International Journal of Innovation and Industrial Revolution
EISSN: 2637-0972
IJIREV
Volume 8 Issue 27 (December 2026) PP. 42-68



INTERNATIONAL JOURNAL OF INNOVATION AND INDUSTRIAL REVOLUTION (IJIREV)
www.gaexcellence.com/ijirev

DEEP LEARNING FOR WAFER DEFECT DETECTION AND CLASSIFICATION IN SEMICONDUCTOR PHOTOLITHOGRAPHY: A STRUCTURED REVIEW

Noraiza Aqilah Abdul Aziz[1], Zakaria Hussain[2*], Muhammad Khusairi Osman[3], Saiful Zaimy Yahaya[4]

[1]School of Electrical Engineering, Universiti Teknologi MARA (UiTM) Pulau Pinang, Malaysia
noraizaaqilah@gmail.com https://orcid.org/0009-0002-1159-250X
[2]School of Electrical Engineering, Universiti Teknologi MARA (UiTM) Pulau Pinang, Malaysia
zakaria183@uitm.edu.my https://orcid.org/0000-0002-2605-2621
[3]School of Electrical Engineering, Universiti Teknologi MARA (UiTM) Pulau Pinang, Malaysia
khusairi@uitm.edu.my https://orcid.org/0000-0002-5479-6409
[4]School of Electrical Engineering, Universiti Teknologi MARA (UiTM) Pulau Pinang, Malaysia
saiful053@uitm.edu.my https://orcid.org/0000-0002-6906-5086
*Corresponding Author

Article Info:

Article history:
Received date: 30.09.2026
Revised date: 15.10.2026
Accepted date: 04.11.2026
Published date: 01.12.2026

To cite this document:

Aziz, N. A. A., Hussain, Z., Osman, M. K., & Yahaya, S. Z. (2026). Deep Learning for Wafer Defect Detection and Classification in Semiconductor Photolithography: A Structured Review. *International Journal of Innovation and Industrial Revolution, 8*(27), 42-68.

Abstract:

The increasing complexity of semiconductor manufacturing and photolithography processes has intensified the need for accurate, efficient, and scalable wafer defect detection and classification. Conventional inspection approaches may face challenges in identifying complex, low-contrast, mixed-type, and previously unseen defect patterns, creating a growing need for deep learning-based solutions that can support automated and intelligent inspection. This systematic literature review examines the development of deep learning approaches for wafer defect detection and classification in semiconductor photolithography, with particular emphasis on emerging architectures, learning strategies, and industrial implementation. The review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. An advanced search strategy was applied to the Scopus and IEEE Xplore databases using the core keywords defect, detection, semiconductor, deep learning, and photolithography, covering publications from 2019 to October 2026. Following identification, screening, duplication removal, eligibility assessment, and quality evaluation, 15 primary studies were retained for qualitative synthesis. The findings were systematically organized into three themes: Theme 1, Lightweight Network Topologies and

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 27 (December 2026) PP. 42-68

Real-Time Object Detection, highlighting computationally efficient architectures and real-time inspection capabilities; Theme 2, Self-Supervised Representation, Vision–Language Alignment, and Data-Efficient Adaptation, emphasizing approaches that address limited labelled data and improve model adaptability; and Theme 3, Industrial AOI Hardware Integration, Multi-Scale Segmentation, and Benchmark Evaluations, focusing on deployment-oriented inspection systems, precise defect localization, and standardized performance assessment. Collectively, the findings indicate a research trajectory from increasingly efficient deep learning architectures toward data-efficient learning and industrially integrated inspection frameworks. The review provides a structured synthesis of current developments and identifies key methodological directions for advancing reliable, scalable, and intelligent wafer defect inspection in semiconductor photolithography.

DOI: 10.35631/IJIREV.827003

Keywords:

Defect, Deep Learning, Detection, Photolithography, Semiconductor

© The authors (2026). This is an Open Access article distributed under the terms of the Creative Commons Attribution (CC BY NC) (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact ijirev@gaexcllence.com.

Introduction

Deep learning has rapidly transformed the landscape of semiconductor manufacturing, particularly in the domain of wafer defect detection and classification during photolithography—a critical step that directly impacts product yield and reliability. Traditional methods, such as manual inspection and classical image processing, are increasingly being replaced by advanced neural network architectures due to their superior accuracy, adaptability, and efficiency in identifying complex and nano-scale defects (Batool et al., 2021; Dey et al., 2025; Li et al., 2026; Sneha & Vijayaraghavan, 2025). In photolithography, even minor defects caused by particles or process errors can lead to significant yield losses, making robust and automated defect detection essential for maintaining high manufacturing standards (S. Lee & Chun, 2026; Li et al., 2026). Deep learning models, including convolutional neural networks (CNNs) and architectures like EfficientNet and Faster R-CNN, have demonstrated remarkable capabilities in learning intricate image features, enabling them to outperform conventional approaches in both detection and classification tasks (Dehaerne et al., 2022; Hatem et al., 2025; S. Lee & Chun, 2026). These models can process large volumes of wafer images, recognize subtle defect patterns, and adapt to evolving defect types, thus supporting the stringent quality requirements of modern semiconductor fabrication (Dey et al., 2025; Li et al., 2026).

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 27 (December 2026) PP. 42-68

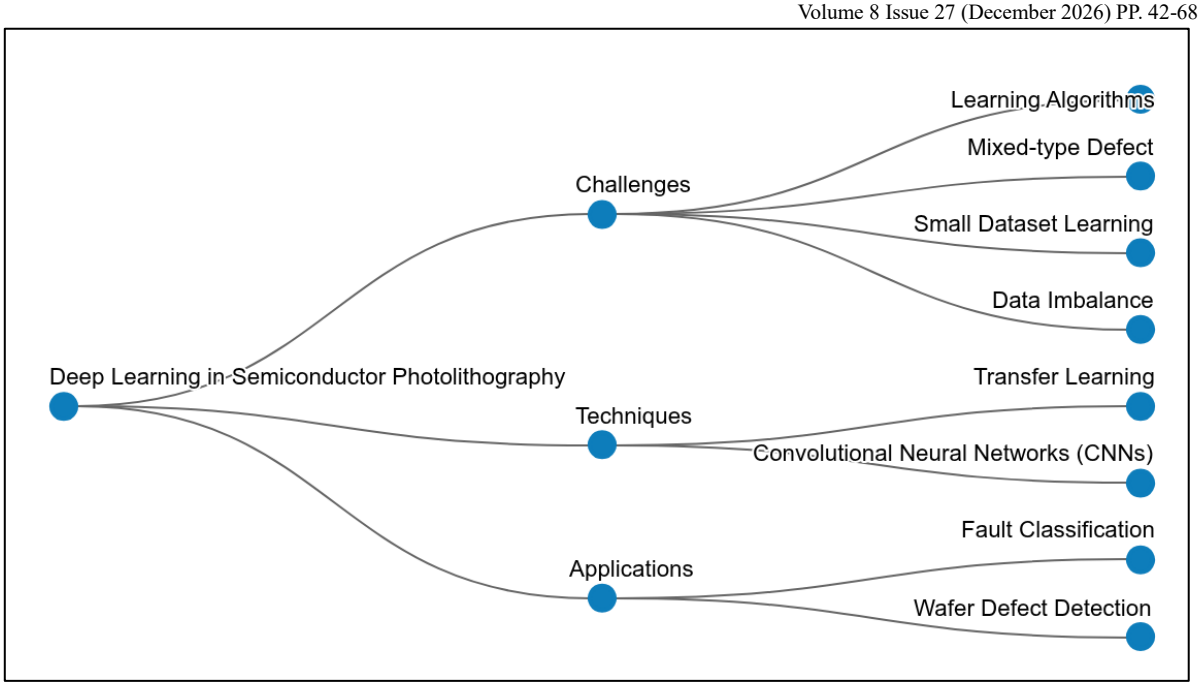


Figure 1: Key Research Themes and Concepts in Deep Learning in Semiconductor Photol

The adoption of deep learning in wafer defect detection and classification offers several key benefits. Automated frameworks not only accelerate the inspection process but also reduce human error and labor costs, leading to more consistent and reliable outcomes (Dey et al., 2022; O'Leary et al., 2020; Taha, 2026). Techniques such as data augmentation, transfer learning, and oversampling are employed to address challenges like limited labeled data and class imbalance, which are common in semiconductor defect datasets (Ganum et al., 2022; Prakash et al., 2026; N. Zhang & Mahmoud, 2025). Furthermore, deep learning-based systems facilitate real-time monitoring and rapid feedback, enabling timely corrective actions that minimize production losses and enhance overall yield (Dey et al., 2022, 2025). As semiconductor manufacturing continues to evolve, the integration of deep learning into photolithography defect inspection is poised to become the industry standard, driving advancements in both process control and product quality (Dey et al., 2025; Li et al., 2026; Sneha & Vijayaraghavan, 2025).

Literature Review

Deep Learning for Wafer Defect Detection

Deep learning for wafer defect detection and classification in semiconductor photolithography has become a central research direction because shrinking device geometries, larger wafers, and more complex process flows have increased defect frequency, visual complexity, and the cost of manual diagnosis (Gan, 2024; T.-K. Kim & Behdinan, 2022; Limam et al., 2024; Shinde et al., 2022, 2025). Across the literature, wafer maps, patterned wafer images, SEM images, and scanner topography maps are all treated as process-signature data that can support automated defect recognition, root-cause analysis, and yield improvement when interpreted accurately (Batool et al., 2021; Bose et al., 2025; Chen et al., 2023; Jeong et al., 2023; C.-Y. Lee et al., 2025). Reviews of the field agree that deep learning is displacing hand-crafted image processing and classical machine learning because it learns features directly from raw inputs,

handles complex nonlinear visual patterns, and scales better to modern inspection demands (Batool et al., 2021; Gan, 2024; Shinde et al., 2022; Yu et al., 2022). At the same time, the literature remains broader than simple classification: current work spans classification, localization, segmentation, mixed-defect recognition, data synthesis, transfer learning, and inline virtual metrology, reflecting the shift from offline wafer-map interpretation to production-facing inspection systems (Batool et al., 2021; Bose et al., 2025; Shinde et al., 2022).

