*Proceedings of the 2025 Winter Simulation Conference*
*E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.*

# PREDICTIVE MODELING IN SEMICONDUCTOR MANUFACTURING: A SYSTEMATIC REVIEW OF CYCLE TIME PREDICTION, ANOMALY DETECTION, AND DIGITAL TWIN INTEGRATION

Claude Yugma[1], Adrien Wartelle[1], Stéphane Dauzère-Pérès[1], Pascal Robert[2], Renaud Roussel[2], Jasper van Heugten[3], and Taki-Eddine Korabi[3]

[1] Mines Saint-Étienne, Univ. Clermont Auvergne, CNRS, UMR 6158 LIMOS, Gardanne, FRANCE
[2]STMicroelectronics, Crolles, FRANCE
[3]Minds.ai, Santa Cruz, CA, USA

## ABSTRACT

Accurate cycle time prediction is critical for optimizing throughput, managing WIP, and ensuring responsiveness in semiconductor manufacturing. This systematic review synthesizes literature from Google Scholar, Web of Science, and IEEE Xplore, covering analytical, statistical, AI-driven, and hybrid approaches. The key contributions are: (1) a structured, comparative evaluation of predictive techniques in terms of accuracy, interpretability, and scalability, and (2) identification of research gaps and emerging directions, such as self-adaptive models, generative AI for data augmentation, and enhanced human-AI collaboration. This review provides insights to support the development of robust forecasting systems aligned with the evolving demands of semiconductor manufacturing.

## 1 INTRODUCTION

The semiconductor manufacturing industry serves as the backbone of modern technological advancement, driving progress across electronics, computing, telecommunications, and other high-tech sectors. With surging global demand fueled by the proliferation of smart devices, artificial intelligence (AI), and the Internet of Things (IoT) manufacturers face mounting pressure to optimize operations, reduce costs, and boost responsiveness to remain competitive (Monostori 2014).

Among the core challenges is accurately predicting cycle time (CT) which, in the case of a wafer fabrication factory (wafer fab), the total time it takes to transform one lot of wafer of silicon into one lot of integrated circuits (IC). This key performance indicator closely tied to the estimation of lead time, e.g. the time between a customer order and its delivery, the management of lot Work-in-Progress (WIP), e.g the occupation level of a wafer fab, lot throughput, and customer satisfaction. CT prediction is tough due to the industry's high variability and reentrant production flows. Reliable CT estimates enable better resource allocation, early identification of bottlenecks, and more agile responses to market fluctuations an imperative underscored by recent global component shortages (Ivanov and Dolgui 2020).

Traditional modeling methods often fail to capture the inherent complexity of semiconductor production. Processes are multi-stage, non-linear, and subject to numerous stochastic influences, including equipment behavior, operator variability, and product mix changes. Moreover, massive volumes of data generated throughout the manufacturing pipeline exceed the capabilities of conventional analytical techniques. This has spurred interest in advanced modeling approaches, notably those based on AI, machine learning (ML), and Digital Twin (DT) technologies (Kang et al. 2016).

Digital Twins, in particular, have emerged as powerful tools, offering real-time synchronization between physical systems and virtual counterparts. Using various degrees of abstraction, these digital models replicate the measured behaviors of a real wafer fab, enabling dynamic forecasting, anomaly detection, and process optimization by continuously integrating real-time production data (Qi et al. 2021).

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

Despite these technological advances, few studies have holistically examined the synergy between CT predictive modeling, real-time anomaly detection, and Digital Twin implementation in real-world manufacturing environments. Existing literature often isolates these components, overlooking their potential integration (Chang and Liao 2006). This review aims to bridge that gap through a structured synthesis of predictive methods and enabling technologies.

The primary objectives of this review are threefold: (1) to evaluate traditional and AI-based CT prediction methods, including Multiple-Factor Linear Combination (MFLC), Artificial Neural Networks (ANN), and Deep Learning; (2) to examine the role of Digital Twins in supporting adaptive, real-time prediction; and (3) to assess the practical performance, scalability, and limitations of various approaches in real manufacturing settings.

The remainder of this paper is organized as follows: Section 2 outlines the systematic review methodology. Section 3 discusses key predictive methods. Section 4 explores Digital Twin integration. Section 5 presents model validation strategies. Section 6 examines influential scientific articles and trends in Explainable AI (XAI). Sections 7 synthesize findings, address current challenges, and propose future research directions and we end the article with Section 8.

## 2 SYSTEMATIC REVIEW METHODOLOGY

This section outlines the systematic methodology employed to collect, screen, and analyze relevant literature on cycle time (CT) prediction in semiconductor manufacturing. The review process was designed to ensure thorough coverage, replicability, and transparency across four main stages: (1) search strategy formulation, (2) article selection and screening, (3) temporal and geographical analysis, and (4) methodological categorization.

