Artificial Intelligence in Semiconductor Manufacturing: A Literature Review

Matthew S. Jones[1] and Houssam Abbas[2]

***Abstract*— This literature review examines the use of artificial intelligence (AI) in semiconductor manufacturing. Key findings indicate an opportunity to leverage AI models, particularly neural networks, to improve scheduling and equipment setup, virtual metrology (VM) and fault detection, and yield prediction. Opportunities include optimizing production paths to maximize throughput given equipment setup and maintenance schedules, expanding the use of VM to reduce reliance on more-expensive conventional metrology, and gaining efficiency through more-accurate yield prediction methods. Furthermore, as AI compute capacity and speed increase, the breadth and depth of value that AI can add to semiconductor manufacturing also increase.**

I. INTRODUCTION

Award-winning science fiction writer Sir Arthur C. Clarke once quipped, "Any sufficiently advanced technology is indistinguishable from magic." [26] Nowhere in modern life is this more true than in semiconductor manufacturing. Combining sand and metal with chemicals to make a functioning integrated circuit containing billions of transistors is the most complicated manufacturing process in the history of mankind. The challenges that must be overcome to do this profitably at scale are as broad as they are technical. Newly cut wafers are polished, materials are added, microscopic patterns are created, different materials are added according to the pattern, the wafer is cleaned, then the cycle repeats on the same tools but with different materials and patterns (known as "re-entry") until processing is complete. The die are then cut and packaged.

An advanced semiconductor manufacturing process can include 1,000 or more steps across a dozen or more machines that can take weeks to complete (See Fig. 1). Ensuring products are completed according to their committed timeline requires precise equipment scheduling, maintenance, and setup procedures; detecting issues early is critical to minimizing wasted time and materials; and accurately predicting yield is crucial for ensuring the correct product quantities without over producing.

This paper covers the latest AI-related research papers in semiconductor manufacturing and groups them into three main categories: scheduling and equipment setup, virtual metrology and fault detection, and yield prediction. Each group has its own problem statement and summary of the papers reviewed.

[1]Matthew S. Jones is a Ph.D. student at Oregon State University and a Data Scientist at Intel Corporation: www.mattjones.ai.

[2]Dr. Houssam Abbas is an Assistant Professor in the Department of Electrical Engineering and Computer Science at Oregon State University: www.houssamabbas.com.

A. Why This Review?

Two papers that review the use of AI in semiconductor manufacturing are among those selected for this one. In 2022, the Proceeds of the Winter Simulation Conference published a panel discussion titled "Production-level artificial intelligence applications in semiconductor manufacturing" which provides the latest (at the time) research and opinions of five experts across the industry [22]. In 2024, the International Conference on Information and Communication Technology Convergence published a paper titled "Bibliometric Study of Artificial Intelligence and Semiconductor Manufacturing Industry" which provides insight into, and trends around, where and how new research was being published [10]. This paper will differ from both of those in that its goal is to summarize the current research across the entire field, and to provide insight into its establishment as a foundation for the future.

Semiconductor manufacturing has always been on the bleeding edge of technological advancement, so it has always been ripe for improvement by artificial intelligence. However, only recently has the compute power available to models been sufficient to run in the factory in real-time, providing opportunities not only for different research, but also for the current research to expand in scope. This makes 2025 the perfect year for another literature review of AI in semiconductor manufacturing.

B. Paper Organization

This paper is organized as follows: Section II describes the review parameters used for the search. Sections III, IV, and V describe the problem statement and the AI-related research for each of those main categories. Section VI concludes the paper by summarizing the key take-aways and proposing future research topics to further the use of AI in semiconductor manufacturing.

II. SEARCH PARAMETERS & RESULTS

Three databases were searched for papers: CASE Proceedings from 2019 through 2023, IEEExplore, and the ACM Digital Library (ACM-DL). The CASE Proceedings contained 1497 papers, 96 of which were reviewed based on title, 53 were read based on abstract, and seven of those were included in this review. The search parameters for IEEExplore required "Semiconductor" and "Manufacturing," "Fab," or "Foundry" in the paper title, and "Artificial Intelligence" or " AI " - *the spaces are intentional to avoid including those consecutive letters within words* - anywhere in the metadata. This search resulted in 61 papers, 11 of which were included

