Comprehensive Framework for Identifying and Visualizing Key Yield Factors in Semiconductor Manufacturing

Gyeunggeun Doh, Jeongwoo Seo, Sanghoun Oh, Minseok Song, *Member, IEEE*

***Abstract*—Semiconductor manufacturing involves complex multistage processes in which product yield is influenced by intricate interactions among machines, materials, recipes, process durations, queue times, and wafer characteristics, such as warpage. Traditional yield analysis methods typically examine these factors in isolation, overlooking their combined effects. To address this limitation, we propose a novel framework based on Transition System (TS) modeling that jointly analyzes high-dimensional process attributes to improve yield understanding and prediction. The framework achieves three goals: (1) quantifying the impact of individual process attributes on yield, (2) identifying interacting attribute combinations that produce best-of-best (BOB) and worst-of-worst (WOW) wafer paths, and (3) visualizing these patterns through an interpretable TS model. We first encode each wafer's event log as a sequence of discrete state attributes - machine, material, recipe, queue time, duration, and warpage - and then construct a TS model that connects adjacent states. Critical attributes are selected through random forest feature importance, followed by association rule mining to extract yield-relevant state combinations. The final model enables classification of wafers into high- or low-yield categories and provides visual insight into the process behavior. Experiments using real production data from a Korean semiconductor facility demonstrate the effectiveness of the framework in uncovering key yield drivers and supporting data-driven process optimization.**

***Index Terms*—Process mining, wafer path, semiconductor manufacturing, transition systems, yield optimization, visualization, multistage manufacturing process.**

I. Introduction

SEMICONDUCTOR manufacturing process is one of the most representative multistage manufacturing processes (MMP) [1]. The system is designed with sequential process stages to produce the final product from the raw material, and alternative machines are aligned at each stage to improve productivity. The product passes through one of the multiple machines in each process stage. Fig. 1 illustrates the concept of semiconductor MMP, where N process stages and K alternative machines with M separate chambers are aligned in each process stage. In the figure, $equipment_{n,k}$ represents the kth machine in the nth step of the process. In addition, $chamber_{n,k,m}$ represents the mth chamber within $equipment_{n,k}$. The connecting line shows the production path, which is the sequence of machines assigned in the manufacturing process.

This work was supported by the National Research Foundation of Korea (NRF) grant (RS-2024-00357330) and by SK hynix AICC (P24.03_TSV Process Optimal Path Discovery and Inefficiency Derivation based on Manufacturing Data)(Corresponding author: Minseok Song.)

Gyeunggeun Doh and Minseok Song are with the Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang 37673, South Korea (e-mail: mssong@postech.ac.kr)

Jeongwoo Seo and Sanghoun Oh are with SK hynix Inc., Icheon 17336, South Korea

After the final processing step, each wafer undergoes an end-of-line (EOL) inspection to assess its final quality. Wafers that fail are marked defective, while those that pass are marked functional. Defective wafers can lead to scrap or rework costs, additional labor, and possible late delivery penalties, all of which can significantly cut margins [2]. Therefore, quickly identifying and addressing defect-causing factors is crucial to remain competitive [3].

Empirical studies have identified two main drivers of the EOL yield. The first is the resource path, which refers to the sequence of machines (e.g., stacking equipment A $\rightarrow$ bonding equipment B $\rightarrow$ lithium equipment C) that a product follows through the manufacturing process. For equipment in the same step, there may be slight differences in performance [4]. The second is the additional attributes of the process, including both quantitative and categorical measurements at each step, such as wafer warpage, process duration, queue time, recipe codes, material lot numbers, and temperature setpoints [5]. These resources and additional process attributes embody the physical and operational conditions experienced by the wafer and often interact non-linearly to influence defect formation. For example, in Fig. 1, $chamber_{3,k,1}$ generally produces high-quality wafers. However, if it has passed through $chamber_{2,k,M}$ in step 2 with an unusually long duration, the final yield still suffers, making that combination undesirable. Together, resource paths and process attributes form a high-dimensional state space, making the joint effects on yield challenging to untangle.

In addition to the challenges posed by high-dimensional complexity, there are other significant issues in yield optimization for semiconductor MMPs. First, the large number of sequential processes and multiple machine options, each with various variants of recipes, make exhaustive optimization of wafer paths computationally infeasible. Second, resource paths are highly dependent on upstream scheduling and equipment availability, leading to multicollinearity that undermines the reliability of traditional statistical and machine learning methods [6]. Third, defects can only be identified after the entire manufacturing process, leaving no opportunity for timely intervention without accurate predictive models. Finally, even the most precise yield analysis methods often lack a clear visualization of their results, making it difficult for process engineers to interpret and use them effectively [7].

To address these challenges, we propose a Transition System (TS) based framework for analyzing and visualizing multistage semiconductor manufacturing, specifically aimed at identifying process factors that influence yield. The methodology includes six phases: data preparation, TS model construction, critical attribute selection, critical attribute combination deriva-

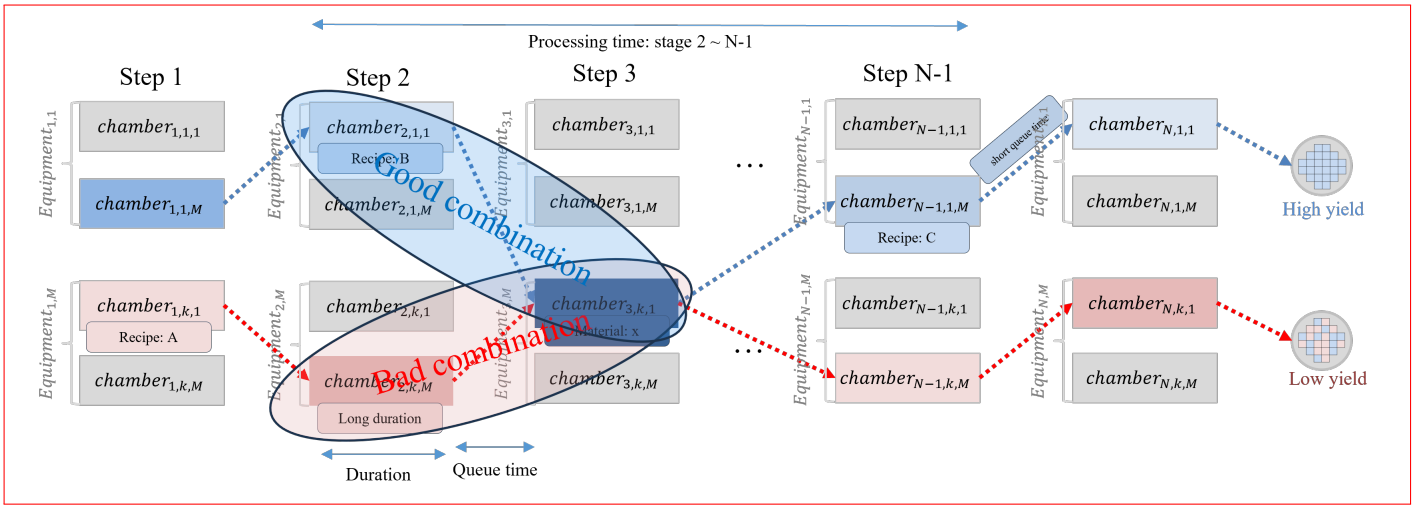

Fig. 1. A conceptual example of a complicated semiconductor manufacturing process.

tion, TS model simplification, and evaluation. Essentially, we develop a TS model from manufacturing data, filter and visualize the most important attributes and their combinations, then simplify the integrated model by focusing on key resources and attributes identified through statistical analysis. Finally, we validate our approach through an empirical study with a leading Korean semiconductor manufacturer.