### 2.1 Search Strategy

The search strategy was built upon both direct and snowballing methods to ensure a comprehensive collection of relevant studies. Direct searches were performed using targeted keyword queries on leading academic databases, including IEEE Xplore, ScienceDirect, SpringerLink, Taylor & Francis, Web of Science, and Google Scholar. The primary keywords combination used were "cycle time prediction AND semiconductor manufacturing," and "lot completion time prediction AND wafer fabrication." To maintain relevance, the scope was limited to publications from 2000 onward.

To complement the database queries, a backward and forward citation analysis (snowballing) was conducted using seminal papers identified in the initial screening. This helped capture influential studies potentially missed by database search filters, thereby increasing coverage.

### 2.2 Article Selection and Screening

The article selection process began with an initial set of 54 publications identified through title and abstract screening across major databases. Each article was then assessed in detail against a set of clearly defined inclusion and exclusion criteria to ensure its relevance, methodological rigor, and contribution to the field of cycle time (CT) prediction in semiconductor manufacturing. To be considered for inclusion, studies had to focus specifically on CT prediction, or a directly connected prediction such as those of remaining CT, completion times and due dates, within the semiconductor domain, it had be published in reputable peer-reviewed journals or high-impact conferences, and include some form of empirical validation such as simulation experiments, industrial datasets, or applied case studies. Furthermore, to reflect current developments, only articles published from the year 2000 onward were eligible.

Conversely, papers were excluded if they addressed unrelated manufacturing processes, lacked methodological depth or empirical evidence, or were identified as redundant or preliminary versions of already included work. After applying these criteria, the dataset was refined into three categories. A total of 29 articles were fully retained as they satisfied all inclusion criteria and demonstrated clear empirical contributions. Fourteen articles were labeled as borderline, requiring further detailed examination due to partial compliance or methodological ambiguity. The remaining 11 articles were excluded on grounds of irrelevance or insufficient rigor.

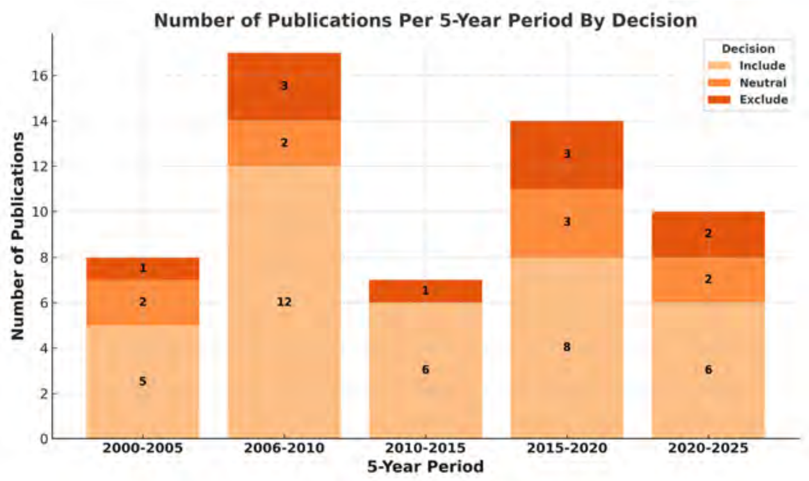

Figure 1: Number of publications per 5-year period categorized by inclusion status.

## 2.3 Temporal and Geographical Trends

An examination of the selected literature across time periods revealed a clear evolution in methodological focus. Between 2000 and 2010, research was primarily centered on foundational statistical and simulation techniques, laying the groundwork for later innovations. The following decade, from 2010 to 2020, saw a marked shift toward artificial intelligence, with increasing adoption of artificial neural networks (ANNs) and fuzzy logic to capture complex relationships in dynamic manufacturing systems. More recently, the period from 2020 to 2025 has been characterized by the integration of hybrid models, the emergence of Digital Twin frameworks, and a growing interest in Explainable AI (XAI), reflecting the industry's need for real-time, interpretable, and resilient predictive tools.

From a geographical perspective, the distribution of contributions shows that research activity is concentrated in several global innovation hubs. As illustrated in Figure 2, the majority of publications originate from China, South Korea, and the United States. These countries have established themselves as leaders in semiconductor research, benefiting from strong academic-industry collaboration, advanced manufacturing capabilities, and strategic national investments in AI and smart manufacturing technologies.

## 2.4 Categorization of Methods and Research Goals

The selected articles were further analyzed and classified according to two essential dimensions: the methodological approach employed and the research objectives pursued. This dual categorization helps clarify the diverse modeling paradigms adopted across the literature and reveals distinct thematic patterns in CT prediction research.

