nattaf et al. (2019), Zhou and Wu (2017), Han and Lee (2023)
		Heuristic control policy	Su (2003), Yurtsever et al. (2009), Yu et al. (2017)
	MIP + CP	Decomposition	Sun et al. (2005), Bixby et al. (2006)
		Exact solution	Maleck et al. (2018)
	CP	Decomposition	Maleck et al. (2019)
	Disjunctive graph	Heuristic control policy	Yugma et al. (2012)
	Kanban system	Simulation	Perraudat et al. (2019)
	Analytical model	Heuristic control policy	Wu et al. (2016b)

Table 2 Classification of the relevant literature based on modeling, approach and objective at a dispatching level

Modeling	Approach	Source
Disjunctive graph	Sampling-based heuristic	Lima et al. (2017a, 2017b, 2019, 2021), Sadeghi et al. (2015)
Feedforward ANN	Heuristic control policy	Chakravorty and Nagarur (2020)
MDP	Numerical calculation algorithm	Wu et al. (2010, 2012a, 2012b, 2016a)
	RL	Altenmüller et al. (2020)
Experiments	Heuristic control policy	Tu et al. (2010), Arima et al. (2015), Kopp et al. (2020), Ciccullo et al. (2014), Kobayashi et al. (2013), Pirovano et al. (2020)
	Plan-based heuristic	Zhang et al. (2016)
BIP	Heuristics	Ham et al. (2011)
MIP/MILP	Exact solution	Maleck and Eckert (2017)
	Heuristic control policy	Chang and Chang (2012), Wang et al. (2018)
	Neural network + heuristic control policy	Li et al. (2012)
	Genetic algorithm	Jia et al. (2013)
ARMA	PE + PI	May et al. (2021c)
ML & ARIMA		May et al. (2021b)
Queueing model	Heuristic control policy	Yang et al. (2015)

ronment in semiconductor manufacturing, though it lacks real-time capabilities and practical application in a real semiconductor setting.

Though the following studies share a common foundation in integer programming models, their objectives, work areas, and utilized parameters differ significantly. Ham et al. (2011) focus on minimizing computation time to provide a real-time dispatching heuristic using binary integer programming (BIP) in a two-machine flow shop. In contrast, Chang and Chang (2012) integrate dispatching rules into a three-stage approach to reduce cycle time. Jia et al. (2013) combine a pull-pull-push-push strategy with a genetic algorithm to develop a closed-loop dispatching heuristic. Multiple dispatching rules incorporating risk factors are proposed by Maleck and Eckert (2017) to account for tool failure probabilities. Li et al. (2012) use a learning-based approach, training a neural network to set the weighted parameters of a dispatching rule based on job due dates and machine workloads. Wang et al. (2018) explore time-link area constraints and develop a control policy for initiating the first process of a time constraint. May et al. (2021c) and May et al. (2021b) apply autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models, respectively. May et al. (2021b) also construct a recurrent neural network (RNN), specifically a Long-Short-Term-Memory (LSTM) network, to compare the performance of different approaches. Both studies combine a single point estimator (PE) with an extended prediction interval to estimate the probability of a lot adhering to a time constraint. Other approaches include using a neural network to predict cycle time by Chakravorty and Nagarur (2020) or queue lengths (May et al., 2021a), starting processing only if the predicted cycle time is within the time limit. Finally, Yang et al. (2015) address the challenge of long setup times for implantation equipment compared to their process times. To prevent continuous production runs from being interrupted by arriving lots with time limits, they propose a novel dispatching algorithm that combines recipe changes with processing arriving time constraint lots.

Contribution

As explained by the literature review, time constraints in semiconductor manufacturing pose a serious trouble to operations and currently are predominantly modeled using mathematical models and heuristics. Many of these approaches make strong simplifications regarding the number of machines or the heterogeneity of the transitions in a wafer fab. Thus, applying them in an existing semiconductor fab is hardly conceivable for this theoretical work. As a result, solutions are oftentimes not verified in real-world sized semiconductor systems let alone in an existing semiconductor manufacturing environment and in particular not in an entire fab. Thus, they fall short of any applicability in real-world environments. An important contribution is hence the demonstration of the proposed model in a real-world semiconductor fab with practical implications. Additionally, machine learning approaches are significantly underrepresented despite their so far promising results. Most notably, the inclusion of uncertainty in artificial intelligence has been neglected. In contrast to alternative approaches, AI based approaches can be enhanced to deal with industrial size problems and replace industrial, human or priority rule based approaches. To that end, the proposed models must signifi-

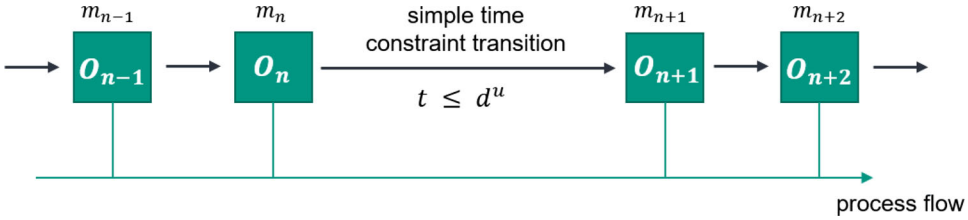

Fig. 1 Simple time constraint: Transition time t between two operations O_n and O_{n+1} limited by an upper limit d^u

cantly outperform state-of-the-art approaches. Therefore, the need to regard time constraints holistically using real-world data from semiconductor manufacturing is derived and a modeling approach is developed accordingly. The model is validated with a real-world wafer fab.

Decision model

We develop a decision model based on real-time data to improve the production control in a complex job shop manufacturing setting. This section defines the formal manufacturing setting, gives a detailed problem description and introduces the model specification.

Manufacturing setting

A complex job shop is a manufacturing setting where $m \in M$ machines, herein also denoted as equipment, manufacture $j \in J$ jobs that can involve several complexities. Each job j contains a process flow $f \in F$ with $f = \{o_1, o_2, \ldots, o_k\}$ with $o \in O$ signifying possible operations. Each machine can perform $O_m \subset O$ operations. First, a re-entrant flow of jobs (i.e. a job is re-entering the manufacturing setting after one pass through) is possible, as for instance a job's process flow f_j contains operations that are performed outside of the manufacturing system such as quality assurance. Furthermore, the setup times are sequence dependent, meaning that the setup time $t^{setup}_{o_1,o_2}$ can differ significantly from the setup times $t^{setup}_{o_3,o_2}$ and $t^{setup}_{o_2,o_1}$ for $o_1, o_2, o_3 \in O_m$. Likewise, the processing time $t^{processing}_{o_k,m_1}$ varies for different jobs while lot sizes for any lot $l_1 = \{j_1, j_2, \ldots\}$ are heterogeneous. Due to frequent and hard to control machine breakdowns, decisions in a complex job shop have to be taken under a high degree of uncertainty (Waschneck et al., 2016). Prescribed due dates and time constraints, that limit the time between operations o_a and o_b with $o_a, o_b \in f$ are a major concern. Failing to hold due dates can result in high costs and low customer satisfaction, while failing to adhere to time constraints can cause insufficient quality and may require to restart manufacturing of this lot from scratch.

Whenever an equipment has to select the next lot to be processed, a decision in terms of gate has to be made. Regarding a time constrained lot, this decision includes the initialization of the time constraint as there is no turning back after the processing is started. The latest possible decision is therefore right before starting processing and targets the minimizing the number of violations of time constraints. This decision is often based on heuristics or the experience of the technicians in charge and is a time consuming and stressful task (Lima et al., 2017b). The objective of this approach is to predict the probability of a lot adhering to its given time constraint to assist the technicians at a dispatching level by using the advanced machine learning techniques (Chen et al., 2023) without using simulation or digital twin based approaches (May et al., 2022).

