Design of a Multi-Stage Decision-Support Pipeline for Semiconductor Manufacturing Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

In semiconductor manufacturing, inspection and metrology decisions are constrained by high costs, limited throughput, and the progressive availability of information over time. As the process advances, richer measurements become accessible, while the opportunity to rework or correct prior steps diminishes. Consequently, the central challenge is not merely accurate defect or yield prediction at individual stages, but the design of a decision-support framework that determines when and where limited inspection resources should be allocated.

This paper proposes a staged decision-support pipeline spanning Stage 0 to Stage 3 that explicitly reflects this temporal structure. From Stage 0 to Stage 2A, stage-wise yield prediction models are constructed using only variables observable at each decision point, thereby preventing information leakage and ensuring operational consistency. Continuous yield predictions are translated into actionable decisions through quantile-based policies that define high-risk and scrap candidates under fixed capacity constraints. Downstream, Stage 2B and Stage 3 demonstrate how wafer map analysis and SEM-based defect morphology assessment can be organized into operationally meaningful candidate selection and triage records under limited inspection capacity.

To ensure claim validity, we introduce a two-layer governance design that separates operational workflow integration from scientifically validated claims. Quantitative performance claims are restricted to the validated core (Stage 0–2A), where same-source ground truth is available, while downstream stages are treated as proxy benchmarks demonstrating functional feasibility without end-to-end causal claims. Experimental results show that incorporating progressively enriched information improves yield prediction accuracy and that the proposed decision agent framework can enrich high-risk wafer selection under fixed inspection budgets. Overall, this work demonstrates that evidence-gated, staged integration provides a practical and claim-safe pathway for deploying machine learning–based decision support in capacity-constrained semiconductor manufacturing environments.

*Use footnote for providing further information, for less known open models (webpage, version)

†Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies

1 Introduction

In semiconductor manufacturing, inspection and metrology are essential for process stabilization. However, due to high costs and limited throughput, it is practically impossible to perform the same level of inspection on every unit. Particularly in the early stages, only a limited range of sensor and equipment data representing process conditions is available. While measurement precision increases as the process progresses, the ability to modify or undo previous process states simultaneously diminishes. Given these characteristics, the core challenge in manufacturing is not merely predicting defects, but designing a framework to determine when, where, and for which targets inspection resources should be allocated.

Existing research based on process data has primarily focused on improving prediction accuracy at individual stages or enhancing the classification performance of specific metrology results. However, this approach has limitations in reflecting the decision-making flow required in actual manufacturing environments. For instance, even if high predictive accuracy is achieved at a specific stage, the result offers limited value if it is not delivered in a form usable for the next stage or if it fails to consider operational constraints such as cost, throughput, and reworkability.

Based on this problem statement, this study proposes an end-to-end decision-support pipeline from Stage 0 to Stage 3, simulating the multi-stage inspection and metrology decision-making process in semiconductor manufacturing. Instead of treating each model output as an independent prediction, the proposed system standardizes and links them into "decision-ready outputs" that include risk scores, economic indicators, core evidence, and contextual information to be passed to the next stage. This allows subsequent decisions to be made with reference to the rationale of previous judgments. Accordingly, the contribution of this research lies in designing a multi-stage decision structure rather than treating manufacturing processes with varying costs as a single analytical step.

The structure of this paper is as follows: Section 2 defines the datasets used in each phase from Stage 0 to Stage 3. Section 3 details the methodology for each stage and the structure of the decision-making agent that integrates the analysis results across stages. Finally, Section 4 discusses conclusions and directions for future research.

2 Dataset

2.1 Stage 0 dataset

The Stage 0 dataset is composed based on sensor and equipment log information obtainable during the early stages of the process, with the objective of early detection of potential yield degradation risks. This stage corresponds to a preliminary phase conducted before in-line metrology or high-cost inspections in actual manufacturing environments; its goal is to screen high-risk candidates using only limited information.

In Stage 0, variables collectable at the start of the process are used as inputs. Numerical variables include pressure, temperature, exposure time, focus offset, dose, implant energy, and tilt angle. Categorical variables include Lot ID, Wafer ID, product type, technology node, and key equipment identifiers (e.g., etch_tool, litho_tool, deposition_tool, implant_tool).