The methodological progression is clear. Earlier learning-based systems depended on manually extracted geometry, density, texture, or Radon-transform features coupled with classifiers such as SVM, decision trees, KNN, random forests, or ensemble voting, but several papers note that the success of these pipelines depended heavily on choosing representative features in advance (He et al., 2025; Shinde et al., 2022; Yu et al., 2022). Deep CNNs changed that workflow by replacing feature engineering with end-to-end representation learning, often after only light preprocessing such as grayscale conversion or median filtering (He et al., 2025; Ingle et al., 2025; Jeong et al., 2023; Yu et al., 2022). CNNs are the dominant architecture in the surveyed literature, but systematic reviews emphasize that CNNs are not the only useful family, since GANs, autoencoders, transfer learning, and hybrid deep-learning-plus-classical models have also been used to address data synthesis, clustering, and weak-label settings (Batool et al., 2021). This diversification matters because different wafer inspection tasks impose different constraints: pattern classification favors compact image classifiers, defect localization favors object detectors, and photolithography metrology increasingly requires models that connect image-level predictions to process tools and die-level yield impact (Bose et al., 2025; Limam et al., 2024; Shinde et al., 2022).

Reported performance is consistently strong, but it depends on task definition and dataset composition. On the widely used WM-811K benchmark, end-to-end CNN variants routinely report accuracies in the mid-90% to high-90% range, including 98.34% for a dual-source DCNN, 96.4% for a dual-channel CNN with ECOC-SVM, 95.4% for the lightweight WM-PeleeNet, 95.09% for a CNN-feature ensemble, 97.05% for a recent classification model, and 99.70% for a compact CNN architecture (Biswas et al., 2022; Chen et al., 2021, 2022; Ingle et al., 2025; C.-Y. Lee et al., 2025; Yu et al., 2022). Performance gains have also come from multi-source fusion and attention mechanisms: a dual-source DCNN outperformed its single-source counterpart, while ESPP-Net, DAP-Net, and DMWMNet were specifically designed to improve recognition of mixed, weak, or overlapping defect patterns (Chen et al., 2022, 2023; He et al., 2025; X. Zhang et al., 2024). For mixed-type wafer maps, DAP-Net reported 98.6% accuracy on Mixed38WM and DMWMNet reported 98.99% accuracy with similarly high precision, recall, F1, and MCC, indicating that recent architectures are targeting the specific difficulty of multi-defect wafers rather than only clean single-label patterns (He et al., 2025; X. Zhang et al., 2024). Even so, classwise results remain uneven under imbalance: one recent CNN study reported much lower F1 for classes such as Scratch and Loc than for Edge-Ring or None, showing that aggregate accuracy can mask persistent minority-class weakness (N. Zhang & Mahmoud, 2025).

A major theme across the literature is that deployment bottlenecks now matter almost as much as raw accuracy. Several papers argue that many high-performing deep models are too large or too slow for embedded or real-time fab settings, motivating lightweight architectures such as WM-PeleeNet, MobileNet-derived models, shallow CNNs, and parameter-reduced networks (Gan, 2024; Yu et al., 2022). WM-PeleeNet is the clearest efficiency case, reducing parameters

by 92.68% and FLOPs by 58.85% while maintaining 95.4% average accuracy and processing 1000 wafer images in 22.99 seconds (Yu et al., 2022). Localization is another practical gap. Much of the older wafer-map literature emphasized classification alone, but YOLO-based studies show that object detection can add spatially actionable information, with YOLOv3 and YOLOv4 exceeding 94% classification accuracy on 19,200 wafer maps while also locating defects in real time, and a later YOLOv8 study reporting 96% overall accuracy on a small multi-class wafer dataset (Chen et al., 2023; Shinde et al., 2022). Segmentation extends this further by delineating defect regions rather than only assigning labels; recent work reported Dice 0.7731 and IoU 0.6590 for wafer defect segmentation alongside classification F1 of 0.9676, indicating that joint pixel-level and image-level modeling is becoming feasible (Ingle et al., 2025).

The most important unresolved problems are data scarcity, class imbalance, geometric variability, mixed defects, noise robustness, and industrial generalization. Reviews consistently identify limited labeled data, unknown defects, multi-label instances, and the need for more realistic industrial frameworks as open challenges (Batool et al., 2021; Gan, 2024). Class imbalance is especially recurrent, and papers address it through oversampling, minority-class augmentation, weighted ensembles, focal-loss variants, and synthetic generation; these strategies usually improve underrepresented classes, but they do not eliminate the gap entirely (C.-Y. Lee et al., 2025; Yu et al., 2022; N. Zhang & Mahmoud, 2025; X. Zhang et al., 2024). Geometric invariance is another targeted issue, because wafer labels are often unchanged by rotation and flip; transformation-invariant CNNs using Radon features, kernel flips, and relevance-propagation explanations improved robustness under scarce and out-of-distribution conditions (Y. Kim et al., 2023). In photolithography specifically, the literature is shifting from wafer-bin-map classification toward inline patterned-wafer and SEM inspection, where defects are smaller, noisier, and less well annotated: synthetic SEM data enabled YOLOv8 to reach 96% mAP and transfer to real data with 84.6% detection of Bridge defects and 78.3% of Break defects, while virtual metrology systems based on scanner topography maps promise full-wafer, stepwise gross defect monitoring, inline yield estimation, and tool-level root-cause ranking (Bose et al., 2025; Shinde et al., 2025). Overall, the literature shows that deep learning already delivers high-performance wafer defect detection and classification, but the next phase in semiconductor photolithography is not just better classifiers; it is robust, data-efficient, interpretable, and inline systems that can operate across scarce labels, tiny defects, mixed patterns, and real production variability (Batool et al., 2021; Bose et al., 2025; T.-K. Kim & Behdinan, 2022; C.-Y. Lee et al., 2025; Limam et al., 2024).

Research Question

The formulation of research questions (RQs) is a critical component of a Systematic Literature Review (SLR), as the questions establish the conceptual scope, analytical direction, and methodological boundaries of the review. Clearly defined RQs guide the systematic identification, selection, assessment, and synthesis of relevant studies, ensuring that the literature search remains focused and aligned with the research objectives. Well-constructed questions also provide a coherent framework for organizing and interpreting evidence, facilitating the identification of established research patterns, emerging developments, methodological approaches, and knowledge gaps. In addition, explicit RQs enhance the transparency and reproducibility of the review by establishing a clear link between the search strategy, eligibility criteria, data extraction, and subsequent synthesis of findings.

International Journal of Innovation and Industrial Revolution
EISSN: 2637-0972 IJIREV
Volume 8 Issue 27 (December 2026) PP. 42-68

Research question formulation is therefore undertaken during the planning stage of an SLR because the scope and structure of the RQs directly influence the methodological decisions applied throughout the review process (Kitchenham, 2007). In the present study, the PICo framework proposed by (Lockwood et al., 2015)was adopted to develop focused and contextually relevant research questions. PICo comprises three components: Population (P), Interest (I), and Context (Co), providing a structured approach for defining the target population, phenomenon or intervention of interest, and specific research context. The framework is particularly appropriate for reviews that seek to synthesize heterogeneous evidence and explore the development of a research domain. For this review, the PICo components were aligned with the identified thematic areas concerning deep learning for wafer defect detection and classification in semiconductor photolithography. Accordingly, three research questions were formulated to systematically examine lightweight network architectures and real-time detection, data-efficient representation and adaptation strategies, and industrial AOI integration, segmentation, and benchmarking, as presented below.

- RQ1: How are lightweight deep learning network topologies applied to improve the accuracy, computational efficiency, and real-time detection of wafer defects in semiconductor photolithography?
- RQ2: How do self-supervised representation learning, vision–language alignment, and data-efficient adaptation approaches address limited labelled data in wafer defect detection and classification?
- RQ3: How are deep learning-based wafer defect detection and classification methods integrated with industrial AOI hardware, multi-scale segmentation, and benchmark datasets to support reliable semiconductor photolithography inspection?

Material And Methods

Systematic literature reviews were conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, an internationally recognized guideline for ensuring transparency, methodological rigor, and reproducibility in evidence synthesis (Page et al., 2021). PRISMA provides a structured protocol for the systematic identification, screening, eligibility assessment, and selection of relevant studies, thereby strengthening the reliability and validity of the review process. In the present study, Scopus and IEEE Explre were employed as the primary bibliographic databases because of their extensive coverage of peer-reviewed scholarly literature and established indexing capabilities. The use of multiple complementary databases also broadens literature coverage and reduces the likelihood of database-specific retrieval bias.

The review process comprised four sequential stages: identification, screening, eligibility, and data abstraction. During identification, a predefined search strategy was applied to retrieve potentially relevant records from the selected databases. Retrieved records were subsequently screened against predetermined inclusion and exclusion criteria, with duplicate and clearly irrelevant records removed. In the eligibility stage, the full texts of potentially relevant studies were assessed systematically to determine their compliance with the predefined criteria. Finally, data abstraction involved the structured extraction, organization, and synthesis of information from the eligible studies to support the subsequent analysis and interpretation. This systematic workflow promotes consistency and traceability throughout the review process, while providing a transparent basis for synthesizing the available evidence and identifying research trends, knowledge gaps, and potential directions for future investigation.