II. Background

This section examines yield optimization in MMPs through various data-driven and model-based approaches. Past research is organized by the types of process features studied, the methods used to manage correlated or interacting factors, and the strategies for visualizing and interpreting models. We align our proposed study with the existing literature and identify research gaps that shape our framework.

A. Process Feature Analysis in Yield Management

Early yield analysis efforts in multistage MMPs focused mainly on resource paths, the specific sequence of machines through which a product passes, and their influence on final quality. Resource-oriented transition systems, labeled with failure rates, identified optimal and suboptimal resource paths [7]. Data mining combined with design-of-experiments techniques pinpointed key recipe settings linked to yield loss [5], and "golden paths" were developed by integrating machine sequence patterns with quantitative attributes [6]. A Gaussian mixture model improved the accuracy of yield prediction using metrology and step-duration data [8]. More recent studies have incorporated sensor-derived signals, such as temperature and pressure, into yield models, using platforms that combine equipment logs with sensor data to improve defect detection [9]. Despite these efforts, no comprehensive analysis has examined both resource paths and the wide range of additional process attributes throughout the manufacturing process.

B. Handling Correlated or Interacting Factors

A common challenge in MMP yield analysis is multicollinearity, as many process attributes (e.g., resource, recipe, duration, queue time, wafer warpage) are highly interconnected. Stream-of-variation (SOV) models address correlated machining parameters through statistical decomposition [10]. State-space models monitor how variation propagates in multistage systems, formally capturing interactions [11]. Linear-variation methods, such as PCA and PLS, help stabilize yield predictions by removing correlations among variables [12]. Pattern mining techniques explicitly examine interactions by predicting failures based on combinations of process steps [13], while association rule mining identifies sets of quality-enhancing factors across production stages [14]. Penalized regression and ensemble methods, including random forests and boosting, effectively manage multicollinearity and capture nonlinear interactions [15]. However, these approaches do not fully account for multicollinearity between resource paths and the wide range of process attributes within a single analytical framework.

C. Visualization and Interpretation of the Process

Even the most precise yield models may have limited influence on process engineers if they lack clear visualization and interpretability. The clarity and ease of interpretation are critical, as the industry's shift toward failure prevention strategies in high-reliability contexts necessitates that analytical findings be fully interpretable and actionable by process engineers for root-cause identification [16]. Process mining directly addresses this problem by automatically discovering and visualizing process flows, enriched with performance metrics, thereby enhancing the interpretability and practical usefulness of yield models [17]. TS models, commonly used in process mining to represent process dynamics, help to visualize manufacturing workflows through "yield maps" that overlay failure rates and wafer counts, offering intuitive insights [7]. Microarray-like visualizations allow quick fault detection by color-coding high-dimensional parameter effects [18]. Interactive dashboards display real-time sensor data, enabling engineers to drill down from overall yield trends to detailed lot histories [9]. Explainable AI visualizations,

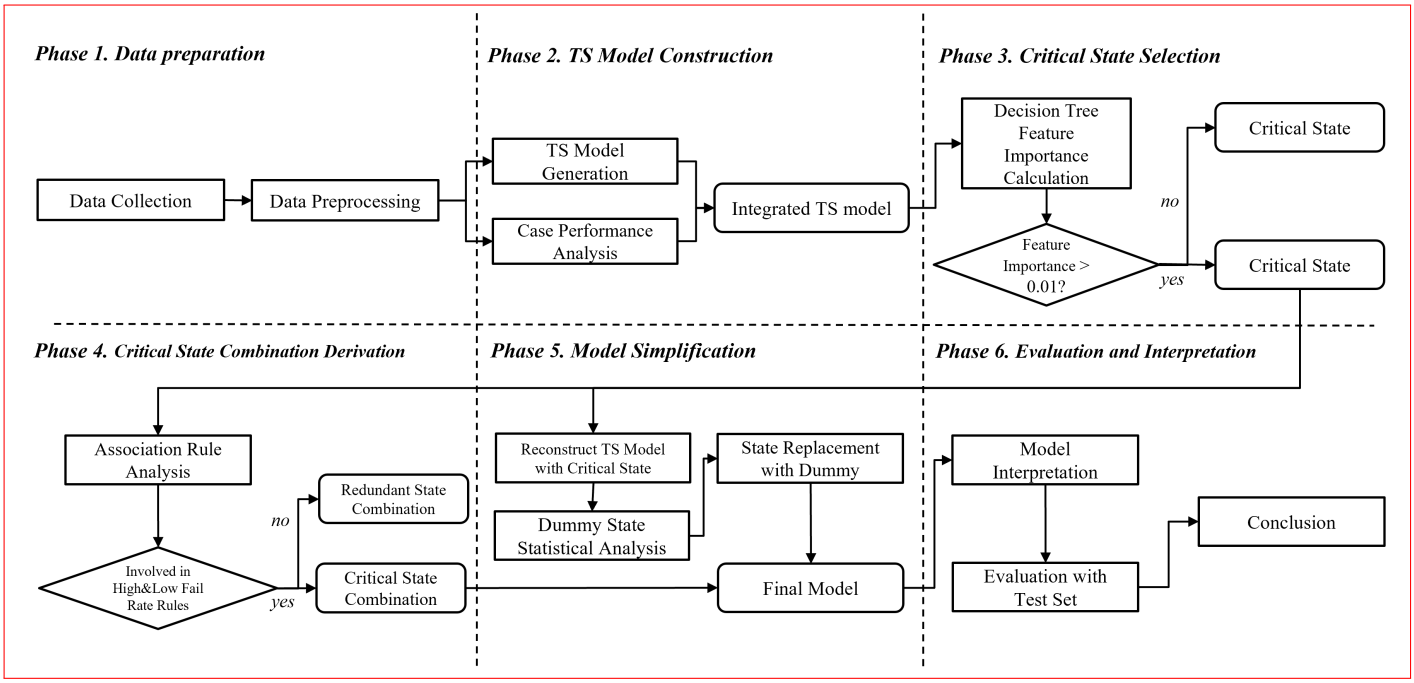

Fig. 2. Overview of the proposed research framework.

especially SHAP value plots, clarify how features impact yield, turning black-box predictions into actionable insights [19]. Overall, process-focused visualizations such as graphs, heatmaps, and dashboards effectively link complex analytics with practical decision-making, supporting ongoing yield improvement. However, these methods have yet to produce integrated, intuitive visualizations that depict both the full process flow and the quality landscape of all key process factors.