Meanwhile, while final yield information is utilized as the target variable during the model training process, it is excluded from the Stage 0 model inputs since it is an unknown value at the time of prediction. Additionally, process dates were excluded as they correlate strongly with Lot ID identifiers, which could lead to redundant reflection of the same process information. Consequently, the Stage 0 model is constrained to perform decision-making strictly within the scope of information available in a real-world operating environment. Table 1 provides examples of the numerical data used in Stage 0, and Table 2 provides examples of the categorical data.

Table 1: Examples of numerical data

lot_id	wafer_id	product_type	technology_node	etch_tool	litho_tool	deposition_tool	implant_tool
LOT_0001	W001	CPU	10nm	ETCH_02	LITHO_01	DEP_03	IMP_01
LOT_0001	W002	CPU	10nm	ETCH_02	LITHO_01	DEP_03	IMP_01

Table 2: Examples of categorical data

pressure	temperature	exposure_time	focus_offset	dose	implant_energy	tilt_angle
149.8087	66.59538	1.576151	0.019829	1.01E+15	49.53053	6.768291
149.005	65.48352	1.400622	0.019902	9.95E+14	49.19151	7.457701

2.2 Stage 1 dataset

The Stage 1 dataset is constructed under the assumption that inline metrology results become available after the early screening performed at Stage 0. This stage corresponds to a phase in the manufacturing process where rework is still feasible, and aims to refine decision-making for the risk candidates identified at Stage 0 by incorporating additional metrology information.

The input data for Stage 1 include all process sensor variables and identification features used in Stage 0, along with additional physical measurement variables obtained through inline metrology. The newly introduced variables consist of critical dimension, oxide thickness, and thickness uniformity. These metrology measurements are not observable at Stage 0 and, in real manufacturing environments, require additional cost and processing time to acquire.

In practical operation, it would be reasonable to perform inline metrology only for the high-risk candidates identified at Stage 0. However, due to the constraints of conducting research based on publicly available datasets, the Stage 1 dataset in this study is constructed by augmenting the full Stage 0 dataset with the corresponding metrology variables. Table 3 presents examples of the additional data incorporated at Stage 1.

Table 3: Examples of data added at Stage 1 relative to Stage 0

critical_dimension	oxide_thickness	thickness_uniformity
21.30141	55.60491	1.673487
22.49313	53.6756	1.416346

2.3 Stage 2A dataset

The Stage 2A dataset is constructed under the assumption that a substantial portion of the manufacturing process has already been completed, such that the feasibility of rework is limited or significantly reduced. The objective of Stage 2A is not to determine whether to proceed with further processing, but rather to provide evidence for deciding which operational action immediate scrapping, additional analysis, or an attempt at rework is most appropriate.

The input data for Stage 2A include all process sensor variables, identification features, and inline metrology variables used up to Stage 1, with additional indicators related to etching and deposition processes. Specifically, etch rate and deposition rate are incorporated as newly added variables. Similar to the construction of the Stage 1 dataset, due to data availability constraints, the Stage 2A dataset in this study is formed by augmenting the full Stage 1 dataset with the corresponding etching and deposition variables. Table 4 presents examples of the Stage 2A data added relative to Stage 1.

Table 4: Examples of data added at Stage 2A relative to Stage 1

etch_rate	deposition_rate
3.559737	2.147921
3.526964	2.137789

2.4 Stage 2B dataset

The Stage 2B dataset is constructed using wafer map data recorded through wafer-level electrical testing after the completion of the manufacturing process. This stage is designed to identify high-risk defective wafers and to facilitate their linkage to scanning electron microscopy (SEM) failure analysis, which is constrained by cost, cycle time, and operational capacity.

At Stage 2B, matrix-form wafer maps are used as the model input. Each wafer map is represented as a two-dimensional grid array with values of 0, 1, and 2, which are normalized to 0, 0.5, and 1, respectively, and subsequently resized before being used as input. The target variable is defined as a pattern label, representing the characteristic failure pattern of the wafer map. Table 5 summarizes the types of pattern labels and their corresponding meanings, and Figure 1 illustrates resized wafer map examples for each label.