The modeling approach presented in this paper specifically targets simple time constraints. As depicted in Fig. 1, the transition time stands for the total time that elapses between the end of operation O_n and the beginning of operation O_{n+1} including transportation, handling and queue time. When the processing of the lot arriving at the first machine m_n is started, the time constraint is initialized and there is no turning back. Only when processing at the subsequent machine m_{n+1} is started within the given time limit d^u, the time constraint is fulfilled and the lot is prevented from exceeding its time constraint.

Therefore, time constraint violations can be identified in retrospect. Hence, any improvement from true positive and true negative predictions have a large effect on the semiconductor fab and can save precious value, energy and time.

Problem description

In semiconductor manufacturing, time constraints can occur limiting the transition time between two or more machines by an upper limit. Time constraints are classified into different levels of complexity according to Klemmt et al. (2008) and Wang et al. (2018) as follows: Simple time constraints put a limit to the transition time between two consecutive machines, i.e. without any intermediate steps, while transitions spanning multiple machines are called timelink area constraints. Time constraints are considered complex when they consist of multiple overlapping or directly successive time constraints. Simple time constraints are specifically regarded in this paper, due to them occurring most frequently in practice.

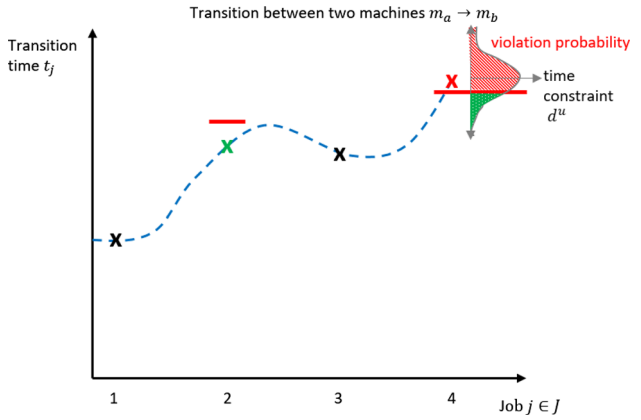

Fig. 2 Decision model using the prediction interval for estimating the probability of adhering to the time constraint

At the moment, dispatching decisions are performed manually by wafer fab operators mostly based on their experience and defined heuristics. In order to provide the operators with meaningful information, the estimated adherence probability of a time constraint transition is predicted. As outlined in "State-of-the-art time constraint control in semiconductor manufacturing" section, machine learning approaches are considerably underrepresented in the literature of time constraint management in semiconductor manufacturing settings. Due to their strong ability to cope with dynamic and complex manufacturing environments, two models are constructed and presented in the following.

Model specification

Generally, neural networks can be treated as black boxes, where a defined input provides a defined output. How or why specific results are produced oftentimes can not be determined leading to a lack of interpretability of the result. Therefore, the point estimators of the neural networks are supplemented by an uncertainty quantification method in order to construct a prediction interval as introduced in May et al. (2021b). Therefore, Monte Carlo dropout is introduced in order to estimate the model uncertainty. Given a specific time limit and a prescribed confidence level, the uncertainty is in turn used to calculate the estimated time constraint adherence probability.

Based on this concept, two models are constructed: For the resource-based modeling approach, the time series data is aggregated based on equipment groups to exploit equipment group-specific characteristics. The transitional modeling approach is fit to the entire data set.

Each transition between any two machines $m_a, m_b \in M$ can potentially contain time constraints. A transition is only interesting and regarded if there is more than one transition of a job $j \in J$ that transitions from machine m_a to m_b. Transitions without any job transitions or with only one job transition do not provide enough data to base the decision on previous observations. The transitions can contain time-constrained jobs that have an upper limit d_u for the transition time and non-constrained transitions. In the transitional modeling approach, each possible transition is treated individually, whereas the resource-based modeling approach pools transitions based on temporal-spatial data. For each model of a transition, or pooled resource transition, a prediction model is formed as time series data on the transition times is available. Each next potential transition uses the currently available (past) data to predict the next transition time. Based on the prediction interval fed by the Monte Carlo dropout, this point predictor of an individual transition is transformed into an estimated probability. The approach is presented in Fig. 2, which shows an exemplary case of four total jobs, the second and forth being time constrained. The second has successfully adhered to the time constraint. Using the past three observations, the estimation for job 4 is performed. As indicated, the time constraint is lower than the expected transition time and also has a low estimated adherence probability (below the bar) compared to a greater estimated violation probability. This can be regarded as the confidence about the transition time being smaller than the time constraint. Selecting this confidence is paramount for a good model and based on operator and expert discussions we selected 90% and 95% as the prescribed confidence. Note, that the confidence is not to be mistaken with the expected share of jobs that violate the time constraint despite being predicted to adhere with the given confidence.

Prediction intervals

A prediction interval is defined as a future value lying between an upper and a lower bound with a given probability (Chatfield, 2001). Since we are only interested in a transition possibly exceeding its defined time limit, the lower bound of the prediction interval is set to $-\infty$. The upper bound can than be calculated by

$$\hat{y} + z_{\alpha/2}\sqrt{Var(e)} \tag{1}$$

where $\hat{y}$ denotes the point estimation, $z_{\alpha/2}$ the appropriate percentage point of a standard normal distribution and $Var(e)$ the variance of the models prediction error (Chatfield, 2001). The first two parameters are directly available from the models prediction and the prescribed confidence interval respectively. The variance of the prediction error consists of the sum of the model-specific variance $\sigma^2_{\hat{y}_i}$ representing the model's uncertainty and the variance of the noise σ^2_{ϵ} in the underlying data (Khosravi et al., 2011). The variance present in the noise can be calculated by division of the sum of the squared differences between the observed values

y_i and the model's prediction $\hat{y}_i$ by the number of samples n_t in the test set (Zhu & Laptev, 2017).

$$\sigma_{\hat{\epsilon}}^2 = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_i - \hat{y}_i)^2 \tag{2}$$

Generally, dropout is a method to prevent NNs from overfitting. At each training step, a collection of neurons is randomly selected from the network and the connections to their successive neurons are cut. The selected neurons are thereby dropped out, thus their output does not influence the output of the network. Applying this process to the fully trained NN during testing is called Monte Carlo dropout and can be used to determine the model uncertainty. It can be shown, that applying dropout before each hidden layer of the NN during testing is equivalent to the approximation of a Gaussian process (Gal & Ghahramani, 2016). The model uncertainty $\sigma_{\hat{y}_i}^2$ can therefore be estimated with sample variance (Zhu & Laptev, 2017). An ensemble of predictions $\hat{y}_{i,n}$ is obtained by performing N stochastic forward passes, from which the model uncertainty can be calculated through the following equations:

$$\hat{y}_i = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_{i,n} \tag{3}$$

$$\sigma_{\hat{y}_i}^2 = \frac{1}{N-1} \sum_{n=1}^{N} (\hat{y}_{i,n} - \hat{y}_i)^2. \tag{4}$$

In order to obtain the point estimator, two modeling approaches are introduced in the following.

Transitional modeling approach

Due to the sequential nature of the data at hand, a transitional modeling approach incorporating a recurrent neural network (RNN) is constructed first in a similar style as May et al. (2021c). Alongside the observed transition times, the current queue meaning the number of lots waiting in front of the upstream equipment is modeled. The result is a sequence of data points which is then split up into subsequences of equal length using a sliding window approach and forming the model's input data. As depicted in Fig. 3, the RNN predicts a transition time based on the input data. Given the model's uncertainty, the prediction can in turn be used to calculate an adherence probability.