D. Research Gaps and Proposed Framework

Building on these advancements and their limitations, our proposed framework directly addresses remaining research gaps. First, we expand feature analysis beyond resource paths to include various process attributes, such as wafer warpage, queue times, and recipes used. Second, we introduce a multicollinearity-aware analytical approach that simultaneously evaluates resource sequences and additional attributes. Third, we utilize a TS model to generate clear, interactive visualizations of both key resource paths and attribute interactions. This unified framework overcomes the limitations of previous studies by combining broader feature sets, correlation-sensitive analysis, and process-focused visualization into a cohesive yield optimization tool.

III. A Framework for Analyzing and Visualizing a Semiconductor Manufacturing Process using a TS Model for Yield Enhancement

This section introduces a systematic framework for analyzing and visualizing semiconductor manufacturing processes. Fig. 2 shows the proposed research framework. Our approach comprises six phases: data preparation, TS model construction, selection of the critical state dimension, derivation of critical state combinations, model simplification, and evaluation.

In phase I, raw event logs are preprocessed to build wafer paths. In phase II, the wafer path sequences are mapped onto a TS model, where each state represents a process attribute, annotated with failure rates and wafer counts to visualize defect patterns. Phase III uses random forest feature importance to identify state dimensions that significantly impact yield, keeping only the most influential ones to simplify the model. In phase IV, association rule mining uncovers key combinations of state values associated with high or low yield outcomes, which are then visualized in the TS model. Phase V further simplifies the model by removing infrequent or redundant states based on variance and coverage criteria. Finally, in phase VI, the method is validated by classifying test wafers using the critical-state combinations, demonstrating that the simplified TS model provides interpretable, actionable insights for yield improvement.

A. Phase I. Data Preparation

During the data preparation stage, the event log from the manufacturing system is extracted. An event log contains a sequence of events associated with process instances (wafer cases). Each event includes a case ID, activity, start and end timestamps, event attributes (such as resource, recipe, material), product attributes (such as wafer warpage), processing time attributes (including aggregated processing times for specific manufacturing steps), and the EOL yield. Table I provides an example of a preprocessed event log, which includes wafer IDs, operations, timestamps, event attributes (equipment, recipes), event time attributes (queue time, duration), product attributes (warpage), processing time attributes,

TABLE I
AN EXAMPLE OF EVENT LOG

Case	Activity (step)	Time stamp		Event attribute		Event Time attribute		Case time attribute	Product attribute	EOL yield
Wafer id	Oper id	Start time	End time	Equipment	Recipe	Queue time	Duration	Processing time 1~2	Warpage	Fail rate
1	1	00:00	00:10	1_2	a	-	Long	Short	Normal	1%
	2	00:15	00:23	2_3	c	Short	Short			
	3	00:30	00:38	3_1	e	Short	Short			
2	1	00:30	00:35	1_1	b	-	Short	Long	High	8%
	2	00:40	01:15	2_3	c	Short	Long			
	3	01:50	02:57	3_2	f	Long	Long			
3	1	01:00	01:05	1_2	a	-	Short	Long	Normal	10%
	2	01:50	02:34	2_1	d	Long	Long			
	3	02:50	03:16	3_2	g	Short	Long			
4	1	02:00	02:15	1_1	C	-	Long	Short	High	2%
	2	02:30	02:54	2_2	F	Short	Long			
	3	03:40	03:50	3_3	I	Long	Short			

TABLE II
AN EXAMPLE OF WAFER PATH EXTRACTED FROM THE EVENT LOG

Path	Product attribute	Step 1				Step 2				Step 3				Case time attribute	EOL yield
Wafer id	Warpage	Queue time	Equipment	Recipe	Duration	Queue time	Equipment	Recipe	Duration	Queue time	Equipment	Recipe	Duration	Processing time 1~2	Fail rate
1	Normal	-	1_2	a	Long	Short	2_3	d	Short	Short	3_1	f	Short	Short	1%
2	High	-	1_1	b	Short	Short	2_3	e	Long	Long	3_2	f	Long	Long	8%
3	Normal	-	1_1	a	Short	Long	2_1	d	Long	Short	3_2	g	Long	Long	10%
4	High	-	1_2	c	Long	Short	2_2	e	Long	Long	3_3	h	Short	Short	2%

and EOL failure rate. Definition 1 formally explains the event log used in this research.

Definition 1 (Event log, Event, case, Activity, Timestamp, Event attribute, Process attribute, Yield): Let L be a manufacturing event log comprising a finite set of events. L can be formalized as $L = \{e|e = 1, \ldots, |L|\}$ and each event $e \in L$ is formally represented as the tuple: $e \mapsto (c(e), act(e), ts(e), te(e), ea(e), eta(e), pa(e), pta(e), y(e))$, where $c(e)$ is the wafer case identifier, $act(e)$ is the executed activity, $ts(e)$ and $te(e)$ are start and end timestamps, $ea(e)$ represents event attributes, $eta(e)$ contains event time attributes (e.g., queue time, duration), $pa(e)$ describes product attributes, $pta(e)$ indicates aggregated process-time attributes, and $y(e)$ is the yield outcome.

Collected event logs are preprocessed to improve data quality and ensure compatibility with TS modeling. Events with missing, inconsistent fields, or duplicate entries are removed. Continuous attributes are discretized using user-defined thresholds to create categorical state dimensions, which are essential for subsequent analysis.

From the discretized logs, wafer paths are created by grouping events according to wafer ID and arranging them in chronological order. Table II shows wafer paths derived from the event log in Table I, illustrating the detailed sequences needed for further analysis. Definition 2 provides a formal explanation of a wafer path.

Definition 2 (Wafer path): Let C represent all wafer cases in the event log L. For each case $c \in C$ with n_c events, let $\{e_{c,1}, e_{c,2}, \ldots, e_{c,n_c}\} \subseteq L$ denote the events with case identifier c, ordered by non-decreasing start timestamp $ts(e)$. A wafer path is defined as $wp(c) = (pa(c), ea(e_{c,1}), ea(e_{c,2}), \ldots, ea(e_{c,n_c}), pta(c))$, and for each wafer path, the sequence of attributes of case c is assigned to the wafer path. Note that $pa(c)$ is the set of product attribute value of wafer c, $ea(e_{c,k})$ is the set of event attribute value of the k_{th} event in the trace of case c, $pta(c)$ is the set of processing time attributes that contains the information of aggregated processing time of the manufacturing step. Thus, $wp(c)$ captures the complete ordered sequence of product attributes, event attributes, and processing time attributes for each wafer.