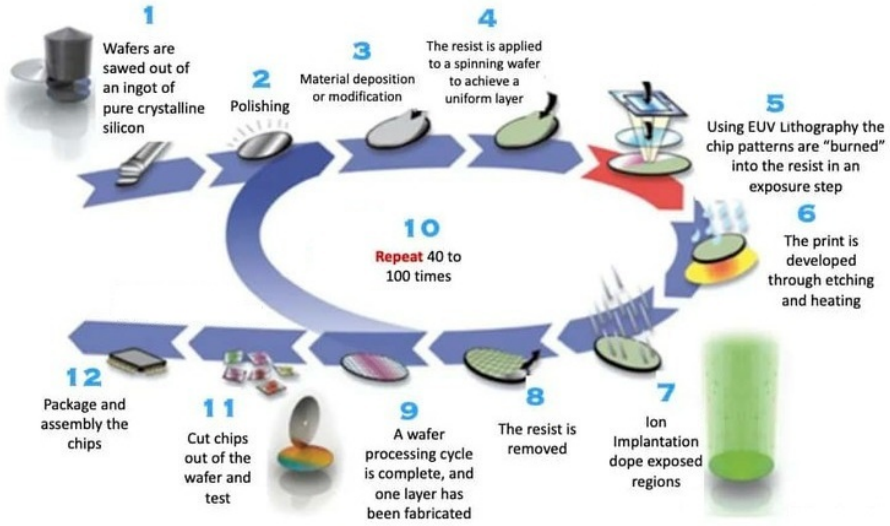

Fig. 1. Chip Fabrication Steps [27]

in this review after filtering by title and abstract. The search parameters for the ACM-DL required "Semiconductor" and "Manufacturing," "Fab," or "Foundry" in the paper title, "Artificial Intelligence" or " AI " anywhere in the metadata, and a publication date in the past five years (2019 through 2024). This search of ACM Full-Text Collection (774,356 records) resulted in 25 papers, seven of which were included in this review after filtering by title and abstract. A date-based restriction was not necessary for IEEExplore. but was for the ACM-DL.

Many of the papers included in this review have co-authors from the semiconductor industry. This led to more realistic data being used in the research, since it could be taken from actual semiconductor factories, but it also led to some obfuscations. In some cases, a specific tool / machine in the factory was the target of the research, but it is not specified in the paper. When tool details were included in the paper, they were included included in the paper summary. Similarly, the model type is stated more generically in some papers than in others. When model specificity is included in the paper, it is included in the summary.

III. SCHEDULING & EQUIPMENT SETUP

A. Problem Statement

Failure to efficiently schedule lots' movement through the manufacturing line can lead to poor machine utilization [3], equipment contamination and increased scrap [1], and lower overall factory yield. While the precise impact of poor scheduling and equipment setup timing cannot be quantified meaningfully because of the extreme variation between individual factories and product mixes, suffice it to say that any improvement is valuable.

Silicon wafers are processed in "lots" and a typical lot contains 25 wafers. The largest wafers are 300mm (˜18") in diameter and can contain hundreds of individual integrated circuit chips. The chips on a wafer, and the wafers in a lot, are supposed to be identical. However, setup and processing times for the different machines vary greatly, and some sequential steps have a minimum and/or maximum lag times between them that must be met to ensure quality [13]. These factors, along with machine down time due to maintenance, make scheduling a complex process.

B. AI-Related Research

While there doesn't seem to be a single model that is dominant, various flavors of neural networks seem to be the most common model used schedule lots through the factory, inform equipment maintenance decisions, and validate the setup parameters. Here are summaries of the papers addressing scheduling and equipment setup:

- A long short-term memory (LSTM) neural network accurately predicted (RMSE = 1.10) the chemical concentration in equipment at numerous points in the factory using information from production lines, equipment, and chemical tanks, which can be used to inform scheduling and equipment maintenance decisions [1].
- A customizable reinforcement learning (RL) agent was coded in Python and modeled the factory as a Markov Decision Process (MDP) to simulate three wafer priority classes (normal, urgent, and super-urgent) [3]. It failed to outperform a standard scheduling process on the SMT2020 dataset.
- A deep-learning neural network (DLNN) was proposed to predict cycle time, equipment utilization, and output quantity given historical data such as wafer quantity, urgency, cycle time, and machine performance; but was not implemented [9].
- A random tree forest model with optimized factory was used to determine the path with the highest yield given data on supposedly identical machines in the same factory line, and led to an maximum R-square value of 0.719 [11].