For the specific implementation of the RNN, a gated recurrent unit cell (GRU) is used. Compared to long short-term memory (LSTM) cells, they require less computational effort. Furthermore, studies have shown that on different tasks, GRU can outperform LSTM in terms of immunity to

Fig. 3 Transitional model architecture

the exploding and vanishing gradient problem (Mateus et al., 2021), prediction errors (Yamak et al., 2019) and overall performance (Zarzycki & Ławryńczuk, 2021). To enlarge the state-of-the-art and compare the proposed approach, we additionally implement the model using a LSTM network in style of May et al. (2021b)

Resource-based modeling approach

Generally, it can be found that when grouping time constraint transitions based on shared equipment groups, transitions with similar properties are aggregated (Lima et al., 2017b). In order to reduce the heterogeneity of the training data, it is decomposed into multiple subsets. For each subset, a feed-forward neural network (FFNN) is trained and the individual predictors are combined into one overall model. Upon input of data, the decider delegates the prediction to the appropriate NN based on the equipment groups and the prediction of that NN than represents the model's output (see Fig. 4).

As proposed by Lima et al. (2017b), the transitions are grouped based on shared machine groups. Each machine is capable of performing one or multiple actions onto the lots they process. Furthermore, machines can share the same actions. First, machines that share actions are assigned the same groups. These initial groups are united upon shared machines resulting in distinct machine groups. Secondly, each transition has a starting and an ending equipment, thus a starting and an ending equipment group. The transitions are partitioned into unique combinations of equipment groups where transitions in one group share the same starting and ending equipment group. Any transition can therefore be associated with a specific combination of a starting and an ending equipment group. The transitions are grouped based on those combinations. When handing input data to the model, it determines the starting and ending equipment group of the transition and dispatches the prediction task to the respective NN based on the given equipment groups.

Numerical experiments

The numerical experiments are performed in a semiconductor manufacturing plant that contains more than one thousand machines, in other words manufacturing equipment. The data set at hands was recorded over the time horizon of several months and contains any positional data for the wafers with respect to performing an operation o_n on any equipment.

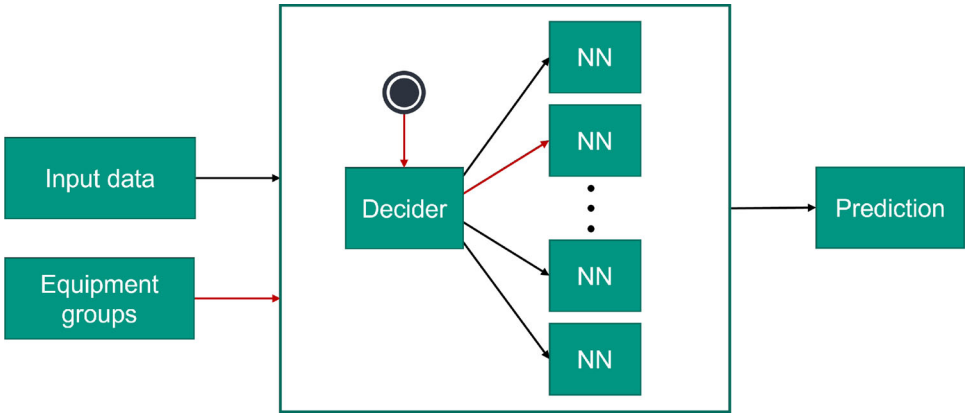

Fig. 4 Resource-based model architecture

This is augmented with information about the wafer-specific time constraints. "Empirical setting" section describes this environment and "Implementation details" section the implementation details.

Empirical setting

Chip manufacturing comprises many different steps making it a highly complex endeavor. It includes circuit design, processing under clean room conditions, testing, and packaging of the wafers (Xiao, 2012). Forming the basis of an integrated circuit (IC), wafers are thin slices of ultra pure single-crystal silicon. Semiconductor manufacturing can be divided into front-end and back-end processes. The front-end consists of the fabrication of the integrated circuits on wafers. To process these wafers, ultra pure silicon wafers are repeatedly undergoing many different process steps to alter the electrical conductivity on a layer by layer basis. By controlling the alteration of the conductivity in a semiconductors due to implemented impurities, tiny conductive paths are realized (Mönch et al., 2013). These make up transistors which themselves are aggregated over many layers to form an integrated circuit. The integrated circuits are then subject to thousands of electrical tests to verify the functionality of the device (STMicroelectronics, 2000). Chips that failed the electrical tests are marked and excluded from further processing. Entering the back-end, the first step is to separate the chips by cutting the wafers. The functional chips are then wired and packaged for easier handling and protection from environmental conditions. Final tests are conducted on each IC by automated test equipment to ensure product quality.

When manufacturing semiconductor products, hundreds of operations must be performed almost impeccably (May & Spanos, 2006), giving rise to several process-related challenges (Mönch et al., 2013). On the process side, wafers have to be processed on the same types of machines multiple times. The resulting recurrent nature of chip manufacturing leads to a re-entrant process flow making production planning and control a challenging task. Besides complex manufacturing control, a key lies in controlling the processes and the alteration of the semiconductor between processes. If conductive paths are changed due to contamination or oxidation on the wafer's surface between and during processes, the integrated circuit will not work properly. To mitigate these negative effects, time constraints are commonly imposed (Wang et al., 2018), limiting the time between two or more consecutive operations by an upper bound. As many ICs are placed on one wafer, only few non working ICs can be tolerated before the whole wafer has to undergo expensive rework or in many cases leads to scrapping (Mönch et al., 2013). Adhering, in other words not violating, the given time constraints therefore is central to effective and efficient semiconductor manufacturing.

Operational data

The dataset collected from the front-end of a semiconductor wafer fab contains transitional data as each processing of a wafer at any equipment, in other words machine, is recorded. During the course of several months, frequent log data from equipment and associated lots has been captured and aggregated to build the dataset of machine events that contain information about starting and ending time of processes for wafers and maintenance. This represents the digital shadow representation of the actual factory for several months containing in the millions of observed entries in a similar style to Table 3.

Not all possible transitions between two machines have been used within that period of time. Changes over time can occur due to changing equipment capabilities and new products—in other words IC designs—that enter the wafer fab. Additionally, transitions of lots used for testing are removed. Using this data, it is possible to infer the transition time between machines, which has to be lower than the prescribed time constraint to avoid scrapping.

When looking at specific starting and ending equipment combination, the transitions in between can be considered realizations of a function. As initially investigated by May

Table 3 Exemplary transition log data

Time stamp	Equipment	Process operation	Lot ID	...
123456	A12	SPS_42	EXAMPLE987	..
..	..	..	..	..

et al. (2021c), some of those functions exhibit a significant autocorrelation. This means, that a specific observation of a transition time depends to some degree on its past realizations. For example, when a long transition time due to a sudden machine down is observed, the successive transition is likely to be long as well and thus correlates with a previously observed value. How this autocorrelation can be exploited to predict transition times is described in the following.

Implementation details

The prediction of the transition times aims at exploiting this autocorrelation of a sequence of observations by feeding a sequence of past observations into the neural network in order for it to predict the potential successive value.

When aggregating the operational data by the lots, a sequence of transitions for each lot can be determined, consisting of a starting and an ending equipment, as well as a time stamp and an identifier for the process performed at the starting equipment. In order to benefit from the autocorrelation between

This data is complemented by a queue feature representing the number of wafers waiting to be processed at the downstream equipment. This is done by increasing the queue count every time a transition towards the downstream equipment is started and by decreasing the queue count every time the downstream equipment finishes processing. The initial queue count is approximated by identifying all lots that are in transition to the respective machine before the time frame under study begins. A correlation analysis of the queue feature reveals that the overall correlation of around 0.25 is only moderate. However, a higher correlation is observed for machines performing single and cluster operations while the correlation is almost non-existent for machines performing batch processes. As during batch processes, multiple lots are processed at once, a longer queue in front of the equipment can possibly be beneficial to arriving lots. Overall, the queue feature can provide additional information in some cases and is therefore added to the input data.