Finally, the wafer path data is divided into training and

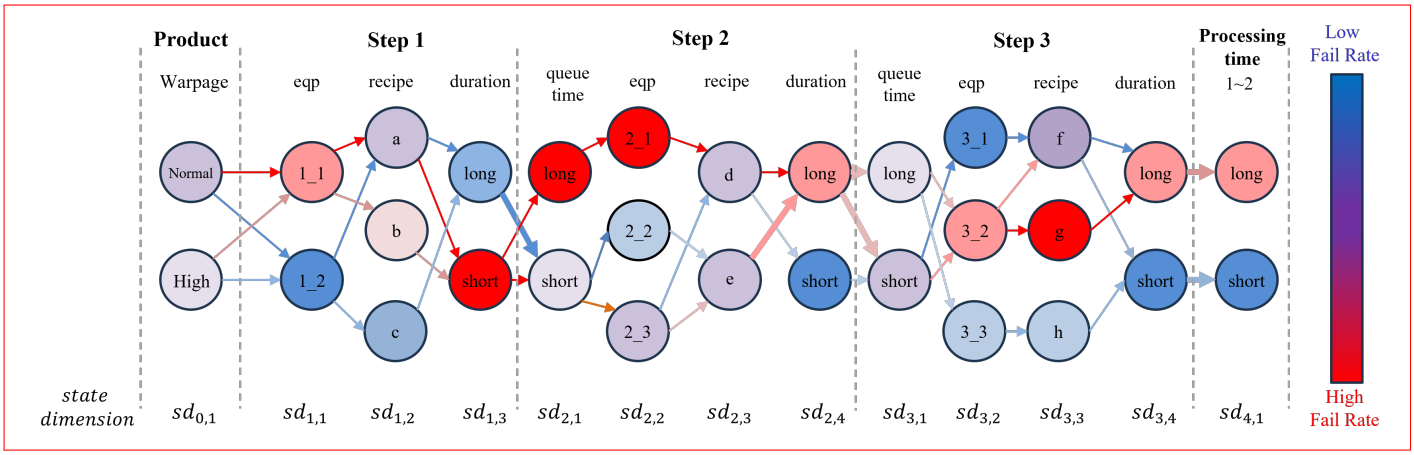

Fig. 3. Running example of the TS model integrated with failure rate annotations.

test sets. The training set helps identify key process attributes and their effects on yield, while the test set evaluates these combinations using yield prediction.

B. Phase II. TS Model Construction

This phase of the framework concentrates on developing a TS model from the cleaned wafer paths produced in Phase I. The objectives are to formalize the multistage process and its attributes as a graph of distinct states and to annotate the TS model with combined quality and volume metrics for analysis and visualization.

A transition system TS is defined as $TS = (S, T)$, where S is the set of states and T is the transition relation, which is a subset of $S \times S$, where $\times$ denotes the Cartesian product To build a TS model, the state and transition need to be extracted from the wafer path. To begin, each wafer path wp_c is decomposed into its step attributes $(v_{c,0}, v_{c,1}, \ldots, v_{c,N}, v_{c,N+1})$, where $v_{c,0} = pa(c)$ captures the product attributes, $v_{c,k} = ea(e_{c,k})$ for $1 \leq k \leq N$ captures the event-level attributes at step k, and $v_{c,N+1} = pta(c)$ captures the aggregated processing time attributes. Here, each $v_{c,k}$ is an attribute-value tuple at step k for wafer case c, consisting of J_k discretized attribute values $(v_{c,k,1}, \ldots, v_{c,k,J_k})$, where each $v_{c,k,j}$ is a single discretized value.

Then, state extraction is performed by gathering all unique attribute values across all wafers and positions along the paths. For each step k and attribute component j, the state dimension $sd_{k,j}$ is constructed as the set of all unique values observed for $v_{c,k,j}$ across all wafer cases c. Each unique value in $sd_{k,j}$ becomes a distinct state in the TS model, and with this state extraction, every product attribute, event attribute at each step, and processing-time attribute is represented as a unique state in the model. After that, all the state dimensions are aggregated into global state set $S = \cup_{k=0}^{N+1} \cup_{j=1}^{J_k} sd_{k,j}$.

After that, transitions between states are extracted to show the successive relationship between process attributes. There are two types of transitions, which are directed arcs. Intra-step transition T_{intra} captures the progression between attribute states within the same step. Inter-step transition T_{inter} capture linkage of states from the final state dimension of one step to the initial dimension of the next. The final transition T is the union of intra/inter step transitions $T = T_{intra} \cup T_{inter}$, which yields a directed transition relation that precisely mirrors each wafer's discrete evolution through the process.

Then the transition system model is constructed as a directed graph $TS = (S, T)$. Thus, the TS node represents all discrete product, event, and processing-time attributes, and those arcs capture every elementary intra-step and inter-step progression experienced by the wafers.

Fig. 3 is an example of the TS model built from the wafer path. A node indicates a state in the TS model, and an arc shows a transition. As shown in the figure, each process attribute of the wafer path becomes a state dimension in the TS model. Additionally, each process attribute value maps to a state in the TS model. For instance, since the product attribute warpage has two values, normal and high, these correspond to two states in the TS model. Similarly, for the equipment in step 1, there are two different pieces of equipment, so two corresponding states are created.

Once the raw transition system is constructed, states and transitions are annotated with two key metrics: the average EOL failure rate for wafers in a state or crossing a transition, and the total wafer count. For visualization, states and transitions are color-coded by failure rate on a blue-to-red gradient, and transition edges are drawn with thickness proportional to the wafer count. The resulting integrated TS model offers a clear, understandable map of both process structure and quality performance.

C. Phase III. Critical State Dimension Selection

In this phase, we leverage supervised learning to identify which of the attributes extracted as state dimensions in phase II have the most significant impact on EOL yield.

First, each wafer path $wp(c)$ is transformed into a feature vector $x_c \in \{0, 1\}^m$ by applying one-hot encoding on every observed state value, and the EOL yield is categorized. The random forest classifier f_{RF} is trained to predict the yield label $y(c)$. After that, using the Gini-importance score $\phi_{k,j,s}$ of the model, each attribute value's contribution to classification accuracy is calculated. To elevate these scores to the level

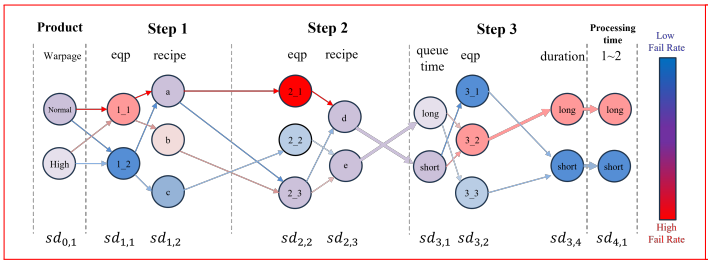

Fig. 4. Running example of the TS model including only critical state dimensions.