Table 5: Types of pattern labels and their meanings

failureType	meaning
Center	Defects are concentrated in the center of the wafer
Donut	The center region is relatively clean, with ring-shaped defects distributed at an intermediate radius
Edge-Loc	Defects are clustered at specific locations near the wafer edge
Edge-Ring	Defects are distributed circumferentially along the wafer edge
Loc	Defects are clustered within a localized region
Near-full	Defects are widely distributed across almost the entire wafer
Random	Defects are scattered irregularly without a distinct structural pattern
Scratch	Defects appear in linear or scratch-like patterns, indicative of physical damage
None	Defect patterns are minimal or barely observable

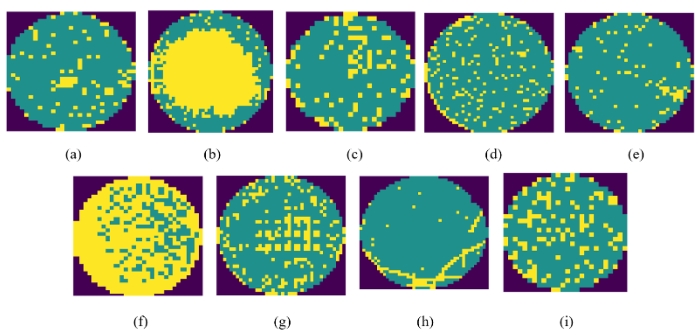

Figure 1: Examples of resized wafer maps for each label. (a) Center, (b) Donut, (c) Edge-Loc, (d) Edge-Ring, (e) Loc, (f) Near-full, (g) Random (h) Scratch (i) None

2.5 Stage 3 dataset

The Stage 3 dataset is constructed to analyze defect morphologies based on SEM images. The dataset consists of a total of 4,591 SEM images, each annotated with one of six defect-type labels. Throughout this study, a unified labeling scheme is adopted across both the manuscript and the codebase, where labels 1–6 correspond to Scratch, Long Scratch, Particle, Pit, Watermark, and No Visible Defect, respectively. Figure 2 presents representative image examples for each label.

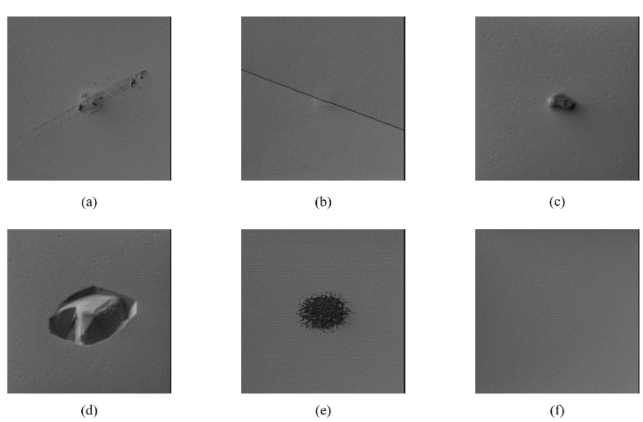

Figure 2: Representative images corresponding to each label. (a) Scratch, (b) Long Scratch, (c) Particle, (d) Pit, (e) Watermark, and (f) No Visible Defect

3 Method

3.1 Stage 0-2A method

This study explicitly reflects the fact that the range of information available in semiconductor manufacturing processes expands progressively over time, and accordingly constructs stage-wise yield prediction models from Stage 0 to Stage 2A. While all stages share the same prediction objective continuous yield estimation each model is restricted to using only the variables that are actually observable at the corresponding decision point. This design prevents discrepancies between model performance and operational timing, thereby ensuring applicability in real manufacturing environments.

The same data splitting strategy and training procedure were applied across all stages. The full dataset consists of 1,250 wafer samples, of which 1,050 were used for model training and 200 were reserved for final evaluation. Within the training data, samples were further split into training and validation sets at an 8:2 ratio, and validation performance was used to monitor the training process. For yield prediction at each stage, an XGBoost-based regression model was employed, which takes input feature vectors and outputs continuous yield values.

The regression performance was evaluated using mean absolute error(MAE) and root mean squared error(RMSE) as evaluation metrics. Table 6 presents the results of validating the yield prediction model's performance on the validation set for each stage.