The categorical inputs, namely starting and ending equipment, are mapped to numerical ones. This is done using entity embedding by assigning a number to each unique equipment.

The input data is split into a training, a validation and a test set. In order to provide the models with a sufficient amount of training samples while reserving enough samples for model validation, a ratio of 0.7 to 0.15 to 0.15 is chosen respectively. The training data is used to fit the neural network by adjusting the weights of the neural network. The validation set is used for regularization during the training phase. After every epoch, the loss on the validation set is tested and an increase eventually triggers early stopping. Finally, the model's performance is evaluated using the test set.

A detailed implementation description for each of the models is introduced in the following.

Resource based models

The individual neural networks are all FFNNs and consist of at least two dense layers. For every NN, different numbers of layers, different kernels, bias regularizers and activation functions are evaluated in a grid search manner. Lastly, Bayesian hyper parameter is applied to optimize dropout rates and number of neurons.

Transitional models

The hyperparameters of the neural networks of the transitional models—consisting of a GRU-based and a LSTM-based model—have to be tuned subsequently. This is done using hyperopt, a Python library for serial and parallel optimization over specified search spaces, originating from Bergstra et al. (2013). The library's search algorithm then minimizes the loss returned by a function given a search space. In order to calculate the loss, a function is defined that trains the model given a specific set of hyperparameters, evaluates its performance and returns the prediction loss. In order to bound the search space, initial experiments are carried out in a grid search manner. The resulting search space is depicted in Table 4. Additionally, the default optimizer of Keras is used.

Finally, the GRU model is built as follows: Two embedding layers of size 92 turn each categorical feature respectively into a dense vector of fixed size. The input data is then fed to a GRU layer consisting of 27 neurons, a dropout rate of 0.250, and a recurrent dropout rate of 0.266. A final dense layer maps the 27 GRU outputs to a single output value. Additionally, early stopping and learning rate reduction on plateaus is enabled. Analogously, the LSTM model consists of an embedding of size 30, an LSTM layer with 32 neurons, and a dropout rate and recurrent dropout rate of 0.2.

In Fig. 5, the training and validation losses are depicted over the course of the training process for the GRU model.

Table 4 Search space and results of the hyperparameter optimization: number of neurons, embedding size, dropout rate (DR), and recurrent DR

	Neurons	Embedding size	DR	Recurr. DR
Parameter range	{10, …, 40}	{20, …, 100}	[0.1;0.35]	[0.1;0.35]
Results GRU	27	92	0.250	0.266
Results LSTM	32	30	0.2	0.2

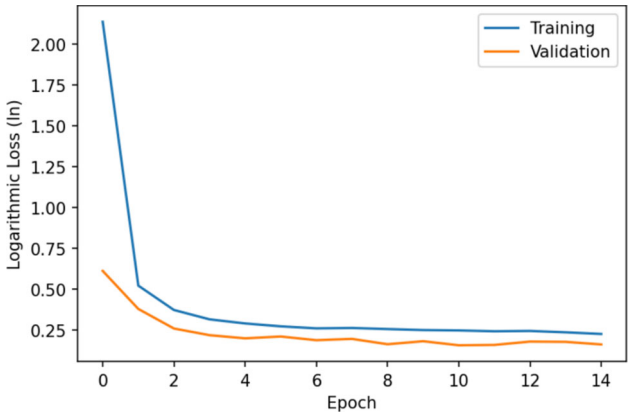

Fig. 5 Training and validation loss of the GRU model

Up until epoch 14, training and validation loss decrease as the training progresses. Training is stopped at epoch 15 due to an increase in validation loss.

Experimental validation of constrained actions

The presented models are implemented and experimentally validated using real-world data. Firstly, performance metrics are introduced in "Evaluation approach" section that are used to evaluate the decision models. Secondly, the evaluation results are presented in "Evaluation results" section, and discussed in "Discussion and managerial insights" section.

Evaluation approach

First, the prediction interval is evaluated. Second, the point estimators and prediction intervals are used to classify transitions into adherences and violations. The appropriate metrics for prediction interval and classification evaluation are introduced in the following. Additionally, the relative absolute error (RAE) is introduced to allow for comparability between models.

Evaluating predictions intervals

In order to objectively assess the quality of the prediction intervals, Khosravi et al. (2011) state that two performance metrics are required. First, the prediction interval coverage probability (PICP) is introduced, stated by the authors to be the most important characteristic of a prediction interval. It is calculated by dividing the number of predictions that are lower or equal to d^u, the given upper limit obtained from the total number of elements in the test set n_t. The PICP should not be lower than $1-\alpha$, the defined confidence level.

$$PICP = \frac{1}{n_t}\sum_{i=1}^{n_t}\hat{y}_i, \text{ with } \hat{y}_i = \begin{cases} 1, & \hat{y}_i \leq d_i^u \\ 0, & \hat{y}_i > d_i^u \end{cases} \tag{5}$$

When constructing a prediction interval, a trade-off has to be found between its width and its coverage probability as increasing the width leads to a higher coverage of predictions. Therefore, the mean prediction interval width (MPIW) is introduced. It is calculated by averaging all upper bounds d_i^u.

$$MPIW = \frac{1}{n_t}\sum_{i=1}^{n_t} d_i^u \tag{6}$$

Evaluating single model performance

Ultimately, the evaluation has to take place in the industrial setting by showing the value generated. By defining a required confidence level for the classification of future transitions, they can be classified by the model into violations and adherences. The goal is to maximize the correct predictions of violations and adherences. The resulting binary classification task can yield four different kinds of predictions: True Positives (TP), True Negatives (TN), False Negatives (FN), and False Positives (FP). This information can be summarized in a confusion matrix, holding all the relevant information of the classification performance (Grandini et al., 2020). Based on that, the traditional metrics of Recall ($\frac{TP}{TP+FN}$), Precision ($\frac{TP}{TP+FP}$) and Accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$) are used to evaluate the classification performance. They can be interpreted in this setting as follows: Recall can be calculated by dividing the number of violations that are detected by the total number of time constraint adherences. It can thus be interpreted as a measure for the detection rate of time constraints. The precision is a metric to measure the share of correctly predicted violations in relation to the total time constraint predicted violations. A high precision thus implies a low false alarm rate. To measure the share of total correct predictions, the accuracy is used.

Comparing multiple models

In order to compare multiple models, an additional metric is required. The absolute error (AE) of a predictor consists of the sum of the absolute residuals of its predictions. The AE of the examined model is therefore calculated by the sum of the absolute differences between the observed values y_i and the model's predictions $\hat{y}_i$. Similarly, the AE of a simple predictor is calculated by the sum of the absolute differences between the observed values y_i and the mean of the observed values $\overline{y}$. The RAE compares the AE of a model to that of a simple predictor always which always predicts the mean of the observations. It can therefore be interpreted as a measure of how much benefit a model provide. It is used to compare multiple models and calculated as described in Eq. (7).

$$RAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} |y_i - \overline{y}|} \tag{7}$$

Evaluation results

Using the performance metrics introduced in "Evaluation approach" section, the models are evaluated based on their point estimators, model uncertainties, and—given a confidence level—the resulting classifications into violations and adherences. The evaluation on the time constraint transitional data is performed using the previously mentioned test set. To obtain a more stable forecast, 500 predictions are made for each observation in the test set in order to form an ensemble prediction. The resulting point estimator yields a MSE of 0.8871 and a mean variance of the prediction error of 1.465 (Table 5).