- A graph representation model was successfully used to differentiate between fab states in a low-dimensional space and aid in process flows, material transfer, equipment setup and maintenance by using training data representing 1,000 fab states [14].
- A neural network model was used to improve the throughput efficiency of a photo-lithography machine by ensuring that the setup state requirements were considered during the prioritization of wafer lots [16]. It showed a 32% improvement in validation MSE using simplified factory model for validation.
- An extended finite state machine (FSM) was used to simulate the movement of wafers through the equipment front end module (EFEM) using the average utilization of the bottleneck chamber as the performance index [18]. Deadlock was shown to be eliminated and utilization was improved in random simulations.
- Statistical process control was used guide process parameter settings (which must be updated before each new lot is processed) using data from a semiconductor foundry [19]. The predicted variance was within 2% of the actual variance.
- A neural network was used to detect unsuccessful ion beam tuning, which is necessary for every setup, using equipment log files. It showed a 65% detection success rate [21].

IV. VIRTUAL METROLOGY & FAULT DETECTION

A. Problem Statement

Metrology is the process of taking measurements at key points in the process to ensure that it stays in control. For example: a particle check that detects and maps the implanted particles after deposition, the thickness of the polished layer after the planar process, the height or depth of the new feature after etching, or the critical dimension after lithography [12]. The purpose of metrology and fault detection is two-fold: First, to ensure that errors do not make it into the final product, and second, to help root-cause manufacturing issues and resolve them [20].

Virtual metrology (VM) is predicting those measurements based on a *different set* of machine parameters and sensor data in the production equipment, as opposed to measuring them physically. Effective VM has the potential to lead to lower reliance on physical metrology, which could lead to higher-density wafers (more chips per wafer), greater layout optimization, and improved machine efficiency.

B. AI-Related Research

Convolutional neural networks (CNN) are a natural choice for virtual metrology as they are particularly well-suited for analyzing visual data, like those from image sensors. Here are summaries of the papers addressing virtual metrology and fault detection:

- A CNN was used to analyze sensor data (collected every second for a recipe lasting 150 seconds) from equipment, including gas flows, pressures, temperatures, and capacitances, and determined whether variables are off-standard with an FI-score of 95% [2].
- A Copula network deconvolution (CND) was used to construct accurate join distribution using the Copula function method using data from a 28nm fab line in Shanghai to accurately measuring the complex non-linear relationship between parameters and determine the root-cause of yield loss [4].
- A CNN was used to estimate parameters, including resistance, capacitance, and inductance from a simple, simulation-generated 150mm microstrip line to non-destructively detect and predict soft and hard failures [6]. It achieved a mean deviation of 0.77% for hard failure localization.
- A Double U-Net Deep Neural Network was used to remove noise and improve contrast in critical dimension scanning electron microscope images taken in a cleanroom at Cornell, and showed results comparable to a commercial system [12].
- A depth-based isolation feature importance model was used on real-world semiconductor manufacturing data to sort the features by importance (necessary for root-cause analysis) and produced results comparable to domain experts [15].
- The process partial least squares method modeled the process flow and variable relationships, used equipment data from seven STMicroelectronics machines (readings from 151 sensors and over 2,000 wafers), to show correlation between machine data for effective VM [17].
- A One-Class Support Vector Machine engine was used to detect out-of-control (OOC) wafers using raw data of the in-line measurements [20]. The engine successfully detected six strong OOC also detected by the Shewhart charts.
- A CNN with a spatial pyramid pooling layer on top of the last convolutional layer was used to improve VM models of multi-chamber production processes using data from other processes leading to training convergence compared to a baseline model [23].
- A CNN was used to predict overlay in photolithography using 18 input variables from 2000 wafers and providing 7 overlay output variables for effective VM [24]. With joint modeling, the average MASE was lower by 1.03.

V. YIELD PREDICTION

A. Problem Statement

Yield in the context of semiconductor fabs is usually "die yield," which is the percentage of "good" die per wafer. This does not tell the whole story. Yield is also a function of time. A factory that produces 100,000 good chips per day has a higher yield as a function of time than one that produces only 80,000, even if the die yield of the former factory is much lower.

Factories aspire to have a high yield under both calculations, but the methods to achieve them will differ. A high die yield is a function of the precision of the process flows, while

a high production yield is a function of throughput, which is impacted by equipment management and maintenance [8]. By effectively predicting yield, a factory is able to reduce cost by removing low-yield wafers from the line before more expenses are incurred.

