Dean&Francis
ISSN 2959-6157

AI-Based Analysis of Yield Losses and Defect Detection in Semiconductor Manufacturing

Shangwei Sun[1, *], **Wutong Huang**[2]

[1]Houston International Institute, Dalian Maritime University, Dalian, 116026, China
[2]Hajim School of Engineering, University of Rochester, Rochester, 14618, United States of America
*Corresponding author: 19945919991@163.com

Abstract:

Wafer yield serves as a critical process indicator in semiconductor manufacturing systems. With the rapid development of artificial intelligence (AI) technology, using AI to predict and control wafer yields has huge potential and a bright prospect. This paper mainly studies the technologies for utilizing artificial intelligence to enhance semiconductor yield and provides a comprehensive and systematic exposition. It describes current and key technologies and presents opinions and outlooks on future work. This paper explores forms and reasons for defects in wafer yield—functional defect loss and parametric defect loss—and elaborates on corresponding methods for dealing with random defects, system defects, and chip production (particularly memory production). This paper also systematically analyzes the application of artificial intelligence in optimizing semiconductor manufacturing yield, focusing on key technologies such as machine learning-based yield prediction, real-time process control, and defect detection. By comparing the performance of different AI approaches, it proposes an intelligent yield management framework for future semiconductor manufacturing, providing theoretical guidance for industrial intelligent upgrading.

Keywords: Semiconductor manufacturing, Yield improvement, Machine learning, Predictive maintenance, Digital twin.

1. Introduction

Semiconductor manufacturing is the current key and bottleneck issue of China's development of semiconductor technologies and increasingly important for promoting national economic growth [1].

As the foundation of modern information technology industry, the manufacturing process is one of the most complex industrial processes in the world today. With process nodes continuously shrinking to 5 nm and beyond, the manufacturing process consists of thousands of steps and has an average production

ISSN 2959-6157

period lasting more than 3 months. The wafer yield is the most important evaluation index for evaluating the manufacturing quality of the wafer, and wafer manufacturers need to effectively and accurately control its level [2]. While the semiconductor manufacturing process constantly advances with its manufacturing processes and technologies, new challenges are arising from these changes in various aspects of the manufacturing process of semiconductor. For instance, as the structures on a chip are minuscule, the chips become extremely sensitive to even the smallest changes in a process step and small particle pollution; defects are mainly concentrated on the chip surface, with significant challenges to the chip production process. In traditional defect identification and analysis system, some problems that may exist during the process were not comprehensively analyzed because it was impossible to fully extract, understand and apply all the wafer parameter data of every measurement, resulting in suboptimal analysis and controls over the wafer yield [3]. In order to cope with the situation, lots of studies appear concentrating on predicting methods for semiconductors by taking measurements data of semiconductor manufactured in an automated manner and so on. Predicting and calculating the semiconductor yield by extracting the massive data of automated semiconductor production equipment enables a further improve of predictive accuracy of the wafer yield. This method can provide prompt feedback when unexpected anomalies happen during production process, and achieve effective control on various process parameter and operation in the production process. Compared to traditional technology manufacturing, artificial intelligence involves machine learning and deep learning, making it much more efficient than those traditional techniques. By analyzing massive amounts of data, learning experience from the repeated trial and error, performing reasoning and making decisions, AI helps people achieve automation, intelligence and optimization within all levels. Leveraging both machine learning and deep learning, we constantly discover patterns or value out of the data streams that are collected in terabytes per second, continually train models, and self-discover the deep correlations within daily stream of numbers and data. There are a number of new and exciting emerging technologies and techniques of artificial intelligence coming out and their wide applications showing promise in a variety of fields of semiconductor manufacturing and improvement for yield prediction. In spite of the great potential of application for emerging artificial intelligence technologies, they still face enormous technical challenges and cost issues, which make it far from reaching widespread adoption in many industries or areas of technology. This paper explores strategies for addressing yield loss in semiconductor industry and leveraging AI, analyses research works status quo and aims at finally offering recommendation on trends and progress of utilizing AI to further improves semiconductor yields.