Table 6: Regression performance of stage-wise yield prediction models on the validation set

stage	MAE	RMSE
stage 0	0.1030	0.1376
stage 1	0.0919	0.1185
stage 2A	0.0957	0.1248

To translate continuous yield predictions into actionable operational decisions, a quantile-based policy was applied to the predicted yield distributions. Across Stage 0, Stage 1, and Stage 2A, wafers in the bottom 20% of predicted yields were defined as high-risk and converted into a binary risk indicator. In addition, at Stage 2A, wafers in the bottom 4% of predicted yields were designated as scrap candidates. This quantile-based policy does not rely on fixed thresholds and therefore maintains consistent operational ratios even under changes in data scale or distribution.

To interpret model decisions and assess operational risk, variable importance analysis, error slice analysis, and sample-level root cause analysis were conducted. Variable importance analysis was used to rank input features according to their contributions to yield prediction, and actual high-risk groups were compared with predicted high-risk groups to decompose true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). In particular, the FN rate defined as the proportion of missed high-risk wafers was used as a key management metric. For individual samples, variables contributing to lower predicted yields were treated as potential risk factors, and the top contributors were identified based on the absolute magnitude of their contributions. However, because contribution values alone do not indicate whether a variable is high or low relative to process norms, samples were further grouped by product type and technology node, normal averages were computed for each group, and each sample was analyzed in terms of whether it deviated above or below the corresponding group mean.

Finally, statistical analyses were performed to examine the assumption that process conditions may differ structurally across lots throughout Stage 0 to Stage 2A. First, a chi-square test was applied to assess the independence between lot identity and risk occurrence, and the results confirmed that the proportion of high-risk wafers varies significantly across lots. Subsequently, within each lot, a binomial test was used to evaluate whether specific process variables were systematically biased in one direction relative to the group average. Because results at the individual lot level may still be influenced by randomness due to limited sample sizes, an additional meta-level analysis was conducted to verify whether the same variables exhibited statistically significant bias repeatedly across multiple lots. The results indicate that certain variables show consistent and significant biases across several lots, suggesting that these variables are associated with lot-level structural characteristics. These findings imply that future process improvement efforts may benefit from differentiating management units on a per-variable basis.

3.2 Stage 2B method

Stage 2B starts from wafer-map-based failureType classification results and connects severity computation and candidate selection policy for prioritizing SEM failure analysis in a single procedure. This stage does not decide candidates from classification results alone; it constructs severity by combining features computed directly from the wafer map with the model confidence, then selects candidates using a Top-% policy over the severity distribution and attaches process-step and root-cause-hypothesis fields through mapping so that the output links directly to operational decision-making.

First, we restricted the training set to samples with valid wafer maps and labels to ensure label quality. During label preprocessing, we removed entries recorded as blanks or meaningless values, and we kept a relaxed minimum-frequency filter to prevent rare patterns from being structurally eliminated in the later candidate-policy stage. Because scratch is a key routing rule that branches to physical damage in the candidate policy, we ensured that this label remains observable during training. We performed a stratified split that preserves label distributions so that training, model selection, and final evaluation are separated, and we retained the original sample indices to reliably re-reference the same samples in later stages.

The model input was defined as a single wafer map. Because failureType is defined by spatial defect shapes on the wafer map, including identifier-like auxiliary columns such as lotName or waferIndex could cause the model to learn spurious cues tied to the data-collection environment or lot composition, which may degrade generalization. Moreover, combining map-like inputs with tabular inputs would require additional missing-value handling, scaling, and fusion-structure design, increasing pipeline complexity; therefore, to keep a consistent selection chain in Stage 2B, we adopted a single-input design. In input preprocessing, we normalized the value range to stabilize training and applied nearest-neighbor resizing so that the discrete state semantics of the wafer map are not distorted by interpolation.

We trained a lightweight CNN to predict nine pattern classes. To mitigate class imbalance, we counted per-class samples in the training data, applied inverse-frequency weights to the loss, and normalized the weights to prevent the overall scale from becoming excessively large and destabilizing training. During training, we selected the best checkpoint based on validation loss to reduce overfitting risk relative to saving the final epoch and to ensure a consistent reference model in downstream stages. After training, we generated per-test-sample predictions and recorded confidence and uncertainty indicators along with pred_label so that the classifier's reliability characteristics can be used in the downstream severity stage. We additionally produced evaluation outputs for checking classification performance.