In terms of methodology, the articles reflected a wide spectrum of modeling strategies. Some studies relied on analytical or queueing models to derive theoretical insights into production dynamics. Others employed statistical techniques such as regression analysis or time-series forecasting to model cycle time based on historical data trends. A significant portion of the literature focused on AI-based methods, including artificial neural networks, deep learning architectures, and fuzzy logic systems, particularly in scenarios requiring non-linear pattern recognition. Case-Based Reasoning (CBR) approaches were also observed, leveraging past operational cases to predict current outcomes. Many of the most recent contributions

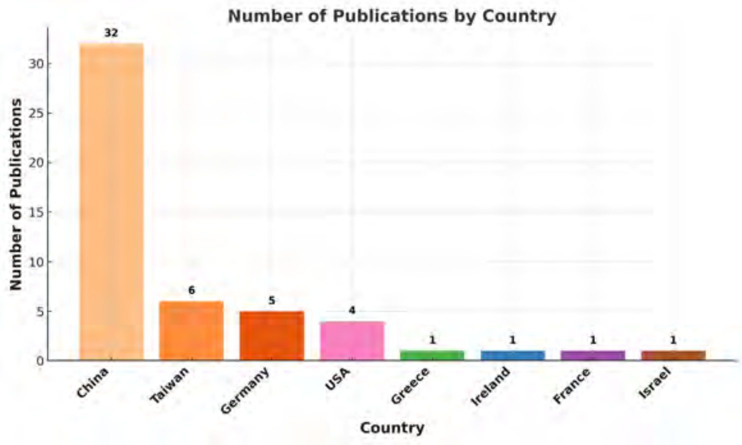

Figure 2: Geographical distribution of publications.

embraced hybrid models, which combine multiple techniques to balance accuracy, interpretability, and adaptability. Finally, simulation-based methods remained popular, offering rich, event-driven representations of production systems.

From the perspective of research objectives, studies typically targeted one or more of the following: prediction of mean full cycle time for completed lots, estimation of remaining cycle time for in-process lots, forecasting the distribution of cycle times to assess variability, predicting lot output quantities, and supporting due-date assignment through upstream CT estimation. These objectives reflect both operational and strategic decision needs within the semiconductor manufacturing environment.

Together, these classifications provide a structured lens through which the subsequent sections analyze modeling performance, integration with Digital Twins, and pathways for industrial implementation.

## 3 PREDICTION METHODS IN SEMICONDUCTOR MANUFACTURING

This section examines the main categories of predictive methods used for cycle time (CT) estimation in semiconductor manufacturing. The literature reveals seven key methodological families: analytical models, statistical techniques, artificial neural networks (ANNs), case-based reasoning (CBR), fuzzy modeling, hybrid approaches, and simulation. Each is described below with its principles, strengths, and limitations.

### 3.1 Analytical Models

Analytical approaches use mathematical and queueing theory formulations to estimate CT, often under simplifying assumptions. Models such as M/M/1 or G/G/1 queues help estimate waiting times and throughput based on arrival and service distributions. Little's Law, which relates average WIP, throughput, and CT, is frequently applied (Little 1961). These models offer interpretability and low computational cost but typically assume system stationarity and struggle with complex dynamics in modern fabs.

### 3.2 Statistical Techniques

Statistical models rely on historical production data to generate regression-based or time-series forecasts. Multiple-Factor Linear Combination (MFLC) models and ARIMA (AutoRegressive Integrated Moving Average) are common examples. While they provide transparency and ease of implementation, their assumptions about linearity and stationarity limit performance in environments with high variability or non-linear dependencies (Wang et al. 2018).

### 3.3 Artificial Neural Networks (ANNs)

ANNs are data-driven models inspired by biological neural networks. They are well-suited for capturing non-linear relationships and learning complex input-output mappings. Backpropagation neural networks are commonly applied for CT prediction. For example, as shown in Chen (2007) and Wang et al. (2018) work, BPNs outperform traditional statistical methods and even some rule-based systems in CT forecasting accuracy, especially when trained on high-dimensional process data. More recently, deep learning variants such as CNNs and RNNs have shown superior performance in recognizing temporal and spatial patterns. However, these models often require large datasets and computational resources, and may lack interpretability (Wang et al. 2018).

### 3.4 Case-Based Reasoning (CBR)

CBR systems retrieve similar past production cases to solve current prediction problems through analogy. This method excels in scenarios with strong historical precedent and can adapt to changes in production context. Notable examples include CBR systems enhanced by expert systems or combined with k-nearest neighbor algorithms (Chang et al. 2001). The main limitation lies in the need for well-maintained case libraries and representative data.

### 3.5 Fuzzy Modeling

Fuzzy logic handles uncertainty and imprecision by allowing variables to belong to multiple fuzzy sets with varying degrees of truth. It is useful for modeling vague concepts such as "high congestion" or "low machine reliability." Fuzzy clustering and fuzzy inference systems have been used to estimate CT under uncertain conditions (Chen 2008). While this approach handles noise and variability well, interpretability may degrade in high-dimensional or rule-intensive systems (Lee and Gao 2021).

### 3.6 Hybrid Approaches

Hybrid methods combine two or more modeling paradigms to capitalize on their respective advantages. For instance, integrating fuzzy systems with neural networks or simulation allows improved adaptability and accuracy. Chen (2007) presented a model combining k-means clustering and fuzzy BPNs, while Seidel et al. (2020) showed how hybrid models integrated into digital twins enhance predictive capability and real-time decision support.