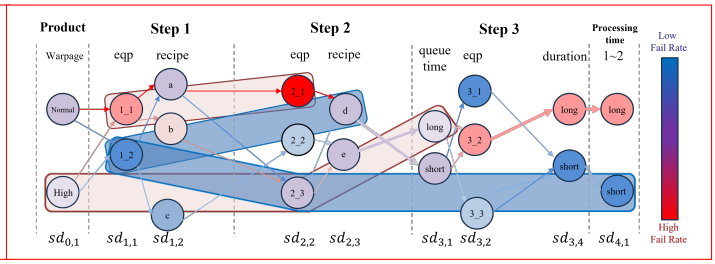

Fig. 5. Running example of the TS model with critical state combinations highlighted.

of state dimensions, we sum up each attribute value's Gini-importance score to calculate the state dimension importance, which is denoted by $I_{k,j} = \sum_{s \in sd_{k,j}} \phi_{k,j,s}$.

After that, state dimensions that have a state dimension importance higher than θ, a user-defined threshold, are selected as the critical state dimension, denoted as $csd_{k,j} = \{sd \in sd_{k,j} | I_{k,j} > \theta\}$. Note that θ is a user-specified cutoff chosen to retain the top-ranked dimensions in terms of yield importance.

Finally, a new TS model is built solely on the critical state dimension $csd_{k,j}$. This reduced TS model preserves only those dimensions that have a critical effect on EOL yield. This selection both enhances interpretability and concentrates subsequent analysis on the most critical factors influencing process performance. Fig. 4 demonstrates an example of critical state dimension selection and TS model reconstruction. Compared to Fig. 3, the new TS model in Fig. 4 includes only the critical-state dimensions identified by the random forest. This allows the TS model to intuitively present only the process attributes that impact the final yield.

D. Phase IV. Critical State Combination Derivation

During this phase, we identify the reduced transition-system traces for multivariate patterns of process attributes, where their co-occurrence is closely related to extreme yield outcomes.

First, we categorize each wafer into one of three yield groups—low, moderate, or high failure rate—based on its observed failure rate. We then use association rule mining to identify rules in which process attribute combinations serve as antecedents and the low- and high-yield classes as consequents. For each class $p \in \{\text{Low, High}\}$, we extract frequent itemsets from the set of reduced traces from phase III. An itemset is deemed frequent if its support meets the class-specific minimum support threshold $min\_sup(p) = ratio(p) \times 3/100$. Here, $ratio(p)$ denotes the proportion of wafers belonging to class p, allowing the minimum support to be adjusted according to class imbalance. This adjustment tackles class imbalance by adjusting the minimum support for each class based on its frequency.

From these frequent itemsets, we generate association rules that meet the user-defined minimum lift threshold λ, where lift quantifies the strength of association between an attribute combination and a yield class relative to statistical independence. Finally, we evaluate and rank each rule r by a novel rule score that balances rule strength and length.

$$\text{rule score(r)} = \text{lift}(r) * \sqrt[3]{\text{length}_{\text{antecedent}}(r)} \quad (1)$$

We then choose the top 250 rules based on rule score for each class. The antecedents of these top-ranked rules are determined as the critical state combinations L_{csc} and H_{csc}, each representing the state combinations associated with low- and high-yield wafers, respectively.

To visualize these findings, each critical combination in L_{csc} and H_{csc} is overlaid on the TS model, with highlights. This visual enhancement reveals the distinct multivariate pathways that most strongly characterize low- and high-yield behaviors, thereby offering clear, actionable insights into the interactions among key process factors. Fig. 5 illustrates how the derived critical state combinations are emphasized in the TS model shown in Fig. 4.

E. Phase V. Model Simplification

In this stage, the reduced TS model from phase IV is further simplified by merging low-impact or rarely used states into dummy categories, while keeping all critical states.

For each state dimension $sd_{k,j}$, which represents a process attribute, we compute the global mean failure rate $\mu_{k,j}$ and standard deviation $\sigma_{k,j}$ across all states in that dimension, as well as the coverage $\pi(s)$ for each state $s \in sd_{k,j}$. For each state dimension, any states that do not meet the variance range or coverage criteria are replaced with a dummy state and combined into a single dummy state.

This simplification greatly reduces the number of distinct nodes and edges in the TS model by removing statistically redundant or rare conditions, while maintaining all critical state combinations and their supporting transitions. Fig. 6 shows an example of simplifying the TS model using dummy state analysis. As seen in the figure, redundant states are replaced with dummy states and combined into a single dummy state, making the model more compact and easier to interpret, and highlighting the most relevant process conditions for yield analysis.

F. Phase VI. Evaluation and Interpretation

In the final phase, we evaluate the predictive ability and practical usefulness of the streamlined TS model and the resulting critical attribute combinations on a held-out test set of wafer traces. Each test wafer in the test set is classified based on the presence of critical combinations within its wafer

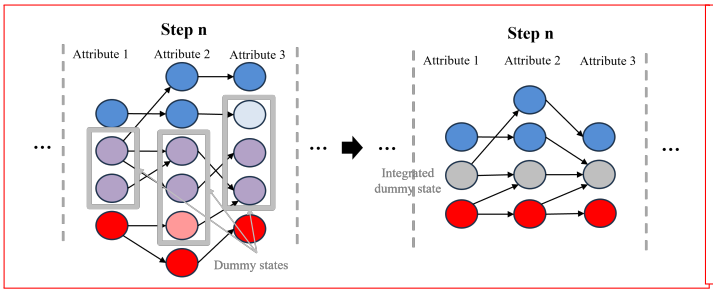

Fig. 6. Running example of simplifying the TS model by dummy state analysis.

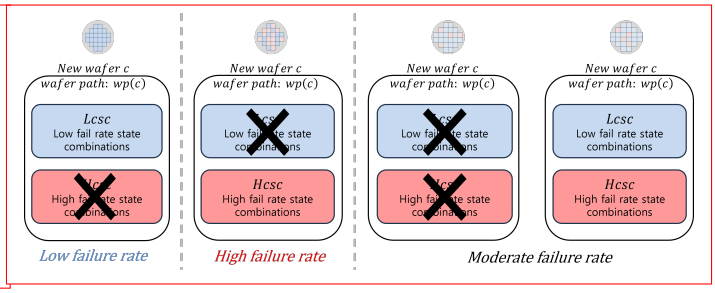

Fig. 7. Classification criteria of the new wafer in the test set.

path. If wp_c contains L_{csc} and is not a member of H_{csc}, it is classified to have a low failure rate. Conversely, the presence of any high failure rate combination H_{csc} and the absence of L_{csc} classifies the new wafer to have a high failure rate. In all other cases, the wafer is classified to have a moderate yield. We then compare these classification results to the actual EOL yield outcomes by using the confusion matrix. Fig. 7 shows the failure rate classification criteria of the new wafer in the test set. The classification can be formalized as follows:

$$\text{yield}_c = \begin{cases} \text{Low failure rate}, & \text{if } \text{wp}_c \in L_{csc}, \\ & \text{wp}_c \notin H_{csc} \\ \text{High failure rate}, & \text{if } \text{wp}_c \notin L_{csc}, \\ & \text{wp}_c \in H_{csc} \\ \text{Moderate failure rate}, & \text{Otherwise} \end{cases} \tag{2}$$

Using the confusion matrix, we can calculate various performance metrics for the classification results. These include balanced accuracy, high failure rate, recall, precision, F1-score, and extreme misclassification rate. These metrics help quantify whether the TS model framework and the extracted critical attribute combinations are valid and contain valuable information about the factors influencing yield.