In the severity computation stage, we combined wafer-map-derived features with confidence to convert classification results into SEM priority. The features were designed as interpretable, map-based signals such as defect amount, positional bias, and 8-neighborhood-based clustering statistics so that severity is not solely dependent on the model output. We formed a raw score via a weighted linear combination of these signals and then converted it to a 0–100 severity scale using a saturating nonlinear mapping. This design mitigates cases where extreme values dominate the distribution and make Top-% thresholds unstable, and it provides scoring that is well suited to percentile-based candidate policies.

The candidate policy is primarily based on Top-% selection over the severity distribution. Scratch is separated and routed to physical damage rather than SEM, maintaining policy consistency with SEM's purpose of identifying process-related root causes. Within the non-scratch pool, samples marked as random-like, blob-like, or cluster-like are included with priority; when budget or throughput constraints exist, the prioritized set is placed first and the remaining slots are filled in descending order of severity to maintain a stable, actionable volume. Table 7 below shows two example rows illustrating the severity computation; random-like, blob-like, and cluster-like flags are generated as markers for prioritized inclusion under limited SEM capacity, but they are omitted from the table for brevity.

Table 7: Examples of numerical data

orig_idx	true_label	pred_label	conf	defect_ratio	edge_bias	cluster_score	severity
679463	Center	Donut	0.531739	0.315241	0	1	94.80402
742536	none	Edge-Loc	0.951462	0.092726	0	1	86.44564

Finally, mapping attaches fields that link the predicted pattern to process steps and root-cause hypotheses so that the result supports downstream failure-analysis decisions beyond being a simple candidate list. Items requiring human review are left blank in the output schema to preserve the boundary between automated policy and operational approval.

3.3 Stage 3 method

Stage3 is not limited to predicting defect morphology from scanning electron microscope images. Instead, it is designed as a decision-support stage that organizes inspection results into case-level operational records, enabling defect review systems to prioritize limited review resources effectively. The objective of Stage3 is not to maximize average classification performance, but to distinguish cases that can be confirmed immediately from those for which confirmation is risky, and to proactively identify cases that require additional review or reacquisition.

To achieve this goal, Stage3 combines morphological class predictions with probability-calibrated confidence signals and entropy-based uncertainty signals, and maps these signals into a predefined triage policy that produces outputs directly usable in defect review workflows. Final confirmation decisions and process actions are explicitly separated from model outputs and are recorded as engineer judgments, ensuring that inference signals and human decision outcomes are not conflated.

In this study, a ResNet18-based classifier was trained on a public SEM defect dataset consisting of 4,591 images annotated with six morphological defect classes. The dataset was split into training 70 percent, validation 15 percent, and test 15 percent sets while preserving class distributions, resulting in 694 images in the test set. Because probability outputs can be interpreted as confirmation strength in defect review environments, temperature scaling was applied to mitigate model overconfidence. Calibration quality evaluated on the validation set yielded an expected calibration error of 0.0971 and a negative log-likelihood of 0.1537.

Uncertainty signals were defined using entropy computed from the calibrated probability distribution. Across the dataset, entropy had a mean of 0.4206 and a median of 0.2776, with the 90th percentile observed at 1.0532. This upper 10 percent region was selected as the uncertainty threshold. By fixing the proportion of cases requiring additional review, this design reduces sensitivity to data distribution shifts or imaging condition changes while directly reflecting the operational constraints of defect review environments with limited review capacity.

The Stage3 triage policy is defined exclusively using signals observable at inference time, including calibrated confidence, entropy, and a brightness-based image quality tag named brightness_tag. Based on these signals, each case is assigned to one of four triage categories: A_strong_evidence, B_overconfidence_warn, C_ambiguous_boundary, and D_acquisition_risk. Any interpretation that uses correctness information is restricted to post hoc analysis on the test set and does not influence triage assignment during operation.

Each triage category is associated with a fixed operational recommendation. A_strong_evidence cases are recommended for immediate confirmation, assigned high review priority, and permitted for automatic confirmation. B_overconfidence_warn cases prohibit immediate confirmation by default and are placed in priority review queues to prevent misconfirmation driven by overconfidence. C_ambiguous_boundary cases trigger recommendations for additional review, such as secondary inspection or expanded region-of-interest analysis. D_acquisition_risk cases prioritize reacquisition or verification of imaging conditions to address potential observation failures. These recommendations are produced as model outputs, while actual confirmation, deferral, and reacquisition decisions are recorded separately as engineer judgments.