2. Yield Losses in Semiconductor Manufacturing: Causes and Impacts

The inherent instability of the film or substrate system and its high sensitivity to external stimuli are the primary causes of its surface complexity [4]. Cause for faults or defects occurring during semiconductor manufacturing may be divided into 2 groups, which are systematic defects and random defects. The random defects refer that they cannot foretell during semiconductor manufacturing. E.g., wafer contamination caused by foreign particles; material defective (defective or variation) and material faulty brought up by variations in atmosphere and surroundings. Systemic defects refer as a rule that these defects will happen again at semiconductor fabrication. These defect occurrences generally cause similar yield loss between batches of wafer. Chief reason of system deviation is the mistake in production processing or equipment defect. They are all the immediate reasons of yield losses, mainly functional loss and parametric loss.

2.1 Functional Yield Loss

Function yield loss: Fault particles that lead the die to not function correctly due to process manufacturing or quality problems and their influencing mechanisms on the dies. Reasons include Particle pollution.Foreign materials could possibly cause circuits short open, etc. For example, in the processing lithography step of the wafer, if there were foreign materials on the surface of the wafer, such as the microscopic particle, then they might interfere with the even and uniform coating of photoresist and fail the exposure, thus causing the defective particles; There are process variation effects caused by process steps etch and deposition. For example, a small error occurring from the lithography could influence pattern dimension changes (i.e., no match to wanted specification). Likewise, there is potential variation in which the conditions of the machine tools influence wafers during their processing. For example, problems from one of the machine tools used for etching such as under-etching, incomplete residue or over-etching. Some of the material aspects of the silicon wafers' process chemical could possibly cause yield losses. For example, minute traces contained within some impurities from the material in wafers or process chemicals could produce problem effects in some subsequent steps applied.

2.2 Parametric Yield Loss

Parametric yield loss refers to chips that function normally, but which fail to meet electrical specification (speed, power consumption and so on). Parameter cause —— The drift of process parameter is one key cause, i.e., change in the temperature, pressure or chemical concentration. Such as in chemical vapor process (CVD), slight changes in temperature can result in film thickness to be non-uniform across the wafer surface. This variation affects the performance of chip significantly.

The second problem is the fidelity of pattern. Slight distortion could happen in pattern during the photolithography process. It might affect transistor threshold voltages and thus affect chip's speed performance. Resistance or capacitance may also vary within metal layers. For instance, increased interconnect resistances may introduce signal delays, which impacts the chip performance as well.

2.3 Impact of Yield Loss on Semiconductor Manufacturing

For the chip industry to provide higher quality chips and ensure yield, high-precision defect detection is required at any time during semiconductor manufacturing. High cost is imposed on a semiconductor company by losses in the yield due to semiconductor manufacturing quality defects or damage. For instance, about 30% of production costs of ICs come from test and yield loss costs; 1/1,000,000,000 yield loss of modern semi-conductor at nanoscale nodes such as 7nm can result in up to 10% yield loss. The modern semiconductor process has over 1,000 steps and lasts over three months in total. Any issues during production, including defects, contamination, and errors, will cause a loss in the yield and thus the chip is not usable for commercial applications. Improving the yield can lead to cost savings and better production rate and corporate competitiveness.

3. Key Technology Analysis and Development Status

3.1 Pre-Production Yield Prediction

Stable and accurate prediction of wafer yield can help to find the problems during the process of wafer producing, improve the quality of chip and reduce the cost of chip production [5]. Jiang et al., proposed a novel scalable, universal, unattended ML framework to realize the prediction of final test(FT)yield. According to this frame, they can warn low yield before two months to correct problem as early as possible [6].