IV. An Empirical Study

In this section, we apply the proposed TS model-based semiconductor analysis and visualization framework to a real-world dataset from the post-processing steps of a semiconductor manufacturing line. We demonstrate how each phase of our method adds to yield insights and predictive power, and we report classification performance on a held-out test set.

Our dataset initially contained 9,240 wafer cases collected over approximately three months. After removing 52 cases due to duplicate records or missing EOL failure-rate labels, we retained 9,188 cases, each undergoing 11 process steps, involving 141 unique equipment or resource identifiers, and characterized by 7 process attributes, including recipe, warpage, step durations, and queue times. All attributes used in this study were selected in consultation with domain experts at the manufacturing site. The experts recommended a set of attributes known to strongly influence EOL failure rates, and we included all of them to minimize potential information loss. In particular, equipment, recipe, queue time, and duration were used as step-wise event attributes for all process steps. Material information (material id) was incorporated only at Step 5 because it is recorded and used only at that step in the production setting. For warpage, we used the wafer-level measurement at the beginning of the post-processing stage, as domain experts indicated that warpage at this stage strongly affects the final yield. In addition, aggregated processing-time attributes were included by grouping steps into three ranges (Steps 1-2, 3-9, and 10-11).

We split the data into training and test sets based on the chronological order of each wafer's process start time, The earlier 80% of wafers in the manufacturing timeline were used for training, and the later 20% were used for testing. The training set includes 7,350 cases with an average failure rate of 0.71%, while the test set contains 1,838 cases with an average failure rate of 0.72%. The similar failure rates between the two sets indicate no noticeable data shift, making this temporal split appropriate for evaluation.

A. Phase I & II. Data Preparation and TS Model Construction

We first applied phases I and II of our method to the training set, constructing wafer-specific paths and developing the full TS model annotated with failure rates and wafer counts. Warpage, resources from steps 1-11, duration, queue time, recipe, material, and processing times of steps 1-2, 3-9, 10-11 are the process attributes extracted as state dimensions.

Fig. 8(a) displays the complete annotated TS model, where each node represents a specific process attribute as a state, and each edge shows the transition. Nodes are color-coded by average defect rate, and edge thickness indicates the number of wafers passing through. The TS model provides an intuitive understanding of the overall process flow, but because there are too many states, a state selection criterion is needed to simplify it.

B. Phase III. Critical State Dimension Selection

Next, to identify which process attributes significantly influence yield, a random-forest classifier was trained. The wafer paths were encoded using all 42 process attributes, which correspond to the initial 42 state dimensions in the full TS model. Using the Gini importance scores, we identified the top 23 most impactful process attributes. The TS model is then reconstructed by retaining only the 23 state dimensions corresponding to these critical attributes. Fig. 8(b) shows the reconstructed TS model using only these critical attribute states. Compared to Fig. 8(a), fewer state dimensions remain