The final outputs of Stage3 are organized as case-level operational records. Each record contains the predicted defect morphology, confidence and uncertainty signals, auxiliary image quality indicators in brightness_tag, the assigned triage category, and the corresponding operational recommendations and review priorities. In addition, a selection_tag is used to indicate how cases are selected for review, where ai denotes cases prioritized by the triage and priority policy, and random denotes cases selected as a baseline for comparison. Ground-truth labels are retained solely for post hoc analysis and are not used in operational triage decisions. Representative examples of these case-level operational records are summarized in Table 8.

To assess the operational validity of the proposed Stage3 design, triage-based case selection was compared against a random selection baseline. The results show that Stage3 more clearly separates immediately confirmable cases from those requiring further inspection or reacquisition than random selection, demonstrating that the combination of calibrated confidence and entropy-based uncertainty provides meaningful operational decision inputs for defect review environments.

Table 8: Example Stage3 operational case records

selection_tag	pred_name_en	triage	calibrated_conf	entropy	brightness_tag	process_improvement_action	dr_priority
ai	Particle / Foreign material	A_strong_evidence	0.9309	0.3605	dark	Confirm	High
random	Pit / Crater	C_ambiguous_boundary	0.3919	1.3326	dark	Re-check + Optional re-acquire	Medium

3.4 Decision Agent Method

Semiconductor inspection and metrology decisions are fundamentally constrained by limited metrology/SEM capacity and strict budget caps. Under these operational constraints, a practical system must achieve two objectives simultaneously: it must (i) operate as a realistic staged decision workflow that engineers can actually use, and (ii) support scientifically valid performance claims. These objectives impose different success criteria. Operational demonstrations prioritize feasibility—human control, auditability, staged routing, and budget-aware execution—whereas scientific claims require strict evidence boundaries, including same-source ground truth, reproducibility, and explicit prevention of leakage or source mixing. To avoid overstating results under heterogeneous data availability, we separate the system into two layers that act as an explicit governance mechanism. Track A provides operational integration by implementing a staged Human-in-the-Loop workflow (Stage 0→1→2A→2B→3) with explicit decision points, budget tracking, and audit trails, while Track B provides the validated core by restricting quantitative claims to Step1 (Stage0–2A), where same-source yield_true ground truth is available, and enforcing reproducible reporting bound to a single run. Figure 3 illustrates the proposed two-layer orchestration that separates operational workflow integration (Track A) from claim-safe scientific validation (Track B).

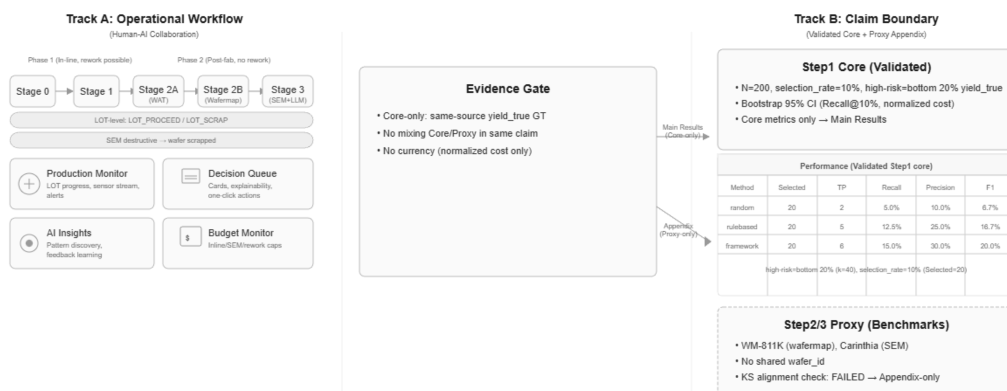

Figure 3: Two-layer governance architecture(workflow vs claims)

Track A implements a staged Human-in-the-Loop pipeline (Stage 0→1→2A→2B→3) with UI-level modules (monitoring, decision queue, insights, and budget control) to reflect realistic fab execution. Track B enforces an evidence gate that restricts main-paper quantitative claims to the Step1 validated core (Stage0–2A) where same-source yield_true ground truth exists, with run-level binding (run_20260131_004542 + sha256 manifest), no currency reporting (normalized costs only), and explicit prevention of Core/Proxy mixing. Downstream Step2/3 modules are evaluated on external benchmarks (WM-811K, Carinthia) and remain Appendix-only when proxy alignment fails.