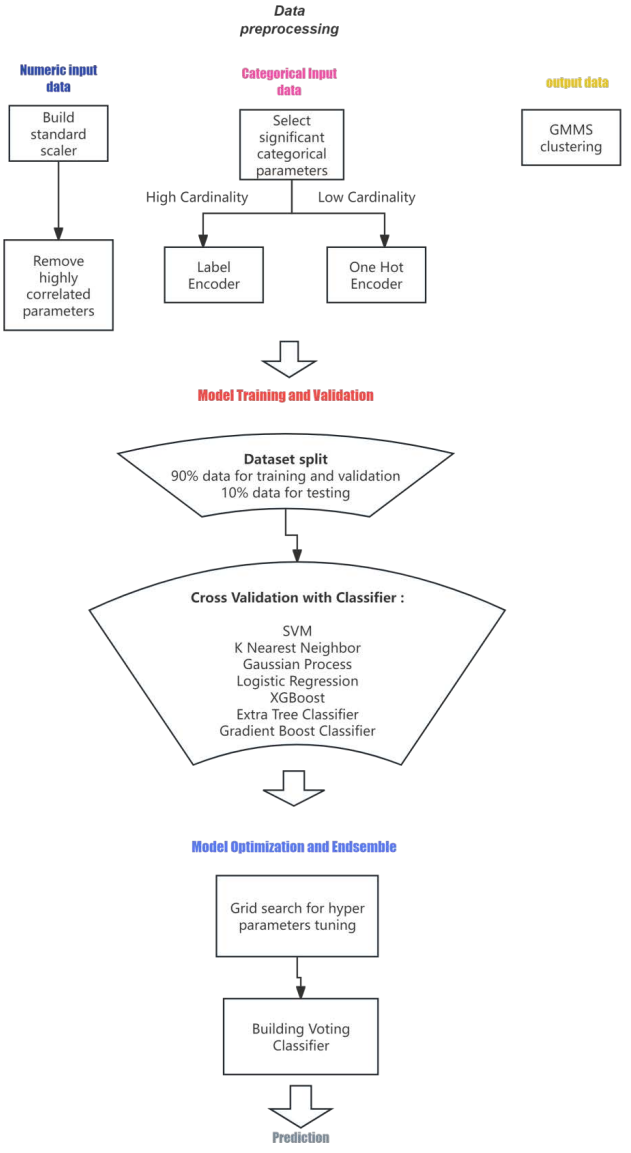

Fig. 1 Automated yield classification framework utilizing Gaussian Mixture Model for unsupervised clustering[6]

The major innovation for this Yield classification method which shown in figure one is that it could process data and make model from start to end fully automatic. One of them was output discrete which is mentioned in the paper. Since FT yield distribution with continuous value was not simple but had many modes, so they refused to give a certain cut point according to manual experience instead. They used GMM to conduct unsupervised clustering operation on all historical FT yield. It could automatically find optimal numbers of clusters which converted regression problem into classification problem.

Additionally, this framework can could do modeling for

ISSN 2959-6157

both numbers of features and others□s(ex. tester type, package type, program version categorical features). For high-cardinality categorical features, the framework intelligently selects between Label Encoder and One-Hot Encoder based on the ratio of data volume to dimensionality growth, thereby achieving a balance between preserving information and avoiding the curse of dimensionality.

Given the high-dimensional and imbalanced nature of semiconductor data, Jiang et al. employed F1-macro as the primary evaluation metric in their experiments. They compared multiple classifiers including SVM, KNN, logistic regression, Gaussian process, and XGBoost tree ensemble models. Then using grid searching optimize hyper-parameters, Finally selected three models having better performance to form voting classifier, on this basis, adopted many classical ensemble methods to improve performance of the framework as well.

Also, Feature importance analysis is another part of this framework. By using correlation measures such as Gini importance, the importance of each feature will be sorted out. This enabled us not only valid current experienced engineering skills but also uncover new hidden valuable factors which could be easily lost among huge sample numbers and led engineers toward true root cause.

Jiang et al. developed an end-to-end machine learning framework requiring no manual tuning to estimate the final test yield at wafer stage and can benefit semiconductor manufacturing efficiency while reducing cost control, serving as a powerful general framework for semiconductor manufacturing yield prediction [6]. The values of this method are that it automatically processes complicated production data without manual guidance caused by some known root causes, and explainability is given as we use model ensemble to produce high accuracy results for predicting semiconductor yield along with some useful insights into what explains them, which provide technical support on how to realize smart manufacturing in this field with important implications for making decisions according to data.

3.2 Real-Time Control in Manufacturing

Real-time process control involves adjusting process parameters on certain steps, such as lithography, etching and deposition, in time by online sensor data and measurements to suppress defect generation and stabilize yield rate, meanwhile offline prediction determines optimal model coefficients that can be loaded onto process equipment offline. The multi-layer ML (machine learning)-based Reinforcement Learning (RL) model proposed by Durowoju & Olowonigba is an overview of how they make it happen[7].