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 27 (December 2026) PP. 42-68

Identification

The identification stage represents the initial phase of the PRISMA-based systematic literature review (SLR), in which potentially relevant records are comprehensively retrieved from predefined bibliographic databases using a structured search strategy. In this study, Scopus and IEEE Xplore were selected as the primary sources because they provide substantial coverage of peer-reviewed research in semiconductor manufacturing, photolithography, defect detection, and artificial intelligence. The search strategy was developed around the core concepts of defect, detection, semiconductor, deep learning, and photolithography, thereby establishing a focused retrieval scope aligned with the review objectives. The search yielded 339 records in total, comprising 183 records from Scopus and 156 records from IEEE Xplore. Scopus contributed approximately 54.0% of the identified records, while IEEE Xplore accounted for 46.0%. The relatively close distribution between the two databases indicates that relying on a single source could potentially omit a substantial proportion of relevant literature, particularly studies published within engineering and technology-oriented venues.

The combined retrieval of 339 records provides a broad initial evidence base for subsequent screening and eligibility assessment. The higher number of records retrieved from Scopus may reflect its multidisciplinary coverage, encompassing research across engineering, computer science, materials science, and manufacturing, whereas IEEE Xplore provides strong representation of technical research in electrical engineering, electronics, computing, and intelligent systems. The use of both databases is therefore justified by their complementary disciplinary coverage and is consistent with the objective of reducing database-dependent retrieval bias. At the identification stage, however, the 339 records should be interpreted as potentially relevant records rather than confirmed eligible studies, as duplicates, irrelevant publications, and studies failing the predefined inclusion criteria have not yet been excluded. These records consequently form the initial pool for the subsequent PRISMA screening and eligibility stages.

Table 1: The Search String

Scopus	TITLE-ABS-KEY (("semiconductor wafer" OR "wafer defect" OR "wafer quality" OR "wafer inspection") AND ("deep learning" OR "machine learning" OR "neural network*" OR AI) AND ("auto detection" OR "automated detection" OR "automatic detection" OR "defect detection")) AND (LIMIT-TO (PUBYEAR , "2019") OR LIMIT-TO (PUBYEAR , "2021") OR LIMIT-TO (PUBYEAR , "2022") OR LIMIT-TO (PUBYEAR , "2023") OR LIMIT-TO (PUBYEAR , "2024") OR LIMIT-TO (PUBYEAR , "2025") OR LIMIT-TO (PUBYEAR , "2026") OR LIMIT-TO (PUBYEAR , "2027")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (OA , "all") OR LIMIT-TO (OA , "publisherfullgold")) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "ENGI") OR LIMIT-TO (SUBJAREA , "MATE")) Date of Access: Oct 2026

IEEE	("wafer defect*" OR "wafer inspection" OR "wafer defect detection") AND (semiconductor* OR "semiconductor manufacturing") AND (photolithograph* OR lithograph*) AND ("deep learning" OR "machine learning" OR "artificial intelligence" OR "neural network*") AND ("image processing" OR "computer vision") Date of Access: Oct 2026

Screening

During the screening stage, the 339 records identified from Scopus and IEEE Xplore were evaluated against predefined inclusion and exclusion criteria to improve the relevance and consistency of the evidence base. Records were excluded if they were non-English publications, published before 2018, classified as books, review articles, or in-press records, outside the Computer Science, Engineering, and Materials Science subject areas, or not available through Gold or Open Access routes. Based on these criteria, 295 records were excluded, leaving 44 records for further assessment, comprising 28 records from Scopus and 16 records from IEEE Xplore. This substantial reduction is justified because the screening criteria were designed to maintain a focused evidence base aligned with the study scope, particularly research addressing technical developments in semiconductor and photolithography-related defect detection. The relatively greater contribution from Scopus was retained because of its broader multidisciplinary coverage, while IEEE Xplore provided complementary engineering and computing literature.

A subsequent deduplication process was conducted to ensure that the same study retrieved from both databases was not counted more than once. Three duplicate records were removed, resulting in 41 unique records proceeding to the subsequent eligibility assessment. This step is essential for preventing duplicate evidence from disproportionately influencing the synthesis and for ensuring that the final study count accurately represents distinct scholarly contributions.

Table 2: The Selection Criterion Is Searching

Criterion	Inclusion	Exclusion
Language	English	Non-English
Time line	2019 – 2026	< 2018
Literature type	Journal (Article) Conference	Book, Review
Publication Stage	Final	In Press
Subject Area	Computer Science, Engineering & Material Science	Besides Computer Science, Engineering & Material Science
Open Access	Gold Open Access	Besides Gold & Open Access

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 27 (December 2026) PP. 42-68

Eligibility

During the eligibility stage, the remaining 41 unique records were subjected to a more detailed assessment to determine their suitability for inclusion in the systematic literature review. Full-text articles were evaluated against the study objectives, with particular attention given to research relevance, scope, and accessibility. A total of 26 records were excluded because they were considered out of field, had titles that were not sufficiently relevant to the research focus, presented abstracts that did not correspond to the study objectives, or lacked accessible full-text content. These exclusions were necessary to ensure that the final evidence base directly addressed the review scope rather than merely matching the search terms. Following this assessment, 15 studies were retained for qualitative analysis, representing approximately 36.6% of the 41 records assessed for eligibility. The resulting dataset provides a more focused body of evidence for qualitative synthesis, while the explicit exclusion criteria strengthen methodological transparency and reduce the inclusion of studies with insufficient topical relevance or inadequate information for systematic evaluation.

Data Abstraction and Analysis

The study employed an integrative analytical approach to systematically examine, compare, and synthesise evidence from the selected literature on wafer defect detection in photolithography using deep learning. This approach was adopted to accommodate methodological and contextual variations across the included studies while enabling the identification of recurring patterns, research themes, and subthemes aligned with the review objectives. Following the eligibility assessment, 15 publications were included in the qualitative synthesis, as presented in Table 3. As illustrated in Figure 2, each publication was examined systematically to extract relevant findings, methodological characteristics, analytical approaches, and empirical evidence concerning the application of deep learning to wafer defect detection in lithography.

The analysis proceeded through several iterative stages. First, the selected articles were reviewed in detail to establish an understanding of their research objectives, datasets, deep learning techniques, evaluation strategies, and principal findings. Relevant evidence was then coded and compared across studies to identify similarities, differences, and emerging patterns. These codes were subsequently grouped into broader thematic categories and refined into coherent themes and subthemes based on their relevance to the research objectives and consistency across the evidence base. Analytical notes were maintained throughout the process to document interpretations, emerging relationships, and decisions made during theme development. The resulting thematic structure was reviewed collaboratively by the authors to assess its coherence and alignment with the underlying evidence. Where differences in interpretation occurred, they were resolved through discussion and consensus, thereby supporting analytical consistency, transparency, and rigour in the qualitative synthesis.

Quality of Appraisal

Following the systematic review guidelines proposed by (Kitchenham, 2007) Kitchenham and Charters (Kitchenham, 2007), a quality assessment was conducted after the selection of the primary studies to evaluate their methodological soundness and establish a consistent basis for comparison across the included evidence. In this study, the quality assessment framework developed by . (Abouzahra et al., 2020) Abouzahra et al. (2020) was adopted, comprising **six**

predefined quality assessment (QA) criteria relevant to the evaluation of the selected studies. Each criterion was assessed using a standardized three-level scoring scheme to ensure consistency and transparency in the appraisal process. A score of **1.0 (Yes, Y)** was assigned when a study fully satisfied the specified criterion, whereas **0.5 (Partly, P)** was assigned when the criterion was only partially satisfied or when identifiable methodological limitations were present. A score of **0 (No, N)** indicated that the study did not satisfy the criterion. This structured scoring procedure enabled the methodological quality of the included studies to be assessed systematically and provided a quantitative basis for comparing the strength and completeness of the available evidence.

The quality assessment for the 15 selected papers in the systematic literature review (SLR) was conducted across the six defined quality criteria:

1. **QA1:** Is the purpose of the study clearly stated?
2. **QA2:** Is the interest and usefulness of the work clearly presented?
3. **QA3:** Is the study methodology clearly established?
4. **QA4:** Are the concepts of the approach clearly defined?
5. **QA5:** Is the work compared and measured with other similar work?
6. **QA6:** Are the limitations of the work clearly mentioned?

The quality assessment (QA) process was conducted using a set of predefined evaluation criteria. Each selected study was independently assessed by three experts, who evaluated the study against each criterion and assigned one of three ratings: Yes (Y), Partly (P), or No (N). The scores assigned by the experts were subsequently aggregated to obtain an overall quality score for each study. To ensure the inclusion of high-quality studies, a study was considered eligible for the subsequent stage of the review only if its total aggregated score, calculated from the assessments of all three experts, exceeded 3.0. This threshold was established to maintain the methodological rigor and reliability of the studies included in the review.