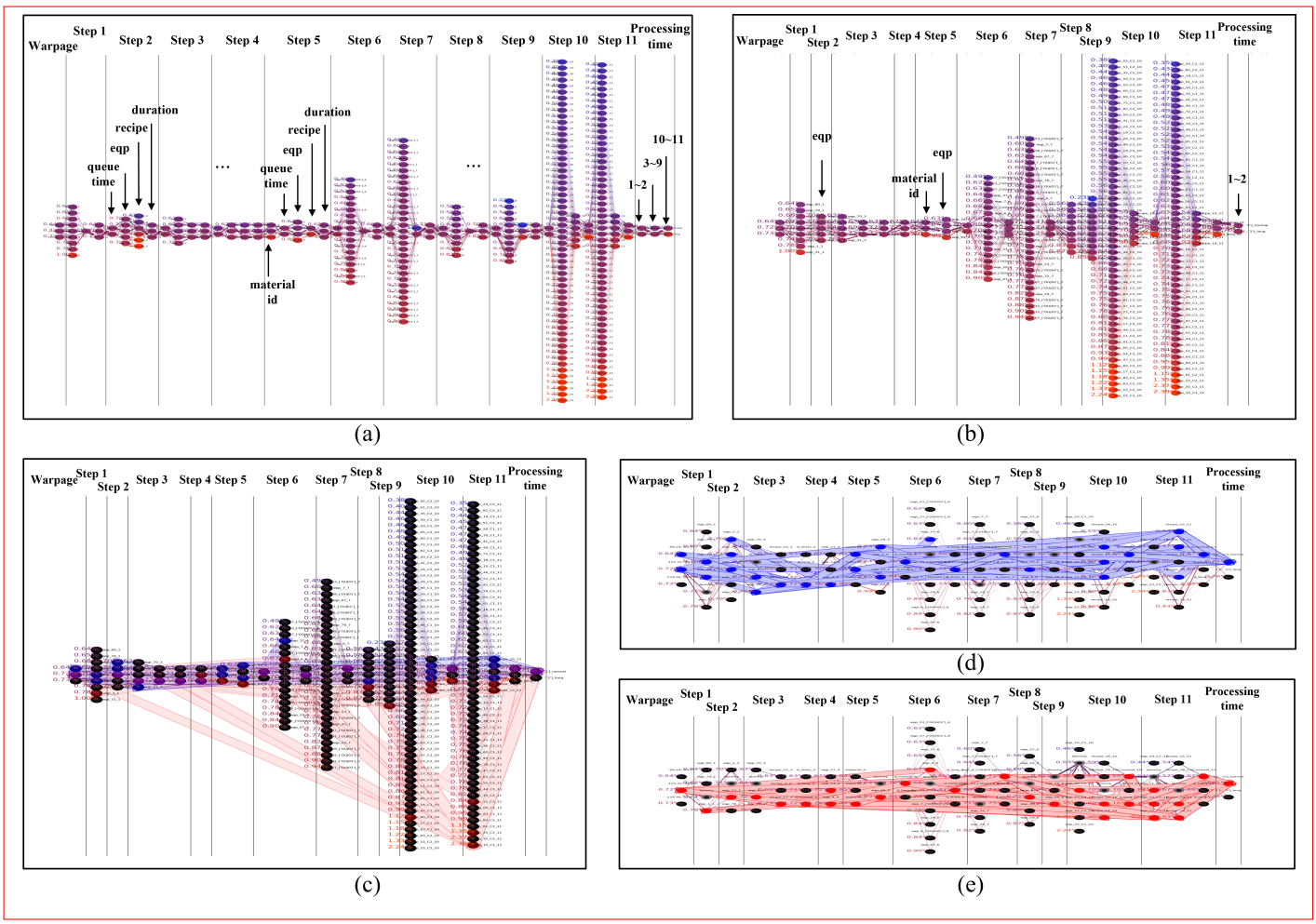

Fig. 8. Transition system of each phase of the methodology: (a) the full, failure-annotated TS model (Phases I&II), (b) the TS model constructed only with critical state dimensions (Phase III), (c) the TS model overlaid with low and high failure state combinations (Phase IV), (d) the simplified TS model highlighted with low failure rule combinations (Phase V), (e) the simplified TS with dummy states and highlighted high failure rule combinations (Phase V).

after selection (e.g., at step 2 only the equipment dimension remains, and at step 5 only the material id and equipment dimensions remain). As a result, by filtering from 42 down to 23 critical state dimensions, the total number of individual states is reduced from 287 to 252. This removes less informative dimensions and all their associated states, keeping only the states that critically affect yield. Engineers found this streamlined graph easier to interpret because it emphasizes only the critical dimensions.

C. Phase IV. Critical State Combination Derivation

In phase IV, association rules were mined separately for low and high failure rate wafer classes, using class-adjusted support thresholds to address imbalance. We extracted and ranked the top 250 rules per class with a custom rule score that balances lift and antecedent length, then overlaid these key combinations onto the TS model. Rules are derived, such as "equipment 4_3 at step 4 + step 4 queue time: short + normal warpage" for low-yield cases or "equipment 3_10 at step 6 + Recipe e at step 3 + step 3 queue time: long + equipment 6_2 at step 6" for high-yield cases.

The resulting TS model is shown in Fig. 8(c). It clearly illustrates how process attributes interact as states. States that seem harmless on their own can become highly or less prone to failure when combined with other states. This direct mapping of rule antecedents onto the TS model allows process engineers to visually trace which combinations of process factors most strongly lead to failure or success.

D. Phase V. Model Simplification

To further simplify the model, all non-critical states with failure rates within $\pm 0.5\sigma$ of their mean or that occurred in fewer than 0.2% of wafers were merged into a single dummy state per dimension. Fig. 8(d) shows the simplified TS model with low failure rate critical state combinations. Fig. 8(e) displays the simplified TS model with high failure rate critical state combinations. The TS model was simplified by combining rarely visited or statistically insignificant nodes into a single dummy state, significantly reducing visual complexity while keeping all critical states and transitions. This step produced a highly streamlined model that still retains full accuracy of the most important process features.

TABLE III
EVALUATION RESULT OF THE PROPOSED METHOD

Method	Feature used	Balanced accuracy	High fail rate recall	High fail rate precision	High fail rate F1-score	Extremity Misclassification Rate
Random selection	-	0.323	0.214	0.147	0.174	0.216
Cho et al., 2021	Eqp	0.375	0.244	0.345	0.286	0.262
ANN	Eqp	0.355	0.216	0.333	0.262	0.142
	Eqp + attributes	0.402	0.304	0.389	0.341	0.130
Dummy coded regression	Eqp	0.330	0.407	0.220	0.286	0.484
	Eqp + attributes	0.366	0.439	0.263	0.329	0.349
Proposed method	Eqp	0.394	0.414	0.356	0.383	0.161
	Eqp + attributes	**0.468**	**0.501**	**0.398**	**0.443**	**0.122**

E. Phase VI. Evaluation and Interpretation

Finally, in phase VI, we applied our three-class classification method to categorize test wafers as low, moderate, or high failure rate based on specific rule combinations. This was tested on a set of 1,838 wafers. We compared the results to four baseline approaches: random selection, the equipment-only TS model [7], an ANN model [20], and dummy-coded regression. To determine if performance improvements mainly resulted from using a broader set of process attributes or from our methodology, we ran each method twice: once with only equipment identifiers and once with the complete attribute set (equipment + all process attributes). This dual-test setup allowed us to assess how much of the predictive gain was due to additional attributes versus our approach.

We evaluate each setting using multiple metrics. Balanced Accuracy (BA) is reported for our three-class setting (Low, Moderate, High) because standard accuracy can be dominated by the majority class under imbalance. Balanced accuracy is the macro-average of class-wise recall and is calculated as follows:

$$\mathrm{BA} = \frac{1}{3} \sum_{c \in \{L,M,H\}} \frac{\mathrm{TP}_c}{N_c}, \tag{3}$$

where TP_c is the number of correctly classified wafers in class c, and N_c is the total number of wafers that truly belong to class c in the test set. By averaging recall across classes, BA reflects predictive performance for all classes in a balanced way regardless of their frequency, providing a more reliable assessment when the class distribution is imbalanced. In addition, we report precision, recall, and F1-score for the high-failure class to evaluate how effectively the models identify high-failure wafers. Finally, we report the Extremity Misclassification Rate (EMR), which measures the proportion of true high-failure wafers incorrectly classified as low-failure. This metric specifically quantifies the model's tendency to commit the most critical operational error, which is to classify high-failure wafers as low-failure. Table III presents the results of the wafer classification task, showing that our method achieved the best performance across all evaluation metrics.

The evaluation results reveal several key insights. The strong classification performance of our proposed method indicates that the critical attributes and attribute combinations identified earlier are accurate predictors of wafer yield outcomes. Specifically, the 23 critical process attributes and the associated interactions identified through association rule mining were confirmed by their high predictive accuracy on the test dataset. Visualizing these findings with the simplified TS model enhances their practical relevance, providing decision-makers with clear and actionable insights. Therefore, the evaluation results demonstrate the effectiveness of our approach, validate the critical yield factors, and emphasize the importance of incorporating these insights into the TS model for straightforward, informed decision-making.

V. CONCLUSION

This study presents a six-phase framework for analyzing and visualizing semiconductor manufacturing processes using a TS model to enhance yield. It combines process mining, data mining, and statistical analysis to identify key process attributes that significantly influence EOL yield. Additionally, it uncovers important combinations of process attributes, addressing multicollinearity often encountered in manufacturing processes. Validation was performed with a real semiconductor post-processing event log. The empirical study demonstrated the effectiveness of our method in pinpointing critical process attributes and their combinations and showed that visualizing these results with the TS model offers valuable support for decision-makers.