B. AI-Related Research

The research on yield prediction, though sparse, is still about improving yield. Long short-term memory networks seem to be effective, though the models still require more validation, particularly for overall wafer yield. Here are summaries of the papers addressing yield prediction:

- A Gaussian mixture model was used to select from other models (Lasso, Support Vector Machine, K-Nearest Neighbor, Random Forrest Regressor, Ada Boost Regressor, and XGBoost Regressor) using three months of production line data to detect low yield problems between wafer fabrication and final test (FT) leading to a 27.2% FT yield improvement [5].
- A conjecture model was used to predict current lot quality based on its sensor data by training it on the same sensor data and the actual lot quality from a semiconductor factory in Taiwan [7]. It achieved an MAPE values under 5%, which implies that the quality prognostics scheme is applicable.
- An LSTM neural network was proposed to predict whether a production lot is likely to contain certain anomalies, which would lead to early defect detection and waste reduction, and could guide maintenance and production scheduling; validation is pending [8].
- An LSTM feed-forward neural network was used to conduct virtual metrology on 3D veritcal-NAND flash memory from South Korea and out-performed other regression models at predicting yield at wafer edges [25].

VI. CONCLUSION

In this paper, we summarized the current AI-related research as it applies to semiconductor manufacturing through a literature review of papers found through targeted searches in recent CASE Proceedings, IEEExplore, and the ACM-DL. The 25 papers were mostly published in simulation-related journals, such as CASE and the Proceedings of the Winter Simulation Conference. The latest Machine Learning-focused conferences, NeurIPS and ICML, did *not* include research on semiconductor manufacturing. Table 1 shows the AI Type, Results, and Validation Data or Method for each paper that provided it. Ref. refers to the reference number and the rows are color-coded by section.

A. Key Takeaways

Key findings indicate current opportunities to leverage AI models, particularly neural networks, to improve scheduling and equipment setup, VM and fault detection, and yield prediction. These predominantly home-grown models were typically trained using actual data from factory sensors, which is the result of industry involvement in the research, particularly in Europe and Asia. As the models improve and are able to identify which data are relevant, a significant benefit will be to stop collecting unnecessary data, which increases costs and reduces efficiency without providing any practical advantage. This will will require a significant mind-shift since the ”more is better” attitude is generally held within factory environments, especially with respect to data collection.

B. Future Research: Closing the Loop

Further research is needed to ”close the loop” from analysis and prediction to system and process control. For example, detecting that the ion beam tuning was unsuccessful can be the first step in re-tuning it. The current state-of-the-art does not include such closed-loop control, likely due to the fact that the models are still not as skilled as trained engineers and compute power has historically not been able handle the load. Now that compute power is unlikely to be the bottleneck, more-robust AI agents should be empowered to make real-time changes to input and control parameters, leading to significant improvement in all areas of semiconductor manufacturing.

Ref.	AI Type	Results	Validation Data or Method
1	LSTM Neural Network	RMSE of 1.10	Factory equipment data
2	Convolutional Neural Network	95% FI-score	Equipment sensor data
3	Customizable Reinforcement Learning Agent	Failed to outperform standard process	SMT2020 dataset
4	Copula Network Deconvolution	Measured relationship between parameters	28nm fab line data
5	Gaussian Mixture Model	27.2% FT yield improvement	Production line data
6	Convolutional Neural Network	Mean deviation of 0.77%	Simulation-generated 150mm microstrip
7	Conjecture Model	MAPE Values under 5%	Factory data
11	Random Tree Forest Model	R-square of 0.719	Optimized wafer data
12	Double U-Net Deep Neural Network	Comparable to a commercial system	Cleanroom data
14	Graph Representation Model	Successful differentiation	Fab state training data
15	Depth-Based Isolation Feature Importance Model	Results comparable to domain experts	Fab manufacturing data
16	Neural Network	32% improvement in validation MSE	Simplified factory model
17	Process Partial Least Squares	Correlation between machine data	STMicroelectronics machine data
18	Neural Network	No deadlock, improved utilization	Random simulation
19	Statistical Process Control	Variance prediction within 2%	Semiconductor foundry data
20	One-Class Support Vector Machine Engine	6 successful predictions	Compared to Shewhart charts
21	Neural Network	65% detection success rate	Equipment log files
23	Convolutional Neural Network	Training convergence to baseline	Data from other processes
24	Convolutional Neural Network	1.03 lower MASE	Wafer data
25	LSTM Neural Network	Out-performed other models	Actual fab data