### 3.7 Simulation-Based Methods

Simulation techniques such as discrete-event simulation (DES) model the detailed behavior of manufacturing processes over time. These models enable scenario testing and CT distribution analysis. DES is highly accurate but resource-intensive to develop and calibrate (Morrison and Martin 2007). Recent work also includes continuous simulation for flow-based processes and agent-based models for system-level behavior (Deenen et al. 2024). Simulation offers deep insight but often lacks real-time responsiveness and scalability.

Each method has its trade-offs in terms of accuracy, interpretability, data requirements, and scalability. Table 1 summarizes their comparative strengths.

## 4 INTEGRATION OF DIGITAL TWINS

Digital Twin (DT) technology represents a transformative innovation in semiconductor manufacturing by enabling real-time synchronization between physical assets and their virtual counterparts. This section details how DTs enhance cycle time (CT) prediction through continuous data integration, predictive accuracy, and dynamic decision support. We discuss three primary areas where Digital Twins are being applied: real-time CT forecasting, lot quantity and due-date estimation, and continuous model calibration and extrapolation.

Table 1: Comparison of predictive modeling approaches.

| Method | Accuracy | Interpretability | Scalability | Data Requirement |
|---|---|---|---|---|
| Analytical Models | Low/Medium | High | High | Low |
| Statistical Models | Medium | Medium/High | Medium/High | Medium |
| ANNs | High | Low | High | High |
| CBR | Medium | Medium | Medium | Medium |
| Fuzzy Models | Medium | Medium | Medium | Medium |
| Hybrid Approaches | High | Medium | Medium | High |
| Simulation | High | High | Low | High |

### 4.1 Real-Time Cycle Time Forecasting

Digital Twins facilitate real-time CT forecasting, particularly for remaining CT, by integrating data from IoT devices, sensors, and manufacturing execution systems (MES). This data-driven environment allows for dynamic simulation and prediction that reflect current shop-floor conditions. For example, Tao et al. (2019) demonstrated that real-time data collected from photolithography tools significantly improves the responsiveness of predictive models. The continuous feedback loop between physical operations and digital replicas supports anomaly detection and rapid schedule adjustments, reducing delays and improving throughput.

### 4.2 Lot Quantity and Due-Date Estimation

Accurately forecasting lot completion quantities and due dates is vital for production planning and customer satisfaction. Digital Twins improve these estimations by leveraging current and historical production data. In highly variable environments, such as those with frequent equipment changes or product mix variability, DTs dynamically adjust their forecasts. Chen, Lin, and Lin (2024) introduced a fuzzy collaborative DT system that updates due-date estimates in real time based on fluctuating WIP levels and capacity changes, demonstrating higher reliability compared to static models.

### 4.3 Model Calibration and Scenario Extrapolation

One of the most powerful features of Digital Twins is their ability to maintain predictive accuracy over time through continuous calibration. As new production data is ingested, model parameters are updated to reflect operational changes. This capability ensures sustained reliability even in evolving manufacturing contexts. Moreover, DTs enable predictive extrapolation—testing “what-if” scenarios such as equipment upgrades, new product introductions, or extreme throughput constraints. Deenen et al. (2024) showed how DTs define validity zones and extend forecasting capabilities into unobserved or future production scenarios.

### 4.4 Implementation Challenges and Research Opportunities

Despite their promise, implementing Digital Twins in semiconductor fabs presents several challenges:

- **Data Quality and Integration**: High-fidelity, synchronized data across multiple systems is a prerequisite for effective DTs. Missing or misaligned data can compromise model reliability.
- **Computational Infrastructure**: Real-time simulation and prediction require substantial computational resources and efficient data pipelines.
- **Technical Expertise**: Integrating AI models, IoT systems, and simulation engines into a coherent DT framework demands specialized interdisciplinary skills.

Future research should focus on overcoming these hurdles by exploring edge computing for real-time responsiveness, developing lightweight DT architectures for resource-constrained environments, and enhancing explainability to support operator trust and human-AI collaboration.

Digital Twins significantly improve CT prediction accuracy, adaptability, and decision-making in semiconductor manufacturing. They support responsive, data-driven production planning but require robust infrastructure and expertise. Their integration marks a shift toward smart, resilient factories capable of adapting to continuous change.

## 5 MODEL ANALYSIS AND VALIDATION

Evaluating the performance and reliability of predictive models is a critical step in ensuring their suitability for real-world semiconductor manufacturing environments. This section examines the different dimensions of model validation, including the metrics used to assess predictive performance, the protocols designed to test generalizability, and the practical considerations necessary for successful implementation.

To assess predictive accuracy, researchers typically rely on quantitative error metrics that compare predicted and actual cycle time values. Among the most widely used is the Mean Squared Error (MSE), which amplifies larger prediction deviations by squaring the error terms, thereby penalizing models that produce substantial outliers. Closely related is the Root Mean Squared Error (RMSE), which retains the benefits of MSE but expresses the error in the same units as the original data, offering more intuitive interpretability for practitioners. The Mean Absolute Error (MAE) is another critical indicator, particularly useful in environments with significant noise or uncertainty, as it measures the average magnitude of prediction errors without over-penalizing large deviations. The Mean Absolute Percentage Error (MAPE) and its symmetric version (SMAPE) offer a relative error measure providing a more meaningful interpretation of the scale of the model quality performances. These metrics are commonly used in semiconductor applications, as seen in Wang et al. (2018) and Chen (2007) works. In addition, researchers often examine the distribution of residual errors to detect systematic bias or instability an approach recommended in robust model development frameworks like those discussed in Meidan et al. (2011) work.