International Journal of Innovation and Industrial Revolution
IJIREV
EISSN: 2637-0972
Volume 8 Issue 27 (December 2026) PP. 42-68

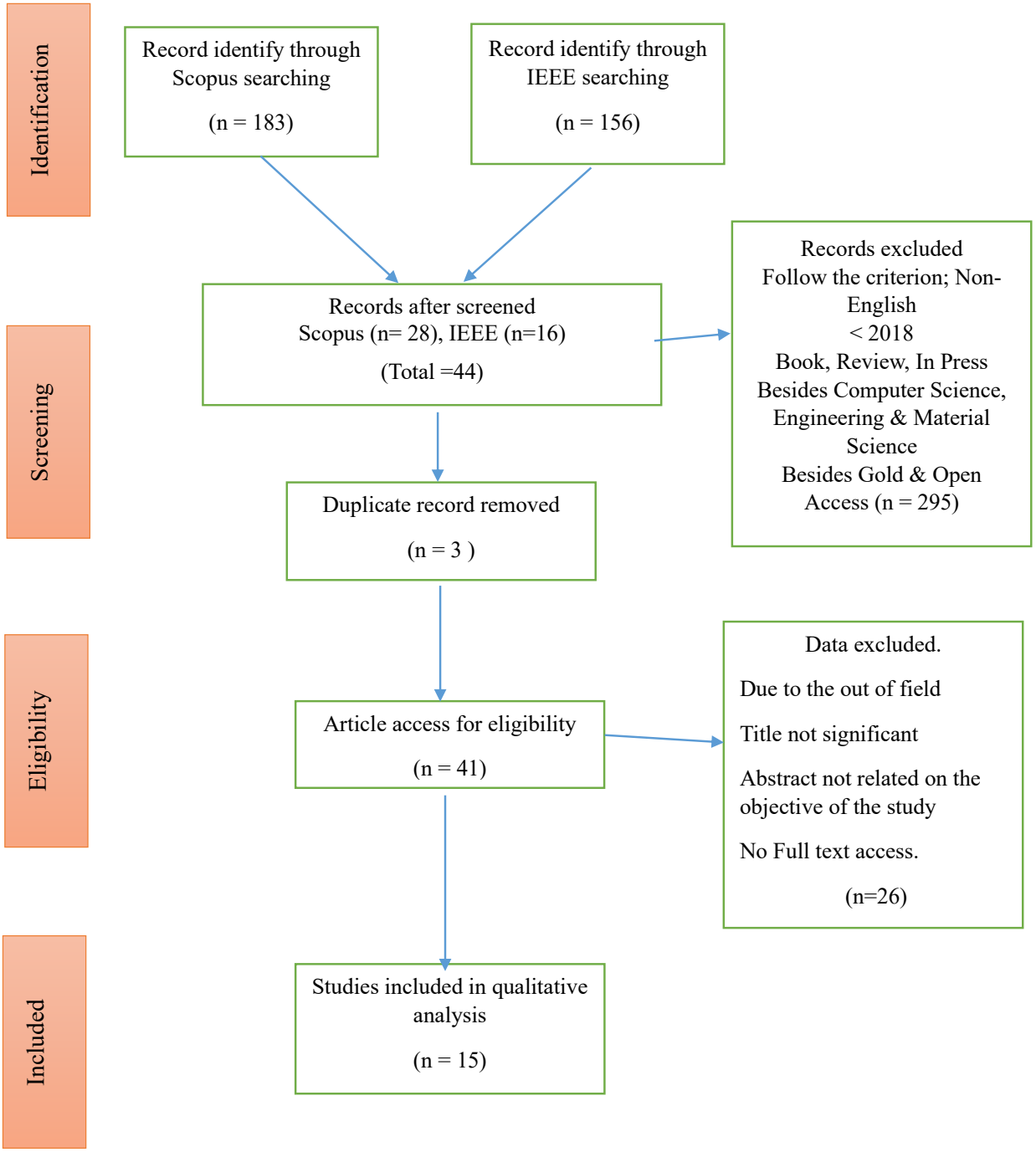


FIGURE 2. Flow Diagram of The Proposed Searching Study Semiconductor Wafer Defect Deep Learning Auto Detection [1]

Result and Discussion

Quality Assessment Analysis

Overall, the selected papers exhibit exceptionally high methodological quality, with every study clearly articulating its core research objectives (QA1) and demonstrating significant

practical utility in automated semiconductor inspection, yield enhancement, and real-time manufacturing quality control (QA2). The studies establish rigorous experimental methodologies (QA3) and precise conceptual frameworks (QA4) by providing detailed neural network architectures, custom loss functions, and transparent dataset partitioning protocols. Furthermore, extensive comparative benchmarking against established machine learning algorithms and state-of-the-art deep learning baselines (QA5) is a consistent strength across the body of literature. While a minor subset of papers addresses constraints only partially, the majority thoroughly examine key research limitations (QA6), explicitly outlining challenges such as extreme class imbalance, computational overhead on edge devices, sensitivity to training schedules, and low-density or mixed-type defect pattern misclassifications.

Quality Assessment Summary Table

Table 3: Number and details of Primary Studies Database

Author	Title and Year	QA1	QA2	QA3	QA4	QA5	QA6	Total Mark	Percentage (%)
Ji-Hee Lee & Jee-Hyong Lee	*A Reliable Defect Detection Method for Patterned Wafer Image Using Convolutional Neural Networks with the Transfer Learning* (2019) (J. Lee & Lee, 2019)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**
Chung-Jen Fu, Hsuan-Lin Chen, & Huo-Yen Tseng	*Application of Artificial Intelligence to Improve Chip Defect Detection Using Semiconductor Equipment* (Fu et al., 2025)	**Y**	**Y**	**P**	**P**	**P**	**N**	**3.5 / 6**	**58.33%**
Naigong Yu et al.	*Wafer map defect patterns classification based on a lightweight network and data augmentation* (Yu et al., 2022)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**
Y. Chi, X. Gong, B. Zhao, & L. Yao	*CISC-YOLO: A Lightweight Network for Micron-Level Defect Detection on Wafers via Efficient Cross-Scale Feature Fusion* (Chi, 2025)	**Y**	**Y**	**Y**	**Y**	**Y**	**P**	**5.5 / 6**	**91.67%**
Dayang N.F. Abdul Hamid et al.	*Detection of Monocrystalline Silicon Wafer Defects Using Deep Learning Transfer*	**Y**	**Y**	**Y**	**Y**	**P**	**P**	**5.0 / 6**	**83.33%**

	Learning (Ganum et al., 2022)								
Jessnor Arif Mat Jizat et al.	*Evaluation of the machine learning classifier in wafer defects classification* (Arif et al., 2021)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**
Muhammad Awais, Octavian Adrian Postolache, & Sancho Moura Oliveira	*Graph-based contrastive learning for self-supervised semiconductor wafer defect detection* (Awais et al., 2026)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**
Kamal Taha	*Observational and experimental insights into machine learning-based defect classification in wafers* (Taha, 2026)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**
Youngjae Kim, Jee-Hyong Lee, & Jongpil Jeong	*Pre-trained CNN-based TransUNet Model for Mixed-Type Defects in Wafer Maps* (Y. Kim et al., 2023)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**
Jessnor Arif Mat Jizat et al.	*Evaluation of the machine learning classifier in wafer defects classification* (Arif et al., 2021)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**
Ivo Façoco, Rafaela Carvalho, & Luís Rosado	*Towards efficient wafer visual inspection: Exploring novel lightweight approaches for anomaly detection and defect segmentation* (Façoco et al., 2025)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**
Monnet et al. / Int. J. Data Sci. Anal.	*Vision-based wafer inspection in semiconductor manufacturing: a case study* (Jesús et al., 2026)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**
Prashant P. Shinde, Priyadarshini	*Wafer Defect Localization and Classification Using*	**Y**	**Y**	**Y**	**Y**	**Y**	**P**	**5.5 / 6**	**91.67%**

P. Pai, & Shashishekar P. Adiga	*Deep Learning Techniques* (Shinde et al., 2022)								
Chen Tang, Lijie Yin, & Yongchao Xie	*Wafer surface defect detection with enhanced YOLOv7* (Tang et al., 2024)	**Y**	**Y**	**Y**	**Y**	**Y**	**P**	**5.5 / 6**	**91.67%**
Z. Wang et al.	*YOLO-LA: Prototype-Based Vision–Language Alignment for Wafer Defect Recognition* (Wang et al., 2026)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**6.0 / 6**	**100%**

Thematic Analysis

The thematic taxonomy for this systematic literature review (SLR) was synthesized through an in-depth analysis of the 15 selected source entries in the project, examining their research objectives, underlying neural network topologies, dataset characteristics, and industrial deployment paradigms. By analyzing recurring methodological patterns, theoretical formulations, and practical manufacturing bottlenecks across the full body of literature, three major, non-overlapping research themes were derived: (1) **Lightweight Network Topologies and Real-Time Object Detection**, which focuses on spatial bounding-box localization, parameter compression, and high-throughput edge deployment; (2) **Self-Supervised Representation, Vision–Language Alignment, and Data-Efficient Adaptation**, which addresses severe label scarcity, complex mixed-type defect patterns, and cross-domain knowledge transfer; and (3) **Industrial AOI Equipment Integration, Multi-Scale Anomaly Segmentation, and Benchmark Evaluations**, which centers on physical hardware integration, overkill/false-alarm reduction, high-resolution pixel-level segmentation, and systematic algorithmic benchmarking. Each of the 15 source entries (comprising 14 unique research studies and one duplicate dataset entry) has been assigned to its dominant theme based on its primary contribution and core findings.

Theme	Focus Area	Number of Articles	Main Research Direction
Theme 1: Lightweight Network Topologies and Real-Time Object Detection	Spatial bounding-box localization, parameter compression, cross-scale feature fusion, and high-throughput single-stage detection.	**5**	Designing compact single-stage object detectors (e.g., YOLO variants, WM-PeleeNet) and synthetic training pipelines to achieve low-latency, high-accuracy spatial localization of wafer surface defects.
Theme 2: Self-Supervised Representation, Vision–Language	Unsupervised pre-training, multimodal vision–language prototypes, graph neural	**5**	Overcoming severe label scarcity, extreme class imbalance, and mixed-defect complexity by leveraging

Alignment, and Data-Efficient Adaptation	networks, and transfer learning fine-tuning.		geometric deep learning (GNNs), vision–language alignment, and contrastive pre-training.
Theme 3: Industrial AOI Hardware Integration, Multi-Scale Segmentation, and Benchmark Evaluations	Equipment-level deployment, overkill/false-alarm reduction, pixel-level semantic segmentation, and systematic algorithmic benchmarking.	**5**	Integrating deep learning modules into physical Automated Optical Inspection (AOI) equipment, benchmarking classic vs. deep classifiers, and applying high-resolution anomaly segmentation for yield management.