TABLE I

SUMMARY: SCHEDULING & EQUIPMENT SETUP, VM & FAULT DETECTION, YIELD PREDICTION

REFERENCES

[1] H.-M. Cho, K.-H. Lee, P. Shim, and A. Park, "A Chemical Monitoring and Prediction System in Semiconductor Manufacturing Process Using Bigdata and AI Techniques," in 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Apr. 2021, pp. 488–491. doi: 10.1109/ICAIIC51459.2021.9415241.

[2] P. Tchatchoua, G. Graton, M. Ouladsine, and M. Juge, "A Comparative Evaluation of Deep Learning Anomaly Detection Techniques on Semiconductor Multivariate Time Series Data," in 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), Aug. 2021, pp. 1613–1620. doi: 10.1109/CASE49439.2021.9551541.

[3] B. Kovács, P. Tassel, M. Gebser, and G. Seidel, "A Customizable Reinforcement Learning Environment for Semiconductor Fab Simulation," in Proceedings of the Winter Simulation Conference, in WSC '22. Singapore, Singapore: IEEE Press, 2023, pp. 2663–2674.

[4] H. -W. Xu, R. -Z. Tan, Q. -L. Chen, Y. Zhou, and W. Qin, "A Framework of Direct Correlation Identification for Wafer Fault Detection," in 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE), Aug. 2023, pp. 1–6. doi: 10.1109/CASE56687.2023.10260446.

[5] D. Jiang, W. Lin, and N. Raghavan, "A Gaussian Mixture Model Clustering Ensemble Regressor for Semiconductor Manufacturing Final Test Yield Prediction," IEEE Access, vol. 9, pp. 22253–22263, 2021, doi: 10.1109/ACCESS.2021.3055433.

[6] S. Kamm, K. Sharma, N. Jazdi, and M. Weyrich, "A Hybrid Modelling Approach for Parameter Estimation of Analytical Reflection Models in the Failure Analysis Process of Semiconductors," in 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), Aug. 2021, pp. 417–422. doi: 10.1109/CASE49439.2021.9551454.

[7] Y.-C. Su, F.-T. Cheng, G.-W. Huang, M.-H. Hung, and T. Yang, "A quality prognostics scheme for semiconductor and TFT-LCD manufacturing processes," in 30th Annual Conference of IEEE Industrial Electronics Society, 2004. IECON 2004, Nov. 2004, pp. 1972-1977 Vol. 2. doi: 10.1109/IECON.2004.1431887.

[8] K. S. S. Alamin et al., "An AI-Enabled Framework for Smart Semiconductor Manufacturing," in 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE), Mar. 2024, pp. 1–6. doi: 10.23919/DATE58400.2024.10546768.

[9] M. S. K. Pheng and L. G. David, "Artificial Intelligence in Back-End Semiconductor Manufacturing: A Case Study," in 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Apr. 2022, pp. 1–4. doi: 10.1109/ICDCECE53908.2022.9792976.

[10] K. K. won, M. Lee, and E. Park, "Bibliometric Study of Artificial Intelligence and Semiconductor Manufacturing Industry," in 2024 15th International Conference on Information and Communication Technology Convergence (ICTC), Oct. 2024, pp. 1977–1981. doi: 10.1109/ICTC62082.2024.10827173.

[11] P. Stich, R. Busch, M. Wahl, C. Weber, and M. Fathi, "Branch selection and data optimization for selecting machines for processes in semiconductor manufacturing using AI-based predictions," in 2021 IEEE International Conference on Electro Information Technology (EIT), May 2021, pp. 1–6. doi: 10.1109/EIT51626.2021.9491836.

[12] S. Ding et al., "Double U-Net based Virtual Metrology on Plasma-Etch CD-SEM Images: AM: Advanced Metrology," in 2023 34th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), May 2023, pp. 1–5. doi: 10.1109/ASMC57536.2023.10121128.