However, evaluating performance on historical data alone is not sufficient. Cross-validation techniques, such as k-fold cross-validation, are essential for assessing a model's ability to generalize to unseen data. This method is widely adopted in the predictive modeling literature, including Lee and Gao (2021) work, where hybrid models are rigorously tested across multiple folds. More advanced validation protocols include empirical testing in live or near-live environments, as emphasized in Seidel et al. (2020) work, where simulation-integrated Digital Twins are tested against real fab operations. These validations account for practical constraints like data latency, missing values, and shifting process dynamics, supporting the concept of validity zones a method detailed in Deenen et al. (2024) work to ensure predictions remain reliable under defined operational conditions.

Model validation also entails evaluating operational aspects beyond statistical accuracy. One key consideration is computational efficiency. In high-throughput fabs, predictions must be generated quickly, as discussed in Chen (2008) work, which highlights real-time forecasting requirements for lot output times. Additionally, interpretability is crucial: while deep learning models like CNNs and RNNs offer high accuracy (see Wang, Zhang, and Wang (2018)), their black-box nature can hinder adoption unless paired with explainability tools—a challenge explored in (Meidan et al. 2011).

Data quality is another major factor influencing model reliability. Semiconductor data often suffer from missing values, noise, and inconsistencies due to sensor or logging issues. Effective preprocessing, including outlier detection and normalization, is emphasized in Chen (2007) and Wang, Zhang, and Wang (2018) works, where feature selection and data cleaning led to significant accuracy improvements.

Scalability concerns are addressed in studies like Seidel et al. (2020) and Deenen et al. (2024) works, where predictive systems are tested across multiple fabs or production segments. These implementations often involve ensembles of models, each tailored to specific tools or products, supported by monitoring systems and automated retraining protocols.

Finally, successful deployment depends on human-AI collaboration. Studies like Chen (2007) and Wang, Zhang, and Wang (2018) emphasize the need for interfaces that offer confidence levels, alert prioritization, and explainable insights, helping engineers and planners integrate predictions into their workflows. This interaction between models and users is central to achieving long-term value.

A holistic model validation process encompasses statistical performance, cross-scenario robustness, and operational feasibility. These considerations ensure that CT prediction models can deliver reliable, actionable insights in the dynamic environment of semiconductor manufacturing.

## 6 EXPLORATION OF SCIENTIFIC ARTICLES

This section explores key scientific contributions that have shaped the development of predictive modeling for cycle time estimation in semiconductor manufacturing. The analysis highlights both historically influential articles and emerging trends that signal the future direction of research.

Among the most impactful studies, Wang, Zhang, and Wang (2018) proposed a domain-specific data-driven model for cycle time prediction, incorporating feature selection tailored to semiconductor fabrication. Their methodology significantly improved prediction accuracy and demonstrated the importance of incorporating industry-specific knowledge into model design. Similarly, Chen (2007) developed a hybrid framework that integrated statistical learning with neural networks to predict wafer lot output time. Their work showcased the advantages of combining multiple techniques to enhance model robustness and applicability across diverse production settings.

Anomaly detection has also received considerable attention, particularly through hybrid approaches. For instance, Lee and Gao (2021) presented a fuzzy clustering-based system combined with genetic algorithms and machine learning techniques. Their method proved highly effective in identifying deviations and providing stable CT forecasts in the presence of noisy or uncertain data. These developments underscore the growing importance of robustness and adaptability in prediction systems.

The emergence of Digital Twins has marked a pivotal shift in modeling strategies. Tao, Qi, Wang, and Nee (2019) provided foundational work on Digital Twin applications for manufacturing, illustrating how virtual-physical synchronization could be leveraged for real-time CT prediction and equipment monitoring. Their framework laid the groundwork for more recent studies that tightly integrate simulation, sensor data, and AI to deliver predictive and prescriptive analytics. Likewise, Chen and Wang (2022) surveyed hybrid modeling approaches and emphasized how Digital Twin architectures can unify various modeling paradigms to provide richer, context-aware insights.

Explainable AI (XAI) has become increasingly prominent in recent literature. Chen, Wang, and Tsai (2009) introduced a model that integrates SHAP (SHapley Additive exPlanations) to clarify feature contributions in deep learning-based CT prediction. Their approach bridged the gap between high-performance AI and human-centric decision-making. Wang, Chen, and Chiu (2022) further demonstrated that interpretability can be achieved without sacrificing model accuracy, a critical consideration for industrial adoption. In this context, Chen, Lin, and Lin (2024) proposed a fuzzy collaborative forecasting system that uses explainable mechanisms to manage prediction uncertainty and support operator engagement.