Lightweight Network Topologies and Real-Time Object Detection

Lightweight network topologies and real-time object detection models have become crucial for semiconductor wafer inspection due to memory and latency constraints in industrial manufacturing environments. Heavy convolutional backbones and multi-stage detectors frequently exhibit parameter redundancy and high floating-point operations, preventing real-time execution on embedded edge hardware. To overcome these computational bottlenecks, (Shinde et al., 2022) Shinde et al. evaluated single-stage YOLO architectures, demonstrating that single-pass grid regression achieves real-time wafer defect prediction while eliminating the computational load of two-stage region proposal networks. Further optimizing lightweight feature extractors, (Yu et al., 2023) Yu et al. constructed WM-PeleeNet by pruning the PeleeNet backbone, employing small convolution kernels in a stem block, dense connectivity, transition layers, and a Convolutional Block Attention Module to compress parameters down to 0.169M and computational load to 0.316 GFLOPs. Similarly, (Chi, 2025) Chi et al. designed CISC-YOLO on the YOLOv8 framework by integrating lightweight IRB-GhostConv-C2f modules into the backbone, which doubles channel dimensions via cheap linear transformations and depthwise convolutions to minimize feature redundancy down to 1.92M parameters and 6.0 GFLOPs. In another structural refinement, Tang et al. enhanced YOLOv7 by incorporating Space-to-Depth Convolution into the MP module, replacing strided convolutions to prevent critical low-level feature loss during sub-sampling while maintaining high operational efficiency. These collective architectural innovations reflect a unified research movement toward replacing parameter-heavy networks with low-latency convolutional backbones specifically engineered for high-throughput wafer inspection platforms.

Beyond basic image classification, spatial bounding-box localization, multi-scale feature aggregation, and attention mechanisms are essential for identifying fine morphological defects across complex wafer surfaces. Shinde et al. demonstrated that object detection algorithms like YOLOv3 and YOLOv4 provide direct spatial coordinates through grid-based bounding box regression, providing vital location information for root-cause diagnosis that non-localizing networks such as ResNet50 and DenseNet121 fail to offer. To enhance detection capability across diverse defect scales, Tang et al. integrated a Convolutional Block Attention Module into the backbone of YOLOv7, allocating channel and spatial attention weights while adopting the angle-aware SIoU loss function to improve bounding box regression accuracy for minor defects. Addressing multi-scale feature representation, (Chi, 2025) Chi et al. introduced a

CNN-based cross-scale feature fusion necking network paired with a spatial and channel synergistic attention module, alongside a dynamic detection head incorporating deformable convolution v3 to capture intricate geometric transformations of micro-defects. Furthermore, Monnet et al. implemented a hierarchical two-stage inspection strategy utilizing automated bounding box extraction from synthetic HSV mask renders, wherein Stage 1 isolates wafer presence and Stage 2 performs fine-grained scratch defect localization. As emphasized across these studies, combining multi-scale feature fusion with specialized attention mechanisms significantly reduces false positives and missed detections in complex manufacturing environments.

Quantitative benchmark evaluations demonstrate that lightweight single-stage detectors achieve exceptional trade-offs between recognition accuracy, model complexity, and inference speed. On the WM-811K benchmark dataset, (Shinde et al., 2022) Shinde et al. reported that YOLOv4 achieves over 94% classification accuracy and a mean Average Precision of 0.758, outperforming DenseNet121 at 92% and ResNet50 at 89% accuracy while providing real-time execution. Evaluating model compression on WM-811K, (Yu et al., 2023) Yu et al. established that WM-PeleeNet achieves an average classification accuracy of 93.6% with a rapid testing duration of 22.99 seconds per 1,000 wafer maps, outperforming MobileNetV2 and ShuffleNetV2 while reducing parameters by 92.68% compared to the unpruned PeleeNet baseline. Testing on the Wafer Surface Defect Dataset, (Chi, 2025) Chi et al. demonstrated that CISC-YOLO achieves 93.7% mAP50, 92.7% precision, and 90.8% recall at an inference speed of 140.5 frames per second, surpassing baseline YOLOv8n, MobileNetV3, and EfficientViT in both parameter economy and detection precision. Similarly, (Tang et al., 2024) Tang et al. confirmed that enhanced YOLOv7 reaches 92.5% mAP, 92.23% precision, 94.1% recall, and a processing frame rate of 136 frames per second on WM-811K, outperforming standard YOLOv7, YOLOv5, and Faster R-CNN by a factor of six in operational throughput speed. These empirical findings consistently prove that modern lightweight object detection architectures fulfill the dual industrial demands of low computational latency and high-precision quality control.

Addressing severe data scarcity, class imbalance, and manual labeling costs, procedural data simulation and advanced augmentation techniques play a fundamental role in bolstering model generalization and preventing overfitting. (Yu et al., 2023) Yu et al. implemented a convolutional autoencoder that incorporates Gaussian noise into the latent feature space to generate diverse synthetic wafer maps, expanding minority classes to 5,000 samples and boosting average classification accuracy on WM-811K by 1.8%, with notable accuracy gains in Donut and Scratch patterns. Expanding on synthetic data generation (Jesús et al., 2026) Monnet et al. developed a procedural pipeline utilizing Blender 3D rendering and domain randomization—varying lighting intensity, camera angles, and scratch geometries—to train YOLOv8m and YOLOv11m models without manual annotation, achieving 1.0000 F1-score in wafer presence detection and 0.9685 F1-score in real wafer scratch localization. However, Monnet et al. also observed that training beyond 100 to 150 epochs causes performance degradation due to synthetic feature hyper-specialization, highlighting the necessity of early stopping checkpoints. In parallel, (Chi, 2025) Chi et al. expanded their Wafer Surface Defect Dataset from 1,021 to 3,256 images via rotation, scaling, brightness adjustment, and flipping, demonstrating that structured dataset expansion stabilizes loss convergence and accelerates learning speed. Similarly, Shinde et al. underscored the necessity of balanced sample selection across imbalanced defect classes to minimize the generalization gap between training and validation loss curves. Collectively, these methodologies establish synthetic-to-real transfer

and systematic data augmentation as essential paradigms for scalable automated defect inspection.

Theme 1: Self-Supervised Representation, Vision–Language Alignment, and Data-Efficient Adaptation

The persistent scarcity of human-annotated defect images and severe class imbalances in semiconductor fabrication necessitate data-efficient paradigms that learn robust representations from limited or unannotated wafer datasets. To address the heavy manual labeling burden and long training durations required when introducing new wafer patterns, (J. Lee & Lee, 2019) Lee and Lee demonstrated that deep transfer learning with pre-trained convolutional neural networks achieves 97% classification accuracy using merely 10% of the training data and 5% of the initial training time on chemical mechanical planarization inspection images. Similarly, Iskandar et al. established that fine-tuning an ImageNet-pretrained MobileNetV2 architecture with unfrozen network parameters and global average pooling yields a 98.99% testing accuracy across six monocrystalline silicon wafer defect categories while drastically mitigating training data requirements and processing overhead in automated optical inspection workflows. Extending data efficiency into multimodal domain adaptation, (Wang et al., 2026) Wang et al. introduced YOLO-LA, a lightweight vision-language alignment framework that couples frozen YOLO visual backbones with text encoders like MiniLM to project natural language defect descriptions into fixed semantic prototypes, achieving superior classification accuracy on low-resolution wafer maps with fewer than 0.2 million trainable parameters. Furthermore, (Awais et al., 2026) Awais et al. addressed extreme label scarcity by engineering a self-supervised graph neural network that combines 8-connectivity die representation with domain-driven contrastive pre-training, maintaining a high classification accuracy of 95.7% on the WM-811K benchmark dataset even when restricting labeled training supervision to only 25% of the total dataset corpus.

Self-supervised pre-training and contrastive learning frameworks offer powerful mechanisms for extracting rich latent representations from unlabeled wafer maps, thereby eliminating dependence on labor-intensive manual annotations and subjective expert judgments in semiconductor quality control. Awais et al. developed a multi-criteria contrastive learning framework using Graph Isomorphism Networks with jumping knowledge connections, utilizing spatial pattern similarity, defect density proximity, structural pattern descriptors, and production lot relationships to establish positive-negative pairs without human labeling, which achieved an adjusted rand index of 0.89 in unsupervised clustering and 0.98 after supervised fine-tuning. In a complementary self-supervised approach for complex wafer patterns, (T.-K. Kim & Behdinan, 2022) Kim et al. integrated a multi-class N-pair contrastive loss function to pre-train the convolutional encoder of a hybrid TransUNet architecture, successfully enforcing feature embeddings of identical defect patterns to cluster closer together in latent space prior to cross-entropy fine-tuning. Incorporating textual semantic anchors into representation learning, (Wang et al., 2026) Wang et al. leveraged pre-trained language models to generate prototype embeddings from concise natural-language defect descriptions, optimizing a learnable projection head via cross-entropy contrastive alignment to regularize visual features and resolve visual ambiguities between morphologically similar defect patterns such as Edge-Loc and Edge-Ring. Additionally, (Ganum et al., 2022) Iskandar et al. highlighted that transferring pre-trained feature weights from large-scale visual repositories provides a superior initialization state, enabling rapid optimization convergence, robust pattern discrimination, and preventing feature distortion when adapting networks to specialized industrial wafer defect distributions.