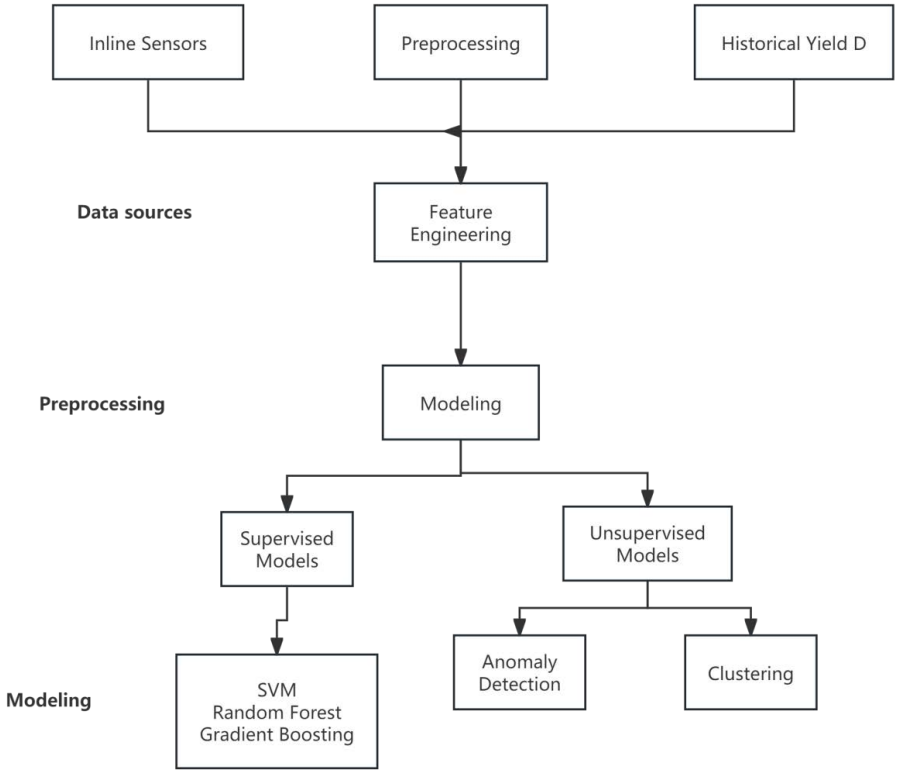

Fig. 2 Multi-layer machine learning architecture integrating supervised, unsupervised, and reinforcement learning [7]

Figure 2 illustrates the architecture components for their proposed ML framework, which has two basic parts. One is processing the incoming sensor data and inline CD-SEM/ellipsometer measurements, called ML model application. In ML model application layer, three learning models are applied with different purposes and roles: Supervised learning, Unsupervised learning and RL reinforcement learning.

Supervised learning uses historical inline sensor measurements and CD-SEM or ellipsometer measures as inputs for training a random forest / XGBoost model for predicting film thickness, CDs or Defect probability from sensor data and predictions.

With the unsupervised layer AutoEncoder + Isolation Forest method for identifying abnormality from sensor sequences with high-dimension features for detecting issues like chamber drift or particle contamination. With RL layer, agent programs were put on continuous process variables like lithography exposure dose, etch endpoint or deposition temperature to perform closed loop tuning in real-time at millisecond level. The Immediate reward in RL was defined with "CD uniformity within the process window" or "defect probability".

Durowoju & Olowonigba deployed the full solution over several months on both DRAM and logic chip production lines. They reported the DRAM line was able to achieve a 3.7% yield increase in a month with implementing RL parameter fine-tuning via exposing a number of products on a test machine under variable real-time conditions. Batch to batch variation was greatly lowered. For logic chips, implementing anomaly detection using AutoEncoder in their DRIE etch process was able to detect gas flow drift 6 hours in advance causing a 58% reduction of etching related defects per product processed. Their system achieved GPU□edge computing node inference time below 3.2s per wafer, satisfying needs of high-takt lines[7].

With the continuous progress of integrated circuits to sub-10-nanometer (i.e. smaller than 10 nm) nodes and beyond, such as 7 nanometers, even small process deviations may lead to functional defects in components and thus cause such issues as a low yield rate. Traditional systems were able to use only static control with fixed threshold values, set from predetermined setting values; they are, however, no longer enough for the actual control, which requires dynamic adjustments and optimizations according to the real-time state of production. This study pioneers the integrated use of supervised learning, unsupervised learning and reinforcement learning to achieve a feasible closed-loop feedback control system, to enable the implementation of rapid yield increase across various semiconductor process nodes at the manufacturing level, and establish a scientific engineering blueprint for transforming semiconductor factories into adapting autonomous intelligent manufacturing.