The low recall of the resource-based model is caused by (in an example case) only one out of five violations being correctly detected. Due to the high false alarm rate, the precision is also low. However, when increasing the confidence level to 95%, four out of five violations are correctly identified leading to an increase in recall to 80% and a slight increase in precision to 0.763% indicating that a higher confidence level favors the resource based model.

The values of the PICP as well as the MPIW for a confidence level of 90% are shown in Table 6. Since the prediction intervals are only limited by an upper bound, higher observations in the test set result in a higher MPIW without effect on the prediction interval coverage probability. In order to compare the mean prediction interval widths, they are put into relation with the mean test observation (MTO).

The PICPs of the three evaluated models are all close to the prescribed confidence level of 90% indicating accurate predictions. At first glance, resource-based modeling and transitional modeling (GRU) approaches appear to have a higher variance than transitional modeling (LSTM) due to the increase in MPIW from transitional modeling (LSTM) to resource-based modeling. However, the ratio between MPIW and MTO are almost identical for all the models indicating equally variable predictions.

To regard the sensitivity of the proposed models it is required to transcend beyond the prescribed 90% level, although for the practical implementation in a given semiconductor fab the local information and selection must be re-adjusted. Figure 6 shows the influence over several time constrained lots that transition over a randomly selected high throughput transition in the given time period. It is worth noting the the time constraints vary in length, which influences the violation probability and thus performance on the confidence level. Nevertheless, it can be seen that the uncertainty inclusion in the models allows for a more thorough analysis of the time constraints than in rule-based approaches. The sensitivity shows that in the particular transition at that time a lower confidence level of approximately 80% would be sufficient to capture all actual violations. However, based on the presented chart and industrial expert validation a stronger risk aversion was chosen. It should be noted that in other transitions, in particular less frequently used transitions, the graph could look vastly different, requiring even higher confidence levels. Thus, as a trade-off 90% was selected.

Discussion and managerial insights

In general the presented models are capable of outperforming state-of-the-art approaches in avoiding time constraint violations as many observed time constraints could have been prevented. Nevertheless, the low precision shows that the models are hardly capable of separating close to violations from actual violations in an always effective manner. In a regular machine learning application, this could be attributed to a lack in generalization capability. However, in the given case, time series are used to condense the behavior within a factory. Clearly, future changes and never before observed circumstances cannot be learned from a time series if complex systems with many non-linearities, as in the semiconductor manufacturing example, are regarded. Much of this perceivable generalization gap can be attributed to the enormous complexity within a semiconductor wafer fab. Main factors are the high volume and ultra high mix production under averse manufacturing conditions with frequent breakdowns and stochastic process behavior and times. Overall, the results are a step towards better controlling product-inherent constraints in complex manufacturing environments such as semiconductor manufacturing operations.

Managerial implications that should be considered start with the precondition for managing product-inherent con-

Table 5 Accuracy, precision and recall for comparing the resource-based modeling approach and the transitional modeling approaches at a confidence level of 90%

	Resource-based approach	LSTM approach	GRU approach
Recall	20.0%	90.9%	42.9%
Precision	0.308	2.42	0.636
Accuracy	61.3%	82.1%	75.9%

Table 6 PICP and MPIW of resource-based modeling, transitional modeling (GRU), and transitional modeling (LSTM)

	Resource-based modeling	Transitional modeling (GRU)	Transitional modeling (LSTM)
MSE	0.4858	0.7280	0.4963
RAE	0.6957	0.8720	0.9996
Variance	0.5215	1.326	0.7553
PICP	0.8941	0.9256	0.9390
MPIW	9.677	8.977	7.671
MTO	8.866	7.876	6.589

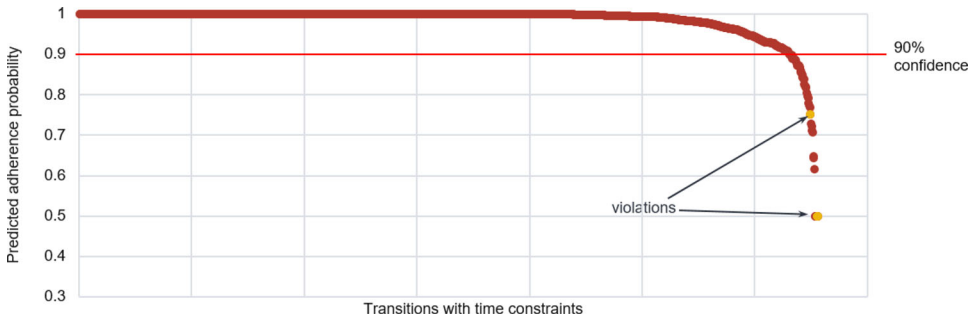

Fig. 6 Exemplary confidence influence on selected transition

straints, the need to identify and track individual products and lots. Hence, data gathering and traceability are key enablers and should be considered early on in an interconnected way. Additionally, the proposed approaches highlights the benefits that come from a shift from traditional operations management to a network perspective based on real-time and historical data. Similarly, the inclusion of uncertainty into decision making and possibility to regard individual, product-inherent constraints to advance well beyond traditional, average based decision making. Clearly, the benefit of decisions based on uncertainty informed artificial intelligence should be factored in when designing future production control.

Conclusions and further developments

This work presents a novel production control model for product-inherent constraints, on the example of time constraints in semiconductor manufacturing. Based on a transition time prediction, the probability of time constraint adherence is estimated with different real-time, data-based machine learning models by considering data and model based uncertainty. By using these modeling approaches, the heavy class imbalance in an industrial setting with less than 1% violations can be resolved. The evaluation shows that data-based production control strongly outperforms the industrial state-of-the-art with the transitional model. Selecting the best performing model in such volatile environments is not time invariant. The real-world industrial semiconductor fab application illustrates the need to move from static, long-term optimization towards data-based, real-time decision making for production control.

One limitation of this study is that it only considers simple time constraints that do not overlap or span several machines. While these account for the majority of time constraints as well as their respective violations, complex, nested time constraints should be included in a holistic approach and will be the focus of further, deeper developments. Additionally, the modeling can be extended with more advanced machine learning techniques that can incorporate uncertainty, such as bayesian networks. Moreover, the extension with both bigger and more exhaustive, i.e. more detailed data can propel the research in this domain. Lastly, the integration of digital-twin based simulations and research on their accuracy could be researched in the future.

Acknowledgements This research work was undertaken in the context of DIGIMAN4.0 project. DIGIMAN4.0 is a European Training Network supported by Horizon 2020, the EU Framework Programme for Research and Innovation (Project ID: 814225).

Author contributions All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

References

Abel, O., Helbig, A., Marquardt, W., Zwick, H., & Daszkowski, T. (2000). Productivity optimization of an industrial semi-batch polymerization reactor under safety constraints. *Journal of Process Control, 10*(4), 351–362.

Altenmüller, T., Stüker, T., Waschneck, B., Kuhnle, A., & Lanza, G. (2020). Reinforcement learning for an intelligent and autonomous production control of complex job-shops under time constraints. *Production Engineering, 14*(3), 319–328. https://doi.org/10.1007/s11740-020-00967-8

Amorim, P., Meyr, H., Almeder, C., & Almada-Lobo, B. (2013). Managing perishability in production-distribution planning: A discussion and review. *Flexible Services and Manufacturing Journal, 25*(3), 389–413.

An, Y. J., Kim, Y. D., & Choi, S. W. (2016). Minimizing makespan in a two-machine flowshop with a limited waiting time constraint and sequence-dependent setup times. *Computers & Operations Research, 71*, 127–136. https://doi.org/10.1016/j.cor.2016.01.017

Arima, S., Kobayashi, A., Wang, Y. F., Sakurai, K., & Monma, Y. (2015). Optimization of re-entrant hybrid flows with multiple queue time constraints in batch processes of semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing, 28*(4), 528–544. https://doi.org/10.1109/TSM.2015.2478281

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning* (pp. 115–123). PMLR.