To handle complex spatial topologies and multi-label defect compositions, advanced neural network encoders combine graph structures, vision transformers, and multimodal projections to capture multi-scale geometric relationships and global structural context across diverse fabrication runs. (Awais et al., 2026) Awais et al. implemented an 8-connectivity graph topological structure where each wafer die serves as a node, allowing message-passing operations to naturally capture irregular wafer boundaries and spatial continuity while preserving multi-scale structural context through jumping knowledge layer aggregations across network depths. For pixel-level spatial inference on complex defect maps, (T.-K. Kim & Behdinan, 2022) Kim et al. combined convolutional neural network feature extractors with transformer self-attention layers within a TransUNet framework, enabling the model to capture both local spatial details and long-range contextual dependencies across multi-type mixed defect distributions, reaching a 0.995 precision score. Enhancing spatial feature representations with natural language domain priors, Wang et al. demonstrated that aligning frozen YOLO visual feature maps with text-derived semantic prototypes through dynamic projection layers provides robust structural regularization, preventing global spatial misinterpretations on low-resolution wafer map inputs. Furthermore, (J. Lee & Lee, 2019) Lee and Lee evaluated deep convolutional backbones including Inception v3, VGG16, and VGG19, proving that deeper receptive fields and multi-scale inception blocks effectively extract high-level pattern representations from complex patterned wafer images, significantly outperforming traditional shallow convolutional networks in model stability, feature transferability, and convergence speed.

Automated data annotation, pseudo-labeling, and fine-tuning strategies significantly streamline model deployment and boost segmentation accuracy across industrial semiconductor inspection pipelines. (T.-K. Kim & Behdinan, 2022) Kim et al. introduced an automatic pseudo-defect masking technique based on 2-connectivity morphological pixel separation using Scikit-image, generating high-quality segmentation masks automatically and achieving an F1-score of 0.995 when segmenting two-type mixed defect patterns on the MixedWM38 dataset without manual labeling labor. Optimizing fine-tuning protocols for transfer learning models, (Ganum et al., 2022) Iskandar et al. demonstrated that unfreezing network parameters during training increased validation accuracy from 91.80% to 97.67% and reduced mean validation loss from 4.35% to 0.94%, confirming that adjusting pre-trained feature maps to domain-specific defect features is essential for peak operational performance. Similarly, (J. Lee & Lee, 2019) Lee and Lee analyzed layer-freezing strategies during fine-tuning, discovering that unfreezing the upper two convolutional blocks of pre-trained VGG and Inception models provided optimal adaptation for new wafer patterns, achieving 90% accuracy on unclassified defect samples that conventional inspection algorithms failed to evaluate. Complementing these fine-tuning strategies, (Awais et al., 2026) Awais et al. established that end-to-end fine-tuning of pre-trained self-supervised graph embeddings yields a classification accuracy of 98.6% and an adjusted rand index of 0.98, demonstrating that minimal supervised adaptation on contrastively learned representations achieves near-perfect alignment with ground-truth defect labels across production manufacturing environments.

Theme 2: Industrial AOI Hardware Integration, Multi-Scale Segmentation, and Benchmark Evaluations

Integrating artificial intelligence directly into industrial Automated Optical Inspection (AOI) equipment and machine vision hardware platforms represents a critical milestone in modern semiconductor yield management and manufacturing process control. Traditional automated optical inspection machinery frequently encounters high false-alarm and overkill rates

exceeding 20% due to highly reflective surface properties, complex pattern interference, and rigid thresholding rules, which causes severe product waste and high manual re-inspection costs. To resolve these hardware-level inspection bottlenecks, (Fu et al., 2025) Fu, Chen, and Tseng integrated deep Convolutional Neural Networks (CNN) directly into semiconductor production equipment pipelines, establishing automated image sorting and feedback re-learning mechanisms that automatically filter misjudged samples, thereby drastically reducing overkill quantities from 200 false non-good chips down to 13 units and increasing overall inspection accuracy to 94.0%. Similarly, (Façoco et al., 2025) Façoco, Carvalho, and Rosado explored hardware-compatible lightweight deep learning models on high-resolution 300mm wafer images with a 3.4 μm/pixel spatial resolution, implementing unsupervised anomaly detection algorithms like EfficientAD that achieve an image-level F1-score of 82.35% with an ultra-low inference latency of only 10.56 milliseconds per frame on industrial GPU setups. In parallel, (Arif et al., 2021) Mat Jizat et al. deployed transfer learning feature embedders on the commercial Jaeger machine vision platform, combining pre-trained InceptionV3 feature extraction with classical machine learning classifiers to streamline real-time industrial inspection without requiring full network retraining. Furthermore, (Taha, 2026) Taha documented how leading semiconductor manufacturers like Intel and GlobalFoundries successfully embedded ResNet50 computer vision systems and XGBoost classifiers directly into assembly and testing lines, replacing slow offline manual reviews with rapid inline defect screening across wafer thinning, sorting, and micro-packaging operations to maximize fab throughput.

To address complex multi-scale defect geometries and distinguish benign surface noise from fatal structural flaws, high-resolution semantic segmentation and dual-stage decision frameworks provide fine-grained spatial boundaries and severity assessments. Standard bounding-box detection models often fail when defects span across entire wafer dies or manifest as microscopic surface scratches that require precise pixel-level boundaries. Addressing this technical limitation, (Façoco et al., 2025) Façoco, Carvalho, and Rosado evaluated dense semantic segmentation architectures, demonstrating that UPerNet with a Swin Transformer backbone achieves superior multi-scale feature fusion with a 57.45 mask-level F1-score and an 81.47% weighted average recall across seven distinct defect categories. To prevent minor, non-fatal surface deviations from triggering unnecessary wafer rejections, (Façoco et al., 2025) Façoco, Carvalho, and Rosado further introduced a non-trainable dual-stage output fusion strategy that combines unsupervised anomaly heatmaps from EfficientAD with supervised defect severity classifications from UPerNet-Swin, thereby refining PASS/FAIL yield decisions based on whether identified defects are critical or recoverable. Highlighting the necessity of multi-scale contextual feature maps in pixel-level defect inspection, (Taha, 2026) Taha analyzed U-shaped CNN-Transformer hybrids like DeepSEM-Net, which combine local convolutional operations with global self-attention mechanisms to accurately segment both localized micro-pits and broad multi-die defect patterns. Furthermore, (Fu et al., 2025) Fu, Chen, and Tseng demonstrated that optimizing physical equipment lighting configurations—such as adjusting side-light angles to highlight surface textures and front-light angles to accentuate structural die outlines—substantially enhances image contrast, allowing deep learning segmentation modules to reliably delineate subtle scratches, discoloration layers, and probe indentations.

Rigorous algorithmic benchmarking and empirical performance comparisons across industrial wafer datasets are essential for selecting optimal model architectures that balance inspection precision with computational execution speed. Investigating classical machine learning

classifiers paired with deep feature representations, (Arif et al., 2021) Mat Jizat et al. evaluated k-Nearest Neighbours, Support Vector Machines, Stochastic Gradient Descent, and Logistic Regression on industrial wafer images processed through InceptionV3 embeddings, establishing that Ridge-regularized Logistic Regression achieves the highest testing accuracy at 88.0% and a validation accuracy of 86.9%. Furthermore, (Arif et al., 2021) Mat Jizat et al. emphasized that evaluating wafer inspection platforms solely on classification accuracy introduces severe bias due to non-defective sample inundation, proving that Logistic Regression maintains an optimal precision-recall balance (88.5% precision and 88.0% recall) with a minimal false-negative bias that prevents defective wafers from passing into downstream assembly. Providing a broader empirical benchmark across 15 classification methodologies, (Taha, 2026) Taha established that Residual Neural Networks (ResNet) achieve peak classification accuracy (99.0%) and F1-score (98.88%) on wafer map datasets, whereas gradient-boosted decision trees like XGBoost offer the most efficient operational compromise for high-throughput lines, attaining 94.8% accuracy with a rapid training duration of only 7.8 minutes. Supporting these benchmarking outcomes, (Fu et al., 2025)Fu, Chen, and Tseng confirmed that integrating CNN models into physical AOI lines elevates overall classification precision from 52.63% to 71.43% and raises the system F1-score from 0.6897 to 0.8333 during secondary feedback retraining tests.

Deploying automated visual inspection systems in real-world semiconductor fabrication facilities requires overcoming dataset class imbalances, representative data partitioning, and strict computational latency limits. Because non-defective wafers dominate actual manufacturing production lines, (Arif et al., 2021) Mat Jizat et al. structured validation datasets with a 2:1 ratio of good-to-defective images, proving that balanced precision-recall performance prevents over-fitting to majority non-defect classes in automated optical inspection tools. Addressing dataset split bias in low-sample manufacturing scenarios, (Façoco et al., 2025) developed an unsupervised dataset-splitting algorithm using DINO-ViT-S/8 feature manifolds and k-nearest-neighbor coverage metrics, achieving 96% visual feature coverage across train/test splits while maintaining strict multi-label class balance. To ensure seamless integration into high-throughput production lines, (Façoco et al., 2025) conducted computational efficiency evaluations, demonstrating that while UPerNet-Swin requires 45.10 milliseconds per image, the complete dual-stage EfficientAD and UPerNet-Swin pipeline executes in 55.66 milliseconds, remaining well within real-time industrial processing windows. Concurrently, Taha highlighted that leveraging hardware acceleration through GPU-backed TensorFlow environments and distributed edge computing enables deep learning algorithms to process high-dimensional wafer images continuously without creating production line bottlenecks. Ultimately, (Fu et al., 2025) Fu, Chen, and Tseng underscored that continuous human-machine collaboration and automated retraining loops are essential for maintaining model robustness when semiconductor manufacturing processes introduce novel defect signatures or modified layer patterns across evolving fabrication cycles.