These articles collectively represent the evolution from traditional predictive modeling toward integrated, interpretable, and operationally viable systems. They reflect the field's increasing emphasis on hybrid intelligence, real-time analytics, and human-AI collaboration. As such, they serve as both benchmarks and blueprints for future research in predictive modeling for semiconductor manufacturing.

## 7 DISCUSSION AND FUTURE PERSPECTIVES

This section synthesizes the key contributions of predictive modeling in semiconductor manufacturing, reflecting on both the strengths of existing methods and the limitations that persist when transitioning from research to practice. The analysis also highlights forward-looking research opportunities that could transform the industrial adoption of CT prediction tools in complex and evolving fabrication environments.

### 7.1 Synthesis of Predictive Modeling Approaches

The diversity of predictive modeling techniques reflects the multifaceted nature of semiconductor manufacturing processes. Analytical and statistical models provide accessible, computationally efficient solutions grounded in queuing theory, regression analysis, or time-series forecasting. These methods are particularly useful when interpretability and quick implementation are prioritized. However, they often fall short in capturing the full complexity of modern fabs, particularly under high variability, nonlinearity, or reentrant flow conditions.

AI-driven models, such as artificial neural networks, fuzzy logic systems, and hybrid intelligent frameworks, have emerged to address these limitations. These methods exhibit strong capabilities in learning complex patterns from historical data, adapting to changing operational dynamics, and outperforming traditional models in terms of predictive accuracy. Yet, they also introduce new challenges, particularly around interpretability, data dependency, and scalability. Their "black-box" nature often makes them difficult to trust or deploy in production environments without sufficient validation and explanation tools.

Hybrid models—combining analytical, statistical, AI, and simulation components—have proven especially effective in balancing accuracy, flexibility, and interpretability. When integrated with real-time systems like Digital Twins, these models gain the added advantage of synchronizing virtual predictions with physical processes. This integration enables dynamic decision-making and continuous process improvement, transforming static forecasting into real-time operational insight. Across all modeling strategies, the review confirms that effectiveness is highly context-dependent. The most successful implementations align modeling techniques with the unique characteristics of the fab, including its data infrastructure, production mix, and decision-making workflows.

### 7.2 Ongoing Challenges in Deployment

Despite advancements in theory and experimentation, real-world deployment of CT prediction models remains limited. One of the most significant barriers is data quality and accessibility. Semiconductor fabs often generate vast quantities of data, yet this data may be noisy, incomplete, or fragmented across disparate systems. Issues like sensor inaccuracies, data latency, and inconsistencies in event logging can undermine model training and operational use.

Another major challenge is the gap between model sophistication and operator usability. While AI models can achieve high accuracy, their lack of transparency creates a barrier to acceptance. Engineers and managers need to understand, trust, and act upon predictions. Without mechanisms for explainability, even the most accurate model may be underutilized. This is particularly true in high-stakes manufacturing contexts where decisions have direct implications on yield, quality, and customer satisfaction.

Scalability is another concern. Models that perform well in laboratory or pilot settings may fail to deliver in large-scale production environments with thousands of tools and dynamic workflows. Moreover, integrating predictive tools into existing manufacturing execution systems (MES), enterprise resource planning (ERP) platforms, or dispatching mechanisms often requires significant customization. These integrations are not merely technical; they demand alignment across IT, operations, and process engineering teams. Maintenance and model retraining also require dedicated resources, further increasing the burden of long-term adoption.

These challenges highlight that successful deployment involves more than choosing the right algorithm—it requires building a complete system that connects data, people, processes, and technology in a cohesive and sustainable manner.

### 7.3 Directions for Future Research

To address these persistent challenges, future research must move beyond algorithmic performance and consider systemic integration, human-machine collaboration, and operational resilience. One promising direction is the development of self-adaptive models that learn continuously from new data and respond to changes in production without manual recalibration. Such models could incorporate reinforcement learning, concept drift detection, or meta-learning techniques to remain robust in volatile manufacturing contexts.

Another emerging opportunity lies in the use of generative AI for data augmentation. Techniques such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) can generate synthetic production scenarios, helping overcome data scarcity, improve model robustness, and explore edge-case conditions that are rarely observed in historical logs but critical for risk-aware planning.

Equally important is the evolution of human-AI interfaces. Model outputs must be communicated in ways that are understandable, actionable, and aligned with operator expectations. Explainable AI (XAI) frameworks can help bridge the gap, especially when embedded into visual dashboards, interactive simulation tools, or Digital Twin environments.

On the architectural side, research should focus on lightweight and modular Digital Twin systems that can be rapidly configured, scaled, and maintained. Technologies such as edge computing and federated learning can enable distributed prediction capabilities that respect data privacy and reduce latency, particularly in highly automated fabs.

Finally, resilience must become a central goal of predictive systems. Models should not only provide accurate forecasts under ideal conditions but also maintain functionality during disruptions—be it equipment failure, supply chain shocks, or process shifts. Indeed, these disruptions change completely the arrival and process rates of a wafer fab as well as its priority policies, resulting in big shifts in congestion states and CT compared to past data observations. This calls for new evaluation benchmarks and robustness testing protocols that better reflect the realities of modern semiconductor operations.