Bixby, R., Burda, R., & Miller, D. (2006). Short-interval detailed production scheduling in 300mm semiconductor manufacturing using mixed integer and constraint programming. In *The 17th annual SEMI/IEEE ASMC 2006 conference* (pp. 148–154).

Chakravorty, S., & Nagarur, N. N. (2020). An artificial neural network based algorithm for real time dispatching decisions. In *2020 31st annual SEMI advanced semiconductor manufacturing conference (ASMC)* (pp. 1–5).

Chang, C. Y., & Chang, K. H. (2012). An integrated and improved dispatching approach to reduce cycle time of wet etch and furnace operations in semiconductor fabrication. In *Proceedings of the 2012 IEEE 16th international conference on computer supported cooperative work in design (CSCWD)* (pp. 734–741).

Chatfield, C. (2001). *Prediction intervals for time-series forecasting, principles of forecasting, 475–494*. Springer.

Chen, T., Sampath, V., May, M. C., Shan, S., Jorg, O. J., Aguilar Martín, J. J., Stamer, F., Fantoni, G., Tosello, G., & Calaon, M. (2023). Machine learning in manufacturing towards industry 4.0: From 'for now'to 'four-know'. *Applied Sciences, 13*(3), 1903. https://doi.org/10.3390/app13031903

Chien, C., & Chen, C. (2007). A novel timetabling algorithm for a furnace process for semiconductor fabrication with constrained waiting and frequency-based setups. *OR Spectrum, 29*(3), 391–419. https://doi.org/10.1007/s00291-006-0062-3

Cho, L., Park, H. M., Ryan, J. K., Sharkey, T. C., Jung, C., & Pabst, D. (2014). Production scheduling with queue-time constraints: Alternative formulations. In *IIE annual conference and expo 2014* (pp. 282–291).

Ciccullo, F., Pero, M., Pirovano, G., & Sianesi, A. (2014). Scheduling batches with time constraints in a job shop system: Developing two approaches for semiconductor industry. In *XIX Summer School "Francesco Turco"* (p. 12).

Eberle, L., Capón-García, E., Sugiyama, H., Graser, A., Schmidt, R., & Hungerbühler, K. (2016). Rigorous approach to scheduling of sterile drug product manufacturing. *Computers & Chemical Engineering, 94*, 221–234.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059). PMLR.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: An overview. arXiv preprint arXiv:2008.05756

Ham, M., Lee, Y. H., & An, J. (2011). Ip-based real-time dispatching for two-machine batching problem with time window constraints. *IEEE Transactions on Automation Science and Engineering, 8*(3), 589–597. https://doi.org/10.1109/TASE.2010.2098867

Han, J. H., & Lee, J. Y. (2023). Scheduling for a flow shop with waiting time constraints and missing operations in semiconductor manufacturing. *Engineering Optimization, 55*(10), 1742–1759.

Higgins, A. (2000). Adhesive bonding of aircraft structures. *International Journal of Adhesion and Adhesives, 20*(5), 367–376.

Hong, T. Y., Chien, C. F., & Chen, H. P. (2023). Unison framework of system dynamics-based technology acquisition decision for semiconductor manufacturing and an empirical study. *Computers & Industrial Engineering, 177*, 109012.

Huang, W. Y., Ke, L., & Shen, T. (2011). Quantify equipment capacity impacts induced by maximum waiting time constraint through simulation. In *2011 e-Manufacturing design collaboration symposium international symposium on semiconductor manufacturing (eMDC ISSM)* (pp. 1–3).

Jia, W., Jiang, Z., & Li, Y. (2013). Closed loop control-based real-time dispatching heuristic on parallel batch machines with incompatible job families and dynamic arrivals. *International Journal of Production Research, 51*(15), 4570–4584. https://doi.org/10.1080/00207543.2013.774505

Jung, C., Pabst, D., Ham, M., Stehli, M., & Rothe, M. (2014). An effective problem decomposition method for scheduling of diffusion processes based on mixed integer linear programming. *IEEE Transactions on Semiconductor Manufacturing, 27*(3), 357–363. https://doi.org/10.1109/TSM.2014.2337310

Kao, Y. T., Zhan, S. C., Chang, S. C., Ho, J. H., Wang, P., Luh, P. B., Wang, S., Wang, F., & Chang, J. (2011). Near optimal fur-

nance tool allocation with batching and waiting time constraints. In *2011 IEEE international conference on automation science and engineering* (pp. 108–113).

Khosravi, A., Nahavandi, S., Creighton, D., & Atiya, A. F. (2011). Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks, 22*(9), 1341–1356.

Kim, H. J., & Lee, J. H. (2017). A branch and bound algorithm for three-machine flow shop with overlapping waiting time constraints. *IFAC-PapersOnLine, 50*(1), 1101–1105. https://doi.org/10.1016/j.ifacol.2017.08.391

Kitamura, S., Mori, K., & Ono, A. (2006). Capacity planning method for semiconductor fab with time constraints between operations. In *2006 SICE-ICASE international joint conference* (pp. 1100–1103).

Klemmt, A., Horn, S., Weigert, G., & Hielscher, T. (2008). Simulations-based and solver-based optimization approaches for batch processes in semiconductor manufacturing. In *2008 winter simulation conference* (pp. 2041–2049).

Klemmt, A., & Monch, L. (2012). Scheduling jobs with time constraints between consecutive process steps in semiconductor manufacturing. In I. Staff (Ed.), *2012 winter simulation conference* (pp. 1–10). IEEE.

Kobayashi, A., Kuno, T., & Arima, S. (2013). Re-entrant flow control in q-time constraints processes for actual applications. In *2013 e-manufacturing design collaboration symposium (eMDC)* (pp. 1–4).

Kopp, D., Hassoun, M., Kalir, A., & Mönch, L. (2020). Integrating critical queue time constraints into smt2020 simulation models. In *2020 winter simulation conference (WSC)* (pp. 1813–1824).

Kuo, C. J., Chien, C. F., & Chen, J. D. (2011). Manufacturing intelligence to exploit the value of production and tool data to reduce cycle time. *IEEE Transactions on Automation Science and Engineering, 8*(1), 103–111. https://doi.org/10.1109/TASE.2010.2040999

Lee, J. (2020). A genetic algorithm for a two-machine flowshop with a limited waiting time constraint and sequence-dependent setup times. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2020/8833645

Li, L., Li, Y. F., & Sun, Z. J. (2012). Dispatching rule considering time-constraints on processes for semiconductor wafer fabrication facility. In *2012 IEEE international conference on automation science and engineering (CASE)* (pp. 407–412).

Liang, B., Turkcan, A., Ceyhan, M. E., & Stuart, K. (2015). Improvement of chemotherapy patient flow and scheduling in an outpatient oncology clinic. *International Journal of Production Research, 53*(24), 7177–7190.

Lima, A., Borodin, V., Dauzère-Pérès, S., & Vialletelle, P. (2017a). Analyzing different dispatching policies for probability estimation in time constraint tunnels in semiconductor manufacturing. In *2017 winter simulation conference (WSC)* (pp. 3543–3554).

Lima, A., Borodin, V., Dauzère-Pérès, S., & Vialletelle, P. (2017b). A decision support system for managing line stops of time constraint tunnels: Fa, i.e. In *2017 28th annual SEMI advanced semiconductor manufacturing conference (ASMC)* (pp. 309–314).