Conclusion

This systematic literature review examined the development and application of deep learning for wafer defect detection and classification in semiconductor photolithography. The main purpose was to synthesize recent research on deep learning techniques that improve detection accuracy, computational efficiency, adaptability, and industrial applicability. The review followed the PRISMA framework and considered studies published from 2019 to 2026. Scopus and IEEE Xplore were used as the main databases,

with predefined inclusion and exclusion criteria. From an initial 339 records, screening, duplicate removal, eligibility assessment, and quality evaluation resulted in 15 studies for qualitative synthesis. The review was guided by three research questions focusing on lightweight deep learning architectures, data-efficient learning strategies, and integration with industrial automated optical inspection systems, segmentation, and benchmark evaluation.

The findings indicate that deep learning has developed from basic image classification toward more comprehensive and deployment-oriented wafer inspection solutions. Convolutional neural networks remain an important foundation for extracting visual features from wafer images. However, the reviewed studies also show increasing use of object detection, segmentation, transfer learning, self-supervised learning, graph-based learning, vision-language alignment, and hybrid architectures. This development indicates that wafer inspection is no longer focused only on determining whether a defect exists. Greater attention is now given to defect location, morphology, spatial characteristics, and industrial relevance. Such capabilities are important because photolithography produces highly detailed wafer patterns where small defects can affect manufacturing quality and yield.

Three main themes were identified from the reviewed studies. The first theme concerns lightweight network topologies and real-time object detection. The literature shows that computational efficiency has become increasingly important alongside detection accuracy. High-capacity models may achieve strong performance but can require considerable computational resources. Therefore, recent studies have explored compact architectures, parameter reduction, efficient convolutional operations, attention mechanisms, feature fusion, and single-stage object detectors. These methods aim to maintain good detection performance while reducing model complexity and inference time. Lightweight architectures are particularly relevant for high-throughput inspection and deployment on industrial or edge computing platforms.

The second theme focuses on self-supervised representation learning, vision-language alignment, and data-efficient adaptation. Limited labelled data is one of the major challenges in semiconductor defect inspection because annotation requires time, expertise, and detailed understanding of defect types. Class imbalance also makes it difficult for models to learn minority defect categories effectively. The reviewed studies address these problems through transfer learning, contrastive learning, self-supervised learning, graph-based representation learning, synthetic data, pseudo-labelling, and multimodal approaches. These techniques provide opportunities to use unlabelled or weakly labelled data and reduce dependence on large manually labelled datasets. The findings therefore show a gradual movement from fully supervised learning toward more data-efficient learning strategies.

The third theme relates to industrial automated optical inspection hardware, multi-scale segmentation, and benchmark evaluation. This theme highlights the difference between achieving good experimental results and achieving reliable industrial deployment. In practical inspection environments, models must operate under limitations related to processing speed, hardware resources, data distribution, and reliability. Segmentation methods are important when accurate defect boundaries and spatial information are required. The reviewed studies also demonstrate increasing integration of deep learning with automated optical inspection

systems to support automated defect recognition and manufacturing feedback. This suggests that research is progressively moving from isolated deep learning models toward more complete inspection frameworks.

The synthesis further indicates that model accuracy alone is not sufficient for evaluating semiconductor inspection systems. Computational complexity, inference speed, minority-class performance, localization capability, false alarms, dataset representativeness, and industrial compatibility also require consideration. The relationship between data strategy and model architecture is particularly important. Techniques such as data augmentation, synthetic data generation, transfer learning, and self-supervised learning can support model development when labelled semiconductor defect data are limited. Therefore, future inspection systems should consider both model design and the availability and quality of training data.

From a practical perspective, deep learning can support more consistent and automated wafer inspection while reducing dependence on manual inspection. Lightweight models can support deployment where computing resources are limited, while data-efficient approaches can reduce the effort required to create labelled datasets. Object detection and segmentation can provide more detailed information about defect location and characteristics. Integration with automated optical inspection systems may also support faster manufacturing feedback. However, practical deployment still requires consideration of class imbalance, false alarms, data drift, hardware limitations, and changes in defect patterns caused by variations in manufacturing processes.

Several limitations should be considered when interpreting the findings. The review mainly used Scopus and IEEE Xplore, meaning relevant studies indexed in other databases may not have been included. The restriction to English-language publications may also have limited the available evidence. In addition, only 15 studies were retained for qualitative synthesis, which limits the extent to which the findings represent the whole research field. Differences in datasets, defect categories, evaluation measures, experimental conditions, and hardware configurations also make direct comparison between studies difficult. Therefore, reported performance should be interpreted according to the specific dataset and inspection task rather than treated as directly comparable across all studies.

Future research should focus on standardized evaluation using larger and more industrially representative datasets. Greater attention is required for unseen defects, changing manufacturing conditions, domain adaptation, and continual learning. Synthetic-to-real approaches should also be further investigated to determine their ability to represent actual photolithography defects. Explainable deep learning may provide additional value where inspection results are used for manufacturing decisions. Benchmarking should also consider inference latency, memory consumption, class-wise performance, false-positive and false-negative rates, and robustness under changing data conditions rather than focusing mainly on overall accuracy.

Overall, the systematic review demonstrates that deep learning has become an important foundation for automated wafer defect detection and classification in semiconductor photolithography. The research direction is gradually moving from high-accuracy classification toward lightweight, data-efficient, spatially aware, and industrially integrated inspection systems. The main challenge is therefore not only improving predictive accuracy but also developing models that remain reliable under limited labelled data, imbalanced defect distributions, complex defect structures, changing manufacturing conditions, and real-time

requirements. Continued development of efficient architectures, advanced representation learning, segmentation, synthetic data, and industrial inspection integration can contribute to more scalable and reliable semiconductor wafer inspection systems.

Acknowledgements:	The authors would like to express their sincere gratitude to Universiti Teknologi MARA (UiTM) Pulau Pinang for providing the necessary resources and support throughout the course of this research. Special appreciation is extended to colleagues and peers who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.
Funding Statement:	No Funding
Conflict of Interest Statement:	The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have contributed to this work and approved the final version of the manuscript for submission to the International Journal of Innovation and Industrial Revolution (IJIREV).
Ethics Statement:	This study did not involve any human participants, animals, or sensitive data requiring ethical approval. The authors confirm that the research was conducted in accordance with accepted academic integrity and ethical publishing standards.
Author Contribution Statement:	All authors contributed significantly to the development of this manuscript. Author 1 was responsible for the conceptualization, methodology, and overall supervision of the study. Author 2 handled data collection, analysis, and interpretation of results. Author 3,4 contributed to the literature review, drafting, and critical revision of the manuscript. All authors read and approved the final version of the manuscript prior to submission.

References

Abouzahra, A., Sabraoui, A., & Afdel, K. (2020). Model composition in Model Driven Engineering: A systematic literature review. Information and Software Technology, 125(May), 106316. https://doi.org/10.1016/j.infsof.2020.106316

Arif, J., Jizat, M., Abdul, A. P. P., Fakhri, A., Nasir, A., Taha, Z., & Yuen, E. (2021). Evaluation of the machine learning classifier in wafer defects classification. ICT Express, 7(4), 535–539. https://doi.org/10.1016/j.icte.2021.04.007

Awais, M., Postolache, O. A., & Oliveira, S. M. (2026). Graph-based contrastive learning for self-supervised semiconductor wafer defect detection. Journal of Intelligent Manufacturing. https://doi.org/10.1007/s10845-026-02906-3

Batool, U., Shapiai, M., Tahir, M., Ismail, Z., Zakaria, N., & Elfakharany, A. (2021). A Systematic Review of Deep Learning for Silicon Wafer Defect Recognition. IEEE Access, 9, 116572–116593. https://doi.org/10.1109/access.2021.3106171

Biswas, S., Palanivel, D. A., & Gopalakrishnan, S. (2022). A Novel Convolution Neural Network Model for Wafer Map Defect Patterns Classification. 2022 IEEE Region 10 Symposium (TENSYMP), 1–6. https://doi.org/10.1109/tensymp54529.2022.9864391

Bose, P., Pandey, V. K., Staib, M., Womach, B., Balla, I., Olmedo, M., Khandelwal, A., Ahmadi, A., & Bisht, G. (2025). Virtual defect metrology system utilizing photolithography scanner topography maps to quantitatively detect gross defects, perform inline yield prediction, and identify defect-generating tools. Journal of Micro/Nanopatterning, Materials and Metrology, 24(4). https://doi.org/10.1117/1.JMM.24.4.041404

Chen, S., Huang, Z., Wang, T., Hou, X., & Jun. (2023). Wafer map defect recognition based on multi-scale feature fusion and attention spatial pyramid pooling. Journal of Intelligent Manufacturing, 36, 271–284. https://doi.org/10.1007/s10845-023-02231-z

Chen, S., Zhang, Y., Hou, X., Shang, Y., & Yang, P. (2022). Wafer map failure pattern recognition based on deep convolutional neural network. Expert Syst. Appl., 209, 118254. https://doi.org/10.1016/j.eswa.2022.118254

Chen, S., Zhang, Y., Yi, M., Shang, Y., & Yang, P. (2021). AI classification of wafer map defect patterns by using dual-channel convolutional neural network. Engineering Failure Analysis. https://doi.org/10.1016/j.engfailanal.2021.105756

Chi, Y. (2025). CISC-YOLO : A Lightweight Network for Micron-Level Defect Detection on Wafers via Efficient Cross-Scale Feature Fusion.