Together, these research directions point toward a new generation of predictive modeling tools—ones that are not only accurate and efficient, but also interpretable, integrable, and resilient in real-world manufacturing environments. The next and final section concludes this review by summarizing its main contributions and outlining the implications for researchers and practitioners alike.

## 8 CONCLUSION

This systematic review has provided a first comprehensive exploration of predictive modeling approaches for cycle time (CT) forecasting in semiconductor manufacturing. By examining a wide range of methodologies—including analytical, statistical, AI-driven, hybrid, and simulation-based techniques—the study has clarified how different approaches align with the diverse operational realities and constraints of semiconductor fabs. In particular, the integration of predictive models with Digital Twin architectures stands out as a promising pathway to support dynamic, real-time decision-making.

Beyond methodology, this review has shown critical dimensions that influence the success or failure of CT prediction models in practice. These include data quality and availability, model interpretability, scalability, and integration within existing manufacturing systems. The analysis emphasizes that predictive accuracy alone is insufficient; operational deployment requires careful consideration of technical, organizational, and human factors.

Looking forward, we need to include and analyze other articles. The future of CT prediction in semiconductor manufacturing lies in the development of systems that are not only precise but also transparent, adaptive, and resilient. Research should focus on enhancing model self-learning capabilities, expanding the role of generative AI. Modular Digital Twin infrastructures and distributed computing technologies such as edge and federated learning will further support these goals by enabling scalable and real-time deployment.

Furthermore, CT prediction models and Digital Twin infrastructures must be integrated in the supply chain context where there are used. Indeed the direct customer of a wafer fab is most often a stock of Available to Promise (ATP) ICs within the same company which decouples the wafer fab production from the final customer. As such, the primary users of the models are Planners which must integrate them within a more global decision strategy accounting for the supply chain demand and the ATP stocks available.

## REFERENCES

Chang, P., and T. Liao. 2006. "Combining SOM and Fuzzy Rule Base for Flow Time Prediction in Semiconductor Manufacturing Factory". *Applied Soft Computing* 6:198–206.

Chang, P.-J., C.-Y. Hsieh, J.-J. Huang, and C.-J. Chang. 2001. "A Case-Based Reasoning Approach for Due-Date Assignment in a Wafer Fabrication Factory". *International Journal of Advanced Manufacturing Technology* 18(8):603–608.

Chen, T. 2007. "An intelligent Hybrid System for Wafer Lot Output Time Prediction". *Advanced Engineering Informatics* 21:55–65.

Chen, T. 2008. "An Intelligent Mechanism for Lot Output Time Prediction and Achievability Evaluation in a Wafer Fab". *Computers & Industrial Engineering* 54(1):77–94.

Chen, T., and Y.-C. Wang. 2022, October. "A Two-Stage Explainable Artificial Intelligence Approach for Classification-Based Job Cycle Time Prediction". *The International Journal of Advanced Manufacturing Technology* 123(11).

Chen, T., Y.-C. Wang, and H.-R. Tsai. 2009. "Lot Cycle Time Prediction in a Ramping-up Semiconductor Manufacturing Factory with a SOM–FBPN-Ensemble Approach with Multiple Buckets and Partial Normalization". *The International Journal of Advanced Manufacturing Technology* 42:1206–1216.

Chen, T.-C., C.-W. Lin, and Y.-C. Lin. 2024. "A Fuzzy Collaborative Forecasting Approach Based on XAI Applications for Cycle Time Range Estimation". *Applied Soft Computing* 151:111122.

Deenen, P. C., J. Middelhuis, A. Akcay, and W. M. P. van der Aalst. 2024. "Data-Driven Aggregate Modeling of a Semiconductor wafer fab to Predict WIP Levels and Cycle Time Distributions". *Flexible Services and Manufacturing Journal* 36:567–596.

Ivanov, D., and A. Dolgui. 2020. "Viability of Intertwined Supply Networks: Extending the Supply Chain Resilience Angles Towards Survivability. A Position Paper Motivated by COVID-19 Outbreak". *International Journal of Production Research* 58:2904–2915.

Kang, H. S., J. Y. Lee, S. Choi, H. Kim, J. Park, J. Y. Son, *et al.* 2016. "Smart manufacturing: Past research, present findings, and future directions". *International Journal of Precision Engineering and Manufacturing-Green Technology* 3:111–128.

Lee, G. M., and X. Gao. 2021. "A Hybrid Approach Combining Fuzzy c-Means-Based Genetic Algorithm and Machine Learning for Predicting Job Cycle Times for Semiconductor Manufacturing". *Applied Sciences* 11(16):7428.

Little, J. D. 1961. "A proof for the queuing formula: L = lambda * W". *Operations Research* 9:383–387.