Lima, A., Borodin, V., Dauzère-Pérès, S., & Vialletelle, P. (2019). Sampling-based release control of multiple lots in time constraint tunnels. *Computers in Industry, 110*, 3–11. https://doi.org/10.1016/j.compind.2019.04.014

Lima, A., Borodin, V., Dauzère-Pérès, S., & Vialletelle, P. (2021). A sampling-based approach for managing lot release in time constraint tunnels in semiconductor manufacturing. *International Journal of Production Research, 59*(3), 860–884. https://doi.org/10.1080/00207543.2020.1711984

Maleck, C., & Eckert, T. (2017). A comparison of control methods for production areas with time constraints and tool interruptions in semiconductor manufacturing. In *2017 40th international spring seminar on electronics technology (ISSE)* (pp. 1–6).

Maleck, C., Nieke, G., Bock, K., Pabst, D., Schulze, M., & Stehli, M. (2019). A robust multi-stage scheduling approach for semiconductor manufacturing production areas with time contraints. In *2019 30th annual SEMI advanced semiconductor manufacturing conference (ASMC)* (pp. 1–6).

Maleck, C., Nieke, G., Bock, K., Pabst, D., & Stehli, M. (2018). A comparison of an cp and mip approach for scheduling jobs in production areas with time constraints and uncertainties. In *2018 winter simulation conference (WSC)* (pp. 3526–3537).

Maleck, C., Weigert, G., Pabst, D., & Stehli, M. (2017). Robustness analysis of an mip for production areas with time constraints and tool interruptions in semiconductor manufacturing. In *2017 winter simulation conference (WSC)* (pp. 3714–3725).

Markowich, P. A., Ringhofer, C. A., & Schmeiser, C. (2012). *Semiconductor equations*. Springer.

Mastrangelo, M., Magnanini, M. C., & Tolio, T. A. M. (2024). Control policy for production capacity modulation with waiting-time-constrained work in process. In L. Carrino, L. Galantucci, and L. Settineri (Eds.), *Selected topics in manufacturing: emerging trends from the perspective of AITeM's young researchers*, (Napoli, Italy, 13th–15th Sep. 2023) (pp. 159–175). Springer.

Mateus, B. C., Mendes, M., Farinha, J. T., Assis, R., & Cardoso, A. M. (2021). Comparing LSTM and GRU models to predict the condition of a pulp paper press. *Energies, 14*(21), 6958.

May, G. S., & Spanos, C. J. (2006). *Fundamentals of semiconductor manufacturing and process control* (pp. 1–463).

May, M. C., Albers, A., Fischer, M. D., Mayerhofer, F., Schäfer, L., & Lanza, G. (2021a). Queue length forecasting in complex manufacturing job shops. *Forecasting, 3*(2), 322–338. https://doi.org/10.3390/forecast3020021

May, M. C., Behnen, L., Holzer, A., Kuhnle, A., & Lanza, G. (2021b). Multi-variate time-series for time constraint adherence prediction in complex job shops. *Procedia CIRP, 103*, 55–60. https://doi.org/10.1016/j.procir.2021.10.008

May, M. C., Maucher, S., Holzer, A., Kuhnle, A., & Lanza, G. (2021c). Data analytics for time constraint adherence prediction in a semiconductor manufacturing use-case. *Procedia CIRP, 100*, 49–54. https://doi.org/10.1016/j.procir.2021.05.008

May, M. C., Glatter, D., Arnold, D., Pfeffer, D., & Lanza, G. (2024). Iiot system canvas-from architecture patterns towards an iiot development framework. *Journal of Manufacturing Systems, 72*, 437–459.

May, M. C., Kiefer, L., Kuhnle, A., & Lanza, G. (2022). Ontology-based production simulation with ontologysim. *Applied Sciences, 12*(3), 1608. https://doi.org/10.3390/app12031608

Mönch, L., Fowler, J. W., Dauzère-Pérès, S., Mason, S. J., & Rose, O. (2009). Scheduling semiconductor manufacturing operations: Problems, solution techniques, and future challenges. In *4th multidisciplinary international conference on scheduling: theory & applications*. Citeseer.

Mönch, L., Fowler, J. W., Dauzère-Pérès, S., Mason, S. J., & Rose, O. (2011). A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations. *Journal of Scheduling, 14*(6), 583–599. https://doi.org/10.1007/s10951-010-0222-9

Mönch, L., Fowler, J. W., & Mason, S. J. (2013). *Production planning and control for semiconductor wafer fabrication facilities: Modeling, analysis, and systems* (Vol. 52). Springer.

Nattaf, M., Dauzère-Pérès, S., Yugma, C., & Wu, C. H. (2019). Parallel machine scheduling with time constraints on machine qualifications. *Computers & Operations Research, 107*, 61–76. https://doi.org/10.1016/j.cor.2019.03.004

Ono, A., Kitamura, S., & Mori, K. (2006). Risk based capacity planning method for semiconductor fab with queue time constraints. In *2006*

IEEE international symposium on semiconductor manufacturing (pp. 49–52).

Pappert, F. S., Zhang, T., Rose, O., Suhrke, F., Mager, J., & Frey, T. (2016). Impact of time bound constraints and batching on metallization in an opto-semiconductor fab. In *2016 winter simulation conference (WSC)* (pp. 2947–2957).

Perraudat, A., Lima, A., Dauzère-Pérès, S., & Vialletelle, P. (2019). A decision support system for a critical time constraint tunnel. In *2019 30th annual SEMI advanced semiconductor manufacturing conference (ASMC)* (pp. 1–5).

Pirovano, G., Ciccullo, F., Pero, M., & Rossi, T. (2020). Scheduling batches with time constraints in wafer fabrication. *International Journal of Operational Research, 37*(1), 1–31. https://doi.org/10.1504/IJOR.2020.104222

Sadeghi, R., Dauzère-Pérès, S., Yugma, C., & Lepelletier, G. (2015). Production control in semiconductor manufacturing with time constraints. In *2015 26th annual SEMI advanced semiconductor manufacturing conference (ASMC)* (pp. 29–33).

Somboonwiwat, T., Khompatraporn, C., Miengarrom, T., & Lerdluechachai, K. (2018). A bi-objective environmental-economic optimisation of hot-rolled steel coils supply chain: a case study in Thailand. *Advances in Production Engineering & Management, 13*(1), 93–106.

STMicroelectronics. (2000). Introduction to semiconductor technology.

Su, L. H. (2003). A hybrid two-stage flowshop with limited waiting time constraints. *Computers & Industrial Engineering, 44*(3), 409–424. https://doi.org/10.1016/S0360-8352(02)00216-4

Sun, D. S., Choung, Y. I., Lee, Y. J., & Jang, Y. C. (2005). Scheduling and control for time-constrained processes in semiconductor manufacturing. In ISSM 2005. *IEEE international symposium on semiconductor manufacturing* (pp. 295–298).

Sundaramoorthy, A., & Karimi, I. (2004). Planning in pharmaceutical supply chains with outsourcing and new product introductions. *Industrial & Engineering Chemistry Research, 43*(26), 8293–8306.

Tu, Y., Chen, H., & Liu, T. (2010). Shop-floor control for batch operations with time constraints in wafer fabrication. *International Journal of Industrial Engineering: Theory Applications and Practice, 17*(2), 142–155.