Dehaerne, E., Dey, B., & Halder, S. (2022). A Comparative Study of Deep-Learning Object Detectors for Semiconductor Defect Detection. ICECS 2022 - 29th IEEE International Conference on Electronics, Circuits and Systems, Proceedings. https://doi.org/10.1109/ICECS202256217.2022.9971022

Dey, B., Dehaerne, E., & Halder, S. (2022). Towards Improving Challenging Stochastic Defect Detection in SEM Images Based on Improved YOLOv5. In K. B.S. (Ed.), Proceedings of SPIE - The International Society for Optical Engineering (Vol. 12293). SPIE. https://doi.org/10.1117/12.2645402

Dey, B., Prasad, A., Belgharat, A., Blanco, V., & Halder, S. (2025). Towards Robust Defect Inspection in Advanced Node Semiconductors via Continual Learning. In S. M.J. & S. N.G. (Eds.), Proceedings of SPIE - The International Society for Optical Engineering (Vol. 13426). SPIE. https://doi.org/10.1117/12.3052286

Façoco, I., Carvalho, R., & Rosado, L. (2025). Intelligent Systems with Applications Towards efficient wafer visual inspection : Exploring novel lightweight approaches for anomaly detection and defect segmentation. Intelligent Systems with Applications, 28(April), 200576. https://doi.org/10.1016/j.iswa.2025.200576

Fu, C.-J., Chen, H.-L., & Tseng, H.-Y. (2025). Application of Artificial Intelligence to Improve Chip Defect Detection Using Semiconductor Equipment †. Engineering Proceedings, 98(1). https://doi.org/10.3390/engproc2025098026

Gan, W. (2024). Semiconductor Wafer Defect Detection Based on Machine Learning. Transactions on Computer Science and Intelligent Systems Research. https://doi.org/10.62051/wr098130

Ganum, A., Iskandar, D. N. A., Chin, L. P., & Fauzi, A. H. (2022). Detection of Monocrystalline Silicon Wafer Defects Using Deep Transfer Learning. Journal of Telecommunications and Information Technology, 2022(1), 34–42. https://doi.org/10.26636/jtit.2022.156321

Hatem, H., Essam, A., Samir, H., ElDessouki, A., Wu, S., Ahi, K., Greeneltch, N., Jayaram, S., & Fenger, G. (2025). AI-Powered Defect Detection using Deep Learning: A Pattern-Agnostic Faster R-CNN Approach for SEM Images with GPU Acceleration. ASMC (Advanced Semiconductor Manufacturing Conference) Proceedings. https://doi.org/10.1109/ASMC64512.2025.11010758

He, H., Wei, Y.-C., Lin, X., Zhu, M., & Zhang, H. (2025). Semiconductor Wafer Defect Recognition Based on Improved Coordinate Attention Mechanism. IEEE Access, 13, 46856–46864. https://doi.org/10.1109/access.2025.3546781

Ingle, R., Shahade, A., Gaikwad, M., & Patil, S. (2025). Deep learning driven silicon wafer defect segmentation and classification. MethodsX, 14. https://doi.org/10.1016/j.mex.2025.103158

Jeong, I., Lee, S., Park, K., Kim, I., Huh, H., & Lee, S. (2023). Wafer map failure pattern classification using geometric transformation-invariant convolutional neural network. Scientific Reports, 13. https://doi.org/10.1038/s41598-023-34147-2

Jesús, R., Peñaloza, S., Moghadam, A., & Davoudi, F. (2026). Vision-based wafer inspection in semiconductor manufacturing : a case study on scratch defect detection using synthetic data and YOLO models. 123.

Kim, T.-K., & Behdinan, K. (2022). Advances in machine learning and deep learning applications towards wafer map defect recognition and classification: a review. Journal of Intelligent Manufacturing, 34, 3215–3247. https://doi.org/10.1007/s10845-022-01994-1

Kim, Y., Lee, J.-H., & Jeong, J. (2023). Pre-trained CNN-based TransUNet Model for Mixed-Type Defects in Wafer Maps. WSEAS Transactions on Information Science and Applications, 20, 238–244. https://doi.org/10.37394/23209.2023.20.27

Kitchenham, B. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report, Ver. 2.3 EBSE Technical Report. EBSE.

Lee, C.-Y., Pleva, M., Hládek, D., Lee, C., & Su, M.-H. (2025). Ensemble Learning for Wafer Defect Pattern Classification in the Semiconductor Industry. IEEE Access, 13, 155714–155728. https://doi.org/10.1109/access.2025.3604405

Lee, J., & Lee, J. (2019). A Reliable Defect Detection Method for Patterned Wafer Image Using Convolutional Neural Networks with the Transfer Learning A Reliable Defect Detection Method for Patterned Wafer Image Using Convolutional Neural Networks with the Transfer Learning. https://doi.org/10.1088/1757-899X/647/1/012010

Lee, S., & Chun, K. (2026). Deep Learning Method for Wafer Flaw Detection in Lab-level Photolithography. Indonesian Journal of Electrical Engineering and Informatics, 14(2), 582–594. https://doi.org/10.52549/ijeei.v14i2.7975

Li, M., Su, H., Shi, C., Zhang, Y., Cheng, X., & Xie, Y. (2026). A Review on the Application of Deep Learning in Wafer Defect Detection. 2026 International Annual Conference on

Complex Systems and Intelligent Science, CSIS-IAC 2026, 301–306. https://doi.org/10.1109/CSIS-IAC70275.2026.11585140

Limam, K., Cheema, S., Mouhoubi, S., & Freijedo, F. (2024). Deep Learning-Based Visual Recognition for Inline Defects in Production of Semiconductors. IEEE Journal of Emerging and Selected Topics in Industrial Electronics, 5, 203–211. https://doi.org/10.1109/jestie.2023.3326092

Lockwood, C., Munn, Z., & Porritt, K. (2015). Qualitative research synthesis: Methodological guidance for systematic reviewers utilizing meta-aggregation. International Journal of Evidence-Based Healthcare, 13(3), 179–187. https://doi.org/10.1097/XEB.0000000000000062

O'Leary, J., Sawlani, K., & Mesbah, A. (2020). Deep Learning for Classification of the Chemical Composition of Particle Defects on Semiconductor Wafers. IEEE Transactions on Semiconductor Manufacturing, 33(1), 72–85. https://doi.org/10.1109/TSM.2019.2963656

Page, M. J., McKenzie, J. E., Bossuyt, P., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. Medicina Fluminensis, 57(4), 444–465. https://doi.org/10.21860/medflum2021_264903

Prakash, A., Patwari, U., Warade, S., Pradhan, A. K., & Pradhan, B. B. (2026). Data-Efficient Semiconductor Wafer Defect Identification Using WGAN-GP Augmentation and Lightweight Deep Convolutional Networks. IEEE Access. https://doi.org/10.1109/ACCESS.2026.3730064

Shinde, P., Pai, P., & Adiga, S. (2022). Wafer Defect Localization and Classification Using Deep Learning Techniques. IEEE Access, PP, 1. https://doi.org/10.1109/access.2022.3166512

Shinde, P., Pai, P., Adiga, S., Mayya, K., Seo, Y., Hwang, M., Go, H., & Park, C. (2025). Defect Detection in Photolithographic Patterns Using Deep Learning Models Trained on Synthetic Data. ArXiv, abs/2505.1. https://doi.org/10.1016/j.heliyon.2025.e43377

Sneha, A., & Vijayaraghavan, V. (2025). Deep Learning Approaches for Identifying Defective Silicon Wafers. 2025 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies, INSPECT 2025. https://doi.org/10.1109/INSPECT67393.2025.11350598

Taha, K. (2026). Observational and experimental insights into machine learning-based defect classification in wafers. Journal of Intelligent Manufacturing, 37(1), 45–95. https://doi.org/10.1007/s10845-024-02521-0

Tang, C., Yin, L., & Xie, Y. (2024). Wafer surface defect detection with enhanced YOLOv7 Wafer surface defect detection with enhanced Chen Tang , Lijie Yin * and Yongchao Xie. (March).

Wang, Z., Yang, Y., Chu, J., Zang, Y., She, Z., Fang, W., & Wang, R. (2026). YOLO-LA: Prototype-Based Vision–Language Alignment for Silicon Wafer Defect Pattern Detection. Micromachines, 17(1). https://doi.org/10.3390/mi17010067

Yu, N., Chen, H., Xu, Q., Hasan, M. M., & Ouattara, S. (2022). Wafer map defect patterns classification based on a lightweight network and data augmentation. CAAI Trans. Intell. Technol., 8, 1029–1042. https://doi.org/10.1049/cit2.12126

Yu, N., Chen, H., Xu, Q., Hasan, M. M., & Sie, O. (2023). network and data augmentation. (October 2021), 1029–1042. https://doi.org/10.1049/cit2.12126

Zhang, N., & Mahmoud, W. H. (2025). Semiconductor Wafer Map Defect Classification Using Convolutional Neural Networks on Imbalanced Classes. 17th International Conference on Advanced Computational Intelligence, ICACI 2025, 54–62. https://doi.org/10.1109/ICACI65340.2025.11096243

Zhang, X., Jiang, Z., Yang, H., Mo, Y., Zhou, L., Zhang, Y., Li, J., & Wei, S. (2024). DMWMNet: A novel dual-branch multi-level convolutional network for high-performance mixed-type wafer map defect detection in semiconductor manufacturing. Computers in Industry. https://doi.org/10.1016/j.compind.2024.104136