Furthermore, the framework is computationally practical for large-scale industrial deployments. Although the random forest classifier and association rule mining in phase III and IV are the most intensive components, they exhibit favorable time complexity. Random forest scales log-linearly with the dataset size, supported by a parallelizable structure that enables high-speed processing [21]. Additionally, although association rule mining is exponential by nature, its computational cost is manageable through dimensionality reduction and support-based pruning [22]. All other phases possess strictly linear time complexity relative to the volume of data.

In terms of practical positioning, our framework complements existing industrial and process-mining tools. In semiconductor practice, many yield analysis and monitoring tools focus on time-series signals or isolate a limited set of equipment and attributes, which is effective for detecting excursions but can fragment end-to-end interpretation across the manufacturing flow. Conversely, process mining platforms such as Celonis [23] provide an interpretable view of the overall process flow, yet their support for systematically incorporating and jointly analyzing fine-grained process attributes (e.g., equipment, recipe, queue time, duration, and product characteristics) across all steps is less explicit. Our TS-based approach unifies both perspectives by modeling the complete process while simultaneously identifying and visualizing high-impact attributes and their multivariate interactions on the same process-structured view, enabling engineers to trace yield-critical pathways and conditions in an integrated and actionable manner.

However, the framework operates offline using historical event logs and does not yet include real-time sensor signals or

adaptive responses to changing equipment conditions. Future work will develop an online, incremental TS model capable of processing live data streams, dynamically updating state combinations and rule thresholds, and proactively recommending optimal resource paths for closed-loop yield optimization.

Additionally, our current state abstraction, which treats attributes separately, could be improved by combining stage-level states to capture overall stage effects and simplify complexity. Investigating these more comprehensive abstractions to further improve model interpretability, robustness, and insights is an important area for future research.

REFERENCES

[1] Y. Koren, S. J. Hu, and T. W. Weber, "Impact of manufacturing system configuration on performance," *CIRP annals*, vol. 47, no. 1, pp. 369–372, 1998.

[2] W. J. Moore and A. G. Starr, "An intelligent maintenance system for continuous cost-based prioritisation of maintenance activities," *Computers in industry*, vol. 57, no. 6, pp. 595–606, 2006.

[3] N. Orsini, A. Moore, and A. Wolk, "Interaction analysis based on shapley values and extreme gradient boosting: a realistic simulation and application to a large epidemiological prospective study," *Frontiers in Nutrition*, vol. 9, p. 871768, 2022.

[4] D.-H. Lee, C.-H. Lee, S.-H. Choi, and K.-J. Kim, "A method for wafer assignment in semiconductor wafer fabrication considering both quality and productivity perspectives," *Journal of Manufacturing Systems*, vol. 52, pp. 23–31, 2019.

[5] C.-F. Chien, K.-H. Chang, and W.-C. Wang, "An empirical study of design-of-experiment data mining for yield-loss diagnosis for semiconductor manufacturing," *Journal of Intelligent Manufacturing*, vol. 25, no. 5, pp. 961–972, 2014.

[6] C.-H. Lee, D.-H. Lee, Y.-M. Bae, S.-H. Choi, K.-H. Kim, and K.-J. Kim, "Approach to derive golden paths based on machine sequence patterns in multistage manufacturing process," *Journal of Intelligent Manufacturing*, vol. 33, no. 1, pp. 167–183, 2022.

[7] M. Cho, G. Park, M. Song, J. Lee, B. Lee, and E. Kum, "Discovery of resource-oriented transition systems for yield enhancement in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 34, no. 1, pp. 17–24, 2020.

[8] D. Jiang, W. Lin, and N. Raghavan, "A gaussian mixture model clustering ensemble regressor for semiconductor manufacturing final test yield prediction," *Ieee Access*, vol. 9, pp. 22 253–22 263, 2021.

[9] K. Nakata, R. Orihara, Y. Mizuoka, and K. Takagi, "A comprehensive big-data-based monitoring system for yield enhancement in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 30, no. 4, pp. 339–344, 2017.

[10] D. Djurdjanovic and J. Ni, "Stream-of-variation (sov)-based measurement scheme analysis in multistation machining systems," *IEEE transactions on automation science and engineering*, vol. 3, no. 4, pp. 407–422, 2006.

[11] Q. Huang and J. Shi, "Stream of variation modeling and analysis of serial-parallel multistage manufacturing systems," *J. Manuf. Sci. Eng.*, vol. 126, no. 3, pp. 611–618, 2004.

[12] F. Yang, S. Jin, and Z. Li, "A comprehensive study of linear variation propagation modeling methods for multistage machining processes," *The International Journal of Advanced Manufacturing Technology*, vol. 90, no. 5, pp. 2139–2151, 2017.

[13] H. K. Lim, Y. Kim, and M.-K. Kim, "Failure prediction using sequential pattern mining in the wire bonding process," *IEEE Transactions on Semiconductor Manufacturing*, vol. 30, no. 3, pp. 285–292, 2017.

[14] B. Kamsu-Foguem, F. Rigal, and F. Mauget, "Mining association rules for the quality improvement of the production process," *Expert systems with applications*, vol. 40, no. 4, pp. 1034–1045, 2013.

[15] F. Psarommatis, G. May, P.-A. Dreyfus, and D. Kiritsis, "Zero defect manufacturing: state-of-the-art review, shortcomings and future directions in research," *International journal of production research*, vol. 58, no. 1, pp. 1–17, 2020.

[16] C. Bergès, J. Bird, M. D. Shroff, R. Rongen, and C. Smith, "Data analytics and machine learning: root-cause problem-solving approach to prevent yield loss and quality issues in semiconductor industry for automotive applications," in *2021 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*. IEEE, 2021, pp. 1–10.

[17] W. Van Der Aalst, "Process mining: Overview and opportunities," *ACM Transactions on Management Information Systems (TMIS)*, vol. 3, no. 2, pp. 1–17, 2012.

[18] M.-D. Ma, D. S.-H. Wong, S.-S. Jang, and S.-T. Tseng, "Fault detection based on statistical multivariate analysis and microarray visualization," *IEEE Transactions on industrial informatics*, vol. 6, no. 1, pp. 18–24, 2009.

[19] Y. Lee and Y. Roh, "An expandable yield prediction framework using explainable artificial intelligence for semiconductor manufacturing," *Applied Sciences*, vol. 13, no. 4, p. 2660, 2023.

[20] M. Al-Kharaz, B. Ananou, M. Ouladsine, M. Combal, and J. Pinaton, "Quality prediction in semiconductor manufacturing processes using multilayer perceptron feedforward artificial neural network," in *2019 8th international conference on systems and control (ICSC)*. IEEE, 2019, pp. 423–428.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining—a general survey and comparison," *ACM sigkdd explorations newsletter*, vol. 2, no. 1, pp. 58–64, 2000.

[23] Celonis. (2026) Celonis official website. Accessed: Jan. 1, 2026. [Online]. Available: https://www.celonis.com/