Meidan, Y., B. Lerner, G. Rabinowitz, and M. Hassoun. 2011. "Cycle-Time Key Factor Identification and Prediction in Semiconductor Manufacturing Using Machine Learning and Data Mining". *IEEE Transactions on Semiconductor Manufacturing* 24(2):237–246.

Monostori, L. 2014. "Cyber-physical Production Systems: Roots, Expectations and R&D Challenges". *Procedia CIRP* 17:9–13.

Morrison, J. R., and D. P. Martin. 2007. "Practical Extensions to Cycle Time Approximations for the G/G/m-Queue With Applications". *IEEE Transactions on Automation Science and Engineering* 4(4):523–535.

Qi, Q., F. Tao, T. Hu, N. Anwer, A. Liu, Y. Wei, *et al.* 2021. "Enabling Technologies and Tools for Digital Twin". *Journal of Manufacturing Systems* 58:3–21.

Seidel, G., C. F. Lee, A. Y. Tang, S. L. Low, B. P. Gan, and W. Scholl. 2020. "Challenges Associated with Realization of Lot Level Fab Out Forecast in a Giga Wafer Fabrication Plant". In *2020 Winter Simulation Conference (WSC)*, 1777–1788. Orlando, FL, USA.

Tao, F., Q. Qi, L. Wang, and A. Y. C. Nee. 2019. "Digital Twins and Cyber–Physical Systems toward Smart Manufacturing and Industry 4.0: Correlation and Comparison". *Engineering* 5(4):653–661.

Wang, J., J. Zhang, and X. Wang. 2018. "A Data-Driven Cycle Time Prediction With Feature Selection in a Semiconductor Wafer Fabrication System". *IEEE Transactions on Semiconductor Manufacturing* 31:173–182.

Wang, Y.-C., T. Chen, and M.-C. Chiu. 2022, December. "An Explainable Deep-Learning Approach for Job Cycle Time Prediction". *Decision Analytics Journal* 6:100153.

## AUTHOR BIOGRAPHIES

**CLAUDE YUGMA** is Professor the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne in France since 2016 in Manufacturing Sciences and Logistics department. He received the Ph.D. degree from the Institut National Polytechnique of Grenoble, France, in 2003, and his H.D.R. from the Jean-Monnet University, Saint-Etienne, in December 2013. He was a Postdoctoral fellow at the Ecole Nationale Supérieure de Génie Industriel, Grenoble from 2003 to 2004 and from 2005 to 2006 at EMSE. He co-organized several international conferences as for example the 2013 edition of the conference Modeling and Analysis of Semiconductor Manufacturing. His research interests modeling and scheduling in semiconductor manufacturing. He has published more than 20 papers in international journals and has contributed to more than 80 communications in conferences. His email address is yugma@emse.fr.

**ADRIEN WARTELLE** is a Postdoctoral Fellow at the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne since January 2023 in the Manufacturing Sciences and Logistics department. He received the Ph.D degree from the University of Technology of Troyes in 2022 where he studied the modeling and simulation of congestion in healthcare systems in

cooperation with the Hospital Center of Troyes. His research interests relates to the modeling and simulation of complex system in a data-driven operational research context using notably machine learning and artificial intelligence. His email address is adrien.wartelle@emse.fr .

**STÉPHANE DAUZÈRE-PÉRÈS** Professor at Mines Saint-Etienne, France, and Adjunct Professor at BI Norwegian Business School, Norway. He received the Ph.D. degree from Paul Sabatier University in Toulouse, France, in 1992 and the H.D.R. from Pierre and Marie Curie University, Paris, France, in 1998. He was a Postdoctoral Fellow at M.I.T., U.S.A., in 1992 and 1993, and Research Scientist at Erasmus University Rotterdam, The Netherlands, in 1994. His research interests broadly include modeling and optimization of operations at various decision levels in manufacturing and logistics, with a special emphasis on production planning and scheduling, on semiconductor manufacturing and on railway operations. He has published more than 100 papers in international journals and contributed to more than 250 communications in national and international conferences. Stéphane Dauzère-Pérès has coordinated numerous academic and industrial research projects, including 4 European projects and more than 30 industrial (CIFRE) PhD theses, and also eight conferences. He was runner-up in 2006 of the Franz Edelman Award Competition, and won the Best Applied Paper of the Winter Simulation Conference in 2013 and the EURO award for the best theory and methodology EJOR paper in 2021. His email address is dauzere-peres@emse.fr .

**PASCAL ROBERT** is an Automation Expert at STMicroelectronics in Rousset (France). His email address is is pascal.robert@st.com.

**RENAUD ROUSSEL** is a Scheduling and Dispatching Full Automation Expert at STMicroelectronics in Crolles (France). He has been working for more than 2 decades in the semiconductor industry in manufacturing science at the frontier between operational management, industrial engineering and data science to make the fab as efficient as possible. His email address is is renaud.roussel@st.com.

**JASPER VAN HEUGTEN** is Chief Technology Officer at Minds.ai. His e-mail address is is jasper@minds.ai.

**TAKI-EDDINE KORABI** is engineer at Minds.ai. His e-mail address is taki@minds.ai.