Tu, Y. M., & Chen, C. L. (2011). Model to determine the capacity of wafer fabrications for batch-serial processes with time constraints. *International Journal of Production Research, 49*(10), 2907–2923. https://doi.org/10.1080/00207541003730854

Tu, Y. M., & Chen, H. N. (2009a). Capacity planning with sequential two-level time constraints in the back-end process of wafer fabrication. *International Journal of Production Research, 47*(24), 6967–6979. https://doi.org/10.1080/00207540802415568

Tu, Y. M., & Chen, H. N. (2009b). Tool portfolio planning in the back-end process of wafer fabrication with sequential time constraints. *Journal of the Chinese Institute of Industrial Engineers, 26*(1), 60–69. https://doi.org/10.1080/10170660909509122

Tu, Y. M., & Chen, H. N. (2010). Capacity planning with sequential time constraints under various control policies in the back-end of wafer fabrications. *Journal of the Operational Research Society, 61*(8), 1258–1264. https://doi.org/10.1057/jors.2009.36

Tu, Y. M., & Liou, C. S. (2006). Capacity determination model with time constraints and batch processing in semiconductor wafer fabrication. *Journal of the Chinese Institute of Industrial Engineers, 23*(3), 192–199. https://doi.org/10.1080/10170660609509008

Uzsoy, R., Lee, C. Y., & Martin-Vega, L. A. (1992). A review of production planning and scheduling models in the semiconductor industry Part I. System characteristics, performance evaluation and production planning. *IIE Transactions, 24*(4), 47–60.

Valet, A., Altenmüller, T., Waschneck, B., May, M. C., Kuhnle, A., & Lanza, G. (2022). Opportunistic maintenance scheduling with deep reinforcement learning. *Journal of Manufacturing Systems, 64*, 518–534. https://doi.org/10.1016/j.jmsy.2022.07.016

Wang, C., & Liu, X. B. (2013). Integrated production planning and control: A multi-objective optimization model. *Journal of Industrial Engineering and Management (JIEM), 6*(4), 815–830.

Wang, H. K., Chien, C. F., & Gen, M. (2014). Hybrid estimation of distribution algorithm with multiple subpopulations for semiconductor manufacturing scheduling problem with limited waiting-time constraint. In *2014 IEEE international conference on automation science and engineering (CASE)* (pp. 101–106).

Wang, H. K., Chien, C. F., & Gen, M. (2015). An algorithm of multi-subpopulation parameters with hybrid estimation of distribution for semiconductor scheduling with constrained waiting time. *IEEE Transactions on Semiconductor Manufacturing, 28*(3), 353–366. https://doi.org/10.1109/TSM.2015.2439054

Wang, M., Srivathsan, S., Huang, E., & Wu, K. (2018). Job dispatch control for production lines with overlapped time window constraints. *IEEE Transactions on Semiconductor Manufacturing, 31*(2), 206–214. https://doi.org/10.1109/TSM.2018.2826530

Waschneck, B., Altenmüller, T., Bauernhansl, T., & Kyek, A. (2016). Production scheduling in complex job shops from an industry 4.0 perspective: A review and challenges in the semiconductor industry. *SAMI iKNOW*, 1–12.

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems, 22*(1), 45–55.

Wu, C. H., Cheng, Y. C., Tang, P. J., & Yu, J. Y. (2012a). Optimal batch process admission control in tandem queueing systems with queue time constraint considerations. In *Proceedings of the 2012 winter simulation conference (WSC)* (pp. 1–6).

Wu, C. H., Lin, J. T., & Chien, W. C. (2012b). Dynamic production control in parallel processing systems under process queue time constraints. *Computers & Industrial Engineering, 63*(1), 192–203. https://doi.org/10.1016/j.cie.2012.02.003

Wu, C. H., Chien, W. C., Chuang, Y. T., & Cheng, Y. C. (2016a). Multiple product admission control in semiconductor manufacturing systems with process queue time (PQT) constraints. *Computers & Industrial Engineering, 99*, 347–363. https://doi.org/10.1016/j.cie.2016.04.003

Wu, K., Zhao, N., Gao, L., & Lee, C. (2016b). Production control policy for tandem workstations with constant service times and queue time constraints. *International Journal of Production Research, 54*(21), 6302–6316. https://doi.org/10.1080/00207543.2015.1129468

Wu, C. H., Lin, J. T., & Chien, W. C. (2010). Dynamic production control in a serial line with process queue time constraint. *International Journal of Production Research, 48*(13), 3823–3843. https://doi.org/10.1080/00207540902922836

Wurster, M., Michel, M., May, M. C., Kuhnle, A., Stricker, N., & Lanza, G. (2022). Modelling and condition-based control of a flexible and hybrid disassembly system with manual and autonomous workstations using reinforcement learning. *Journal of Intelligent Manufacturing*, 1–17.

Xiao, H. (2012). *Introduction to semiconductor manufacturing*. SPIE Press.

Yamak, P. T., Yujian, L., & Gadosey, P. K. (2019). A comparison between arima, lstm, and gru for time series forecasting. In *Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence* (pp. 49–55).

Yang, K. T., Ke, L., & Shen, T. (2015). Modeling and dispatching refinement for implantation to reduce the probability of tuning beam. In *2015 26th annual SEMI advanced semiconductor manufacturing conference (ASMC)* (pp. 190–194).

Yin, M., Huang, M., Qian, X., Wang, D., Wang, X., & Lee, L. H. (2021). Fourth-party logistics network design with service time constraint

under stochastic demand. *Journal of Intelligent Manufacturing*, 1–25 .

Yu, T. S., Kim, H. J., Jung, C., & Lee, T. E. (2013). Two-stage lot scheduling with waiting time constraints and due dates. In *2013 winter simulations conference (WSC)* (pp. 3630–3641).

Yu, T. S., Kim, H. J., & Lee, T. E. (2017). Minimization of waiting time variation in a generalized two-machine flowshop with waiting time constraints and skipping jobs. *IEEE Transactions on Semiconductor Manufacturing, 30*(2), 155–165. https://doi.org/10.1109/TSM.2017.2662231

Yuan, S., Li, T., & Wang, B. (2021). A discrete differential evolution algorithm for flow shop group scheduling problem with sequence-dependent setup and transportation times. *Journal of Intelligent Manufacturing, 32*, 427–439.

Yugma, C., Dauzère-Pérès, S., Artigues, C., Derreumaux, A., & Sibille, O. (2012). A batching and scheduling algorithm for the diffusion area in semiconductor manufacturing. *International Journal of Production Research, 50*(8), 2118–2132. https://doi.org/10.1080/00207543.2011.575090

Yurtsever, T., Kutanoglu, E., & Johns, J. (2009). Heuristic based scheduling system for diffusion in semiconductor manufacturing. In *Proceedings of the 2009 winter simulation conference (WSC)*, (pp. 1677–1685).

Zarzycki, K., & Ławryńczuk, M. (2021). LSTM and GRU neural networks as models of dynamical processes used in predictive control: A comparison of models developed for two chemical reactors. *Sensors, 21*(16), 5625.

Zhang, T., Pappert, F. S., & Rose, O. (2016). Time bound control in a stochastic dynamic wafer fab. In *2016 winter simulation conference (WSC)* (pp. 2903–2911).

Zhou, L., Lin, C., Hu, B., & Cao, Z. (2019). A cuckoo search-based scheduling algorithm for a semiconductor production line with constrained waiting time. In *2019 IEEE 15th international conference on automation science and engineering (CASE)* (pp. 338–343).

Zhou, Y., & Wu, K. (2017). Heuristic simulated annealing approach for diffusion scheduling in a semiconductor fab. In *2017 IEEE/ACIS 16th international conference on computer and information science (ICIS)* (pp. 785–789).

Zhu, L., & Laptev, N. (2017). Deep and confident prediction for time series at uber. In *2017 IEEE international conference on data mining workshops (ICDMW)* (pp. 103–110). IEEE.

Ziarnetzky, T., Mönch, L., Ponsignon, T., & Ehm, H. (2017). Rolling horizon planning with engineering activities in semiconductor supply chains. In *2017 13th IEEE conference on automation science and engineering (CASE)* (pp. 1024–1025). IEEE.