In this work, “integration” is therefore not defined as end-to-end model fusion across stages. Because downstream modules (Step2/3) are evaluated on external benchmark datasets that do not share wafer identifiers with Step1, any end-to-end causal or utility claim would be invalid. Instead, integration is defined as claim-safe coupling under evidence gates: Track A chains stages operationally (workflow integration), while Track B determines which metrics are allowed to appear in main results (claim integration). A direct consequence of this design is that a failed proxy alignment result is treated as a governance outcome rather than a technical failure—it prevents benchmark-only modules from contaminating core conclusions.

Agent orchestration follows the same operational logic: the fab decision problem is not "maximize classifier accuracy," but "select a limited fraction of wafers for follow-up." We fix the operational selection policy to selection_rate = 10% and evaluate whether true high-risk wafers—defined as the bottom 20% by yield_true—are enriched in this top-k set. Accordingly, the agent optimizes decision policy parameters rather than retraining models: thresholds are tuned on a validation split only (to prevent test leakage), and budget-aware scheduling is performed using normalized unitless costs with a follow-up cost ratio sweep to avoid dependence on absolute currency assumptions. Decisions and their supporting metadata are logged to ensure that selection outcomes can be reproduced and audited.

Within the validated Step1 core (same-source test set, N=200; selection_rate=10%; high-risk defined as bottom 20% by yield_true), this evidence-gated policy optimization yields a measurable enrichment signal over operational baselines. Under identical selection constraints, the framework increases high-risk recall from 0.05 (random) to 0.15, corresponding to +4 additional high-risk wafers captured (TP: 2→6) and −4 fewer misses (FN: 38→34) relative to random selection. We report this as a preliminary improvement signal because the bootstrap 95% CI for ΔRecall includes 0.0, and therefore we do not claim statistical significance. Nevertheless, the observed gains are operationally meaningful in a capacity-limited regime where each additional "catch" consumes scarce metrology budget. Cost comparisons are expressed in normalized units only; under a follow-up cost ratio sweep, the framework exhibits recall-dominant regions (improved recall at matched normalized cost), while absolute cost savings are not asserted.

Given the small fixed test set (N=200) and the risk of overstating significance, we adopt a conservative validation strategy that prioritizes claim safety. Primary evidence is reported using bootstrap 95% confidence intervals for (i) Recall@10% and (ii) normalized cost reduction (%). If a confidence interval includes 0, we explicitly avoid claims of statistically significant improvement and report results as preliminary signals. Additional tests (chi-square, McNemar, yield-distribution comparisons) are treated as supplementary diagnostics only and do not alter primary conclusions. Finally, downstream Step2/3 modules (pattern/SEM) are reported as proxy benchmarks demonstrating functional feasibility but remain appendix-only because they lack same-wafer linkage with Step1. We further test a minimal plausibility condition (distribution alignment) to assess whether proxy outputs can be mapped into the Step1 context; in the current run, this alignment check fails, implying that Step2/3 results remain isolated, no end-to-end claim is made, and the main conclusions remain restricted to Step1. The results remain subject to limitations such as the small fixed test set and potential lot-level leakage (group split not enforced), which motivates holdout-lot evaluations as future work. Overall, the proposed integration is a two-layer, evidence-gated orchestration in which Track A demonstrates a realistic staged decision workflow and Track B constrains quantitative claims to a validated core supported by same-source ground truth and reproducible artifacts, with downstream benchmarks prevented from contaminating core claims when alignment checks fail.

4 Conclusion

This study proposed a staged decision-support framework that aligns machine learning models with the progressive expansion of information in semiconductor manufacturing. By decomposing the process into sequential stages from Stage 0 to Stage 3, we prevented information leakage and ensured operational consistency. The results demonstrate that integrating early-stage yield prediction with downstream defect analysis enhances decision-making quality under fixed capacity constraints. Furthermore, the two-layer governance design provides a practical and scientifically responsible pathway for deploying AI by separating operational integration from core validation.