3.3 Post-Manufacturing Defect Detection

After the wafer fabrication is finished, we need to understand and search defects quickly to do the electrical test. Defect attribution and location are quite crucial for decreasing the cycle of rework and increasing yield rate. Recently, lots of research work attempted to achieve two different purposes, i.e., detecting defect quickly with image and attribution quickly with information from multiple sources.

Ullah et al. designed a framework for defect detection which could be used to process wafers with optical or simplified SEM images based on lightweight CNN, inverse feature matching and masking method. This framework employs MobileNetV2 as its backbone to extract global features, followed by inverse feature matching using a "reference normal template." It measures anomalies by calculating Euclidean distances. Additionally, it performs masking operations on known non-defect areas such as process variations, retaining only genuine defect regions. These are visualized as red masks. Experiments demonstrated that on a total of 493 images of chips with actual defects, this method achieved an accuracy rate of 82.4% and a precision rate of 86.6%. The latency for inference on a single image was 3.8 milliseconds, meeting the cycle time requirements for high-throughput reinspection after manufacturing[8].

This method successfully reduced defect detection from hours of manual visual inspection to mere seconds of algorithmic detection, providing highly reliable defect coordinates and type information for subsequent root cause analysis.

3.4 Root Cause Analysis of Defects

Durowoju & Olowonigba proposed a model built on supervised and unsupervised learning, making use of explainable AI like SHAP and LIME tools to quantify the contributions of process parameters to yield [7].

Lee & Roh built an interpretable framework to attribute yields based on multi-source data, and used machine learning as the main modeling method [9]. The core purpose of their framework is not only to locate each wafer after EDS, but also to interpret what might have caused any defects in the yield. Their framework scales and makes the process comprehensible using inputs from multiple sources to a model that is scalable. It provides a globally explainable view through its SHAP characteristics, which makes its models fully transparent. This uniformly encodes 983 dimensions of different features

ISSN 2959-6157

such as processing condition, equipment IDs, durations, sensor readings, etc. After using Random Forest to predict the yield after EDS, it quantifies the positive and negative contribution of every feature's influence to yield via Tree-SHAP waterfall plots.

Both frameworks leverage machine learning frameworks and explainable AI to provide engineers with an intuitive understanding of "which process parameters affect yield and how," thereby guiding the changes in the manufacturing process and production planning.

4. Advantages and Development Trends of Artificial Intelligence Algorithms

4.1 High Efficiency

We adopt the RL-based parameter tuning engine as core, connect it with Jiang's yield prediction model and a defect detection network using Faster R-CNN[10], map the data in real time into digital twin space, realizing the micro-second closed-loop of predicting- confirming-doing. That is to say, with full-link digital twin, we can perform the zero-delay simulation between the physical field and virtual field, realize synchronization, and greatly improve efficiency.

4.2 Integration

By integrating CNN features from industrial chip images, SHAP attribution for defect samples, and textual information such as engineer notes and manufacturing logs, we train a unified multimodal large model that combines images, time series, and text. Then the unified model is capable of achieving defects detection, defects explanation and defect solving at the same time. This means overall defect troubleshooting time will be reduced. With the help of an unified user interface, engineers can have easier and more convenient access to related functions in the model.

4.3 Autonomous

The real-time control explored in Durowoju & Olowonigba's research, combined with the SHAP explainability mentioned by Lee & Roh, enables AI systems to progressively achieve autonomous process optimization. This evolution shifts AI from a "decision-supporting" role toward a "self-decision-making" state. In next-gen fab, engineers don't have to adjust parameters or inspect the reasons once it is observed; they just type yield target for what they want. Once an anomaly is recognized and corrected by AI, full „look before you jump" cycle takes minutes rather than days: AI could fast identify, inform, execute remediation steps quickly, verify with closed-loop time frame!

4.4 Safety

The yield model can be trained among multiple fabs and a comprehensive summary about factors determining yield is obtained by encapsulating the yield grading model and SHAP attribution method as a federated learning client. For each foundry only needs to be trained once in local, exchanging gradients or SHAP summaries, and no foundry's sensitive data would be revealed, so there are no worries for leaking any foundry's confidential information (e.g. process parameters or layout details), achieving the purpose of "collaborating without compromising efficiency, sharing without compromising confidentiality."