[13] I. Konstantelos et al., "Fab-Wide Scheduling of Semiconductor Plants: A Large-Scale Industrial Deployment Case Study," in Proceedings of the Winter Simulation Conference, in WSC '22. Singapore, Singapore: IEEE Press, 2023, pp. 3297–3308. [Online]. Available: https://ieeexplore.ieee.org/document/10015364

[14] B. Schulz, C. Jacobi, A. Gisbrecht, A. Evangelos, C. W. Chan, and B. P. Gan, "Graph Representation and Embedding for Semiconductor Manufacturing Fab States," in Proceedings of the Winter Simulation Conference, in WSC '22. Singapore, Singapore: IEEE Press, 2023, pp. 3382–3393. [Online]. Available: https://ieeexplore.ieee.org/document/10015297

[15] M. Carletti et al., "Interpretable anomaly detection for knowledge discovery in semiconductor manufacturing," in Proceedings of the Winter Simulation Conference, in WSC '20. Orlando, Florida: IEEE Press, 2021, pp. 1875–1885. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9384026

[16] J.-H. Lee, H.-J. Kim, Y. Kim, Y. B. Kim, B.-H. Kim, and G.-H. Chung,

"Machine learning-based periodic setup changes for semiconductor manufacturing machines," in Proceedings of the Winter Simulation Conference, in WSC '21. Phoenix, Arizona: IEEE Press, 2022.

[17] G. van Kollenburg, R. Verhoeven, M. Holenderski, N. Meratnia, and D. Pagano, "Modeling Multivariate Relations in Multi-block Semiconductor Manufacturing Data Using Process Pls to Enhance Process Understanding," in Proceedings of the Winter Simulation Conference, in WSC '23. San Antonio, Texas, USA: IEEE Press, 2024, pp. 2333–2344. [Online]. Available: https://ieeexplore.ieee.org/document/10408180

[18] C. Hong and T. -E. Lee, "Modeling, Simulation and Supervisory Control of Semiconductor Manufacturing Cluster Tools with an Equipment Front-End Module," in 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), Aug. 2020, pp. 703–709. doi: 10.1109/CASE48305.2020.9216790.

[19] T. E. Korabi, G. Graton, E. M. E. Adel, M. Ouladsine, and J. Pinaton, "Monitoring of a sampled process data under Run-to-Run control: application to a semiconductor process," in 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), Aug. 2019, pp. 930–935. doi: 10.1109/COASE.2019.8843229.

[20] I. Rabhi, A. Roussy, F. Pasqualini, and C. Alegret, "Out-Of-Control Detection In Semiconductor Manufacturing using One-Class Support Vector Machines," in 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), Aug. 2021, pp. 1628–1633. doi: 10.1109/CASE49439.2021.9551477.

[21] A. Laber, M. Gebser, K. Schekotihin, and Y. Yang, "Predicting Ion Beam Tuning Success in Semiconductor Manufacturing," in 2022 14th International Conference on Advanced Semiconductor Devices and Microsystems (ASDAM), Oct. 2022, pp. 1–4. doi: 10.1109/ASDAM55965.2022.9966756.

[22] C.-F. Chien, H. Ehma, J. Fowler, L. Mönch, and C.-H. Wu, "Production-level artificial intelligence applications in semiconductor manufacturing," in Proceedings of the Winter Simulation Conference, in WSC '21. Phoenix, Arizona: IEEE Press, 2022.

[23] R. Clain, V. Borodin, M. Juge, and A. Roussy, "Virtual metrology for semiconductor manufacturing: Focus on transfer learning," in 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), Aug. 2021, pp. 1621–1626. doi: 10.1109/CASE49439.2021.9551567.

[24] T. C. Tin, S. C. Tan, and C. K. Lee, "Virtual Metrology in Semiconductor Fabrication Foundry Using Deep Learning Neural Networks," IEEE Access, vol. 10, pp. 81960–81973, 2022, doi: 10.1109/ACCESS.2022.3193783.

[25] D. Kim, M. Kim, and W. Kim, "Wafer Edge Yield Prediction Using a Combined Long Short-Term Memory and Feed- Forward Neural Network Model for Semiconductor Manufacturing," IEEE Access, vol. 8, pp. 215125–215132, 2020, doi: 10.1109/ACCESS.2020.3040426.

[26] https://www.azquotes.com/author/2936-Arthur_C_Clarke/tag/magic

[27] http://eitc.org/eita/emerging-technologies-research-and-ventures/new-materials-mechanical-engineering-and-smart-manufacturing/photos3/chip-fabrication-steps_050723a
(edited for clarity)