However, certain limitations remain, such as the reliance on fixed test sets and the lack of same-wafer linkage in downstream proxy data. Future work will focus on validating the framework with fully linked multi-stage data, enforcing lot-level holdout evaluation to eliminate residual leakage, and extending the decision agent to optimize policies under dynamic budget and throughput conditions.

AI Co-Scientist Challenge Korea Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [N/A] .
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[N/A] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "AI Co-Scientist Challenge Korea paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly define the paper's main contribution as the design of a staged decision-support framework spanning Stage 0 to Stage 3. Quantitative performance claims are explicitly restricted to the validated core stages (Stage 0–2A), where same-source ground truth is available, while downstream stages (Stage 2B–3) are presented solely as demonstrations of operational feasibility. As a result, the stated claims accurately reflect both the experimental evidence and the intended scope of the paper without overstating generalization or end-to-end effects.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals

are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses several limitations, including the use of a fixed and relatively small test set, the absence of lot-level holdout splits that may allow residual leakage, and the lack of same-wafer linkage for downstream stages (Stage 2B–3). These constraints are directly reflected in the decision to restrict quantitative claims to Stage 0–2A and to treat downstream results as proxy demonstrations. The paper further outlines these issues as directions for future work, ensuring transparency regarding the scope and robustness of the proposed framework.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not present formal theoretical results such as theorems, lemmas, or mathematical guarantees. Instead, it focuses on the design and empirical evaluation of a staged decision-support framework under realistic operational constraints. As a result, assumptions and formal proofs are not applicable to the scope of this work.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses all information necessary to reproduce the experimental results that support its main claims. For the validated core stages (Stage 0–2A), the dataset composition, data splits, fixed test set, model choices, evaluation metrics, and quantile-based decision policies are explicitly described. Downstream stages are clearly separated as proxy benchmarks with fully specified datasets, inputs, and evaluation procedures, ensuring that the scope of reproducibility is transparent and aligned with the claims of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not provide open access to the full data and code. The core datasets are derived from industrial manufacturing processes and cannot be publicly released, and the integrated training and evaluation scripts are not shared. While some downstream experiments rely on publicly available benchmark datasets, executable code and complete reproduction instructions are not included in the paper or supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides sufficient experimental details to understand the reported results. For each stage, it specifies the dataset composition and splits, model types, training and validation procedures, evaluation metrics, and decision policies. For the core stages (Stage 0–2A), the fixed test set and evaluation criteria are clearly described in the main text, with additional stage-specific details provided in the methodology sections and appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports statistical uncertainty for the experiments supporting its main claims using bootstrap-based 95% confidence intervals. When the confidence interval includes zero, the paper explicitly refrains from claiming statistical significance and interprets the results as preliminary improvement signals. This conservative reporting clearly distinguishes statistical significance from observed effect trends.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not explicitly report detailed information about the computational resources used for the experiments, such as the type of compute workers (CPU or GPU), memory specifications, or execution time. As a result, the exact computational cost required to reproduce the experiments is not fully specified.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://nips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research was conducted in accordance with the NeurIPS Code of Ethics. The study respects legal and ethical constraints associated with industrial data, preserves anonymity where required, and avoids overstated or misleading claims by clearly delineating the scope and limitations of the results.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive and negative societal impacts of the proposed work. On the positive side, the framework can improve resource efficiency and decision quality in semiconductor manufacturing by supporting more informed inspection and metrology allocation. On the negative side, the authors acknowledge that erroneous model outputs or misinterpretation could lead to suboptimal operational decisions, and mitigate these risks by adopting a Human-in-the-Loop design and by conservatively limiting the scope of automated claims and actions.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release or distribute data or models that pose a high risk of misuse or dual use, such as pretrained generative models or scraped datasets. The contribution focuses on the design and evaluation of a decision-support framework rather than the public release of high-risk assets, making safeguards not applicable in this context.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper properly cites the original sources of the publicly available datasets used in the experiments. However, the specific licenses and detailed terms of use for these assets are not explicitly stated in the paper or supplemental material. As a result, license information is not fully documented within the manuscript.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release or distribute new assets such as datasets, code, or models. Accordingly, documentation of newly introduced assets is not applicable to this work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted using existing process data and publicly available benchmark datasets without direct human participation, making this question not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects. As no study participants are involved, there are no participant risks to disclose and no Institutional Review Board (IRB) approval is required for this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review