5. Summary

Semiconductor manufacturing processes are getting more and more complicated, so the difficulty of yield management becomes increasingly huge. The rapid development of artificial intelligence (AI), machine learning (ML) and deep learning (DL) is important research directions to improve yields in semiconductor manufacturing fields, attracting much attention and intensive research. The economic impact and causes of the loss of semiconductor production yields were reviewed by this paper, and the technologies and progress related to the application of AI for improving yield as well as the prediction of their future were introduced.

Machine learning framework and technologies for pre-manufacturing process of yield prediction, which can enhance the efficiency of semiconductor manufacturing to a certain extent through the automatic processing of large-scale data with the aid of modeling algorithms, as well as control manufacturing costs to some extent, multi-layer machine learning and reinforcement learning framework technology and the realization methods for real-time process control during semiconductor manufacturing, which can control product defects in real time by using ML algorithm and prevent the generation of defective products, as well as reduce the loss rate of semiconductor production yields in real time. Lightweight CNN combined with inverse feature matching and mask-based defect detection frame technology, and scalable input and SHAP-explainable yield attribution frame technology to achieve a more precise localization position of defects and perform stronger analyses of defects in different scenarios and types so as to make the causes of defects in more complex scenarios clearer to further enhance semiconductor wafer yield rates.

In the future, along with the progress and development

of AI technology, semiconductor manufacturing yield improvement will develop towards an era based on data and intelligence. Further combining and deeply integrating prediction, discovery, explanation and optimization closed-loop intelligent manufacturing whole industry chain will be the next trajectory of future direction.

References

[1] Zhou Xiu. Research on Data-Driven Wafer Manufacturing Yield Prediction Methods. Henan University of Technology, 2024. DOI:10.27791/d.cnki.ghegy.2024.000604.

[2] Zheng Cheng. Research on Data-Driven Wafer Yield Control Methods. Donghua University, 2021. DOI:10.27012/d.cnki.gdhuu.2021.001259.

[3] Yang Fangtao. Application of Artificial Intelligence in Computer Network Technology. Industrial Innovation Research, 2023,(13):91-93. DOI:CNKI:SUN:CYCX.0.2023-13-029.

[4] Lai Andi, Liao Jun, Ou Di, et al. Sensitivity Study of Random Defects in Film/Substrate Systems. Chinese Journal of Solid Mechanics, 2024,45(05):652-664. DOI:10.19636/j.cnki.cjsm42-1250/o3.2024.021.

[5] Xu Hongwei, Zhang Jie, Lü Youlong, et al. Wafer Yield Prediction Method Based on Improved Continuous Deep Belief Network. Computer Integrated Manufacturing Systems, 2020,26(09):2388-2395. DOI:10.13196/j.cims.2020.09.008.

[6] D. Jiang, W. Lin and N. Raghavan, «A Novel Framework for Semiconductor Manufacturing Final Test Yield Classification Using Machine Learning Techniques,» in IEEE Access, vol. 8, pp. 197885-197895, 2020, doi: 10.1109/ACCESS.2020.3034680.

[7] Durowoju E S, Olowonigba J K. Machine Learning-Driven Process Optimization in Semiconductor Manufacturing: A New Framework for Yield Enhancement and Defect Reduction. International Journal of Advance Research Publication and Reviews, 2024, 1(4): 110-130. DOI: 10.55248/gengpi.6.0725.2579.

[8] Waseem Ullah, Samee Ullah Khan, Min Je Kim, Altaf Hussain, Muhammad Munsif, Mi Young Lee, Daeho Seo, Sung Wook Baik, Industrial defective chips detection using deep convolutional neural network with inverse feature matching mechanism, Journal of Computational Design and Engineering, Volume 11, Issue 3, June 2024, Pages 326–336, https://doi.org/10.1093/jcde/qwae019.

[9] Lee, Y., & Roh, Y. (2023). An Expandable Yield Prediction Framework Using Explainable Artificial Intelligence for Semiconductor Manufacturing. Applied Sciences, 13(4), 2660. https://doi.org/10.3390/app13042660.

[10] H. Hatem et al., «AI-Powered Defect Detection using Deep Learning: A Pattern-Agnostic Faster R-CNN Approach for SEM Images with GPU Acceleration,» 2025 36th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), Albany, NY, USA, 2025, pp. 1-5, doi: 10.1109/ASMC64512.2025.11010758.