Sequence-Aware Inline Measurement Attribution for Good-Bad Wafer Diagnosis

DM: Big Data Management and Machine Learning

Kohei Miyaguchi*
IBM Research – Tokyo
Tokyo, Japan
koheimiyaguchi@gmail.com

Masao Joko
IBM Semiconductors
Tokyo, Japan
mjoko@jp.ibm.com

Rebekah Sheraw
IBM Semiconductors
Albany, NY, USA
rebekah.sheraw@ibm.com

Tsuyoshi Idé
IBM Semiconductors
Yorktown Heights, NY, USA
tide@us.ibm.com

Abstract—How can we identify problematic upstream processes when a certain type of wafer defect starts appearing at a quality checkpoint? Given the complexity of modern semiconductor manufacturing, which involves thousands of process steps, cross-process root cause analysis for wafer defects has been considered highly challenging. This paper proposes a novel framework called Trajectory Shapley Attribution (TSA), an extension of Shapley values (SV), a widely used attribution algorithm in explainable artificial intelligence research. TSA overcomes key limitations of standard SV, including its disregard for the sequential nature of manufacturing processes and its reliance on an arbitrarily chosen reference point. We applied TSA to a good-bad wafer diagnosis task in experimental front-end-of-line processes at the NY CREATES Albany NanoTech fab, aiming to identify measurement items (serving as proxies for process parameters) most relevant to abnormal defect occurrence.

Fig. 1. Defect attribution. Given a defect flag y and a set of observed in-line measurements x, compute the responsibility score $s_i(x)$ for each measurement item as a proxy for the underlying process.

I. INTRODUCTION

Root cause analysis (RCA) of wafer defects is a key challenge throughout all stages of semiconductor manufacturing, from process integration to high-volume production. Performing RCA across multiple process steps is particularly difficult due to the complex sequence and combination of physically and chemically diverse processes, which are carried out using various equipment across hierarchical levels such as wafer, lot, and batch.

RCA is often approached as a by-product of virtual metrology (VM) modeling, which aims to predict a quality metric as a function of process variables. Typically, VM is formulated as a regression or classification problem. However, existing approaches still face challenges that hinder their practicality. Regularized linear regression combined with variable selection techniques, a commonly adopted VM approach (e.g., [1]–[3]), essentially performs variable-wise correlation analysis for attribution, disregarding the sequential nature of the processing route. While recurrent neural networks (RNNs) and Transformers can capture complex nonlinear dependencies in sequential processes (e.g., [4]–[8]), they operate as black boxes, making input attribution a non-trivial task.
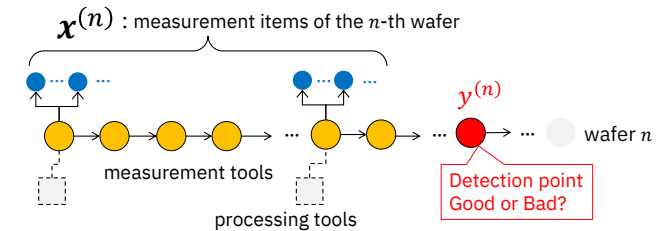
To address the interpretability challenges of deep-learning-based models and facilitate actionable insights, techniques developed in the explainable artificial intelligence (XAI) research community are gaining increasing attention in semiconductor data analytics. One key approach is *Shapley values* (SV) [9]–[11], which compute attribution scores for each input variable in a prediction model, given an observed input-output pair. When applied to a VM model, SV can provide valuable insights into which input variables are most responsible for an observed process quality metric. However, for cross-process defect attribution, existing approaches (e.g., [12]–[15]) have two main limitations: they fail to respect the sequential nature of semiconductor processes and rely on an arbitrarily chosen baseline (or reference) point. These issues stem from the original modeling assumptions of SV.

In this paper, we propose a model-agnostic input attribution algorithm called *Trajectory Shapley Attribution (TSA)*, a novel extension of SV designed for scenarios where input variables form a trajectory. As discussed later, TSA eliminates dependency on an arbitrary baseline point by incorporating a data imputation algorithm that leverages the lot-wafer hierarchy in semiconductor manufacturing. We apply our methodology to in-line measurements from a state-of-the-art front-end-of-line (FEOL) process at the Albany NanoTech fab to demonstrate the utility of our approach.

---

* Kohei Miyaguchi is currently affiliated with LY Research, Japan.

II. RELATED WORK

The overall framework of in-line measurement attribution aligns with advanced process control (APC), which encompasses virtual metrology (VM), variable attribution, and recipe tuning as key sub-tasks. For a recent and comprehensive review of APC, VM, and related tasks, see [16]. However, most existing studies focus on specific processes. While cross-process dependency analysis and attribution of wafer quality issues are crucial for accelerating yield improvement [17], the vision of digital twins—namely, data-driven, semi-automated yield control—remains far from reality. In many mass production fabs, yield ramp-up still requires significant manual effort.

The integration of model-agnostic XAI techniques with advanced quality prediction models is a major research trend in semiconductor analytics. This approach aims to strike an optimal balance between predictive power and interpretability. Among various XAI methods (see, e.g., [18] for a general overview), the Shapley value (SV) is the most widely used in semiconductor analytics. Torres et al. [12] and Stoner et al. [13] were among the first to leverage SV for fab-wide yield control. Lee et al. [14] applied SV to ten different regression models for yield prediction and attribution, demonstrating its model-agnostic nature. Guo et al. [15] further utilized SV for chip probing yield in a good-bad classification setting.

Ironically, despite its widespread adoption, most studies treat SV as a *black-box* software tool, paying little attention to the validity of its modeling assumptions. We argue that the original definition of SV may not be fully appropriate for semiconductor manufacturing due to two key issues: 1) variable coalitions include physically inadmissible process routes, and 2) the default baseline value in SHAP, a popular Python implementation [11], disregards the wafer-lot hierarchy. The proposed Trajectory Shapley Attribution (TSA) addresses these limitations.

Finally, the Transformer architecture has recently gained attention in semiconductor analytics [4], [5] due to its ability to directly model dependencies between sequential input variables via the self-attention mechanism [19]. However, Transformers are known to be even more data-hungry than RNNs. It remains an open question whether Transformers provide reliable predictive performance in semiconductor applications, given the challenge of extremely low effective sample sizes, as discussed in Sec. III-B. For instance, Han et al. [5] evaluated their proposed Transformer-based VM model using only 344 training samples and 36 validation samples—orders of magnitude smaller than the "internet-scale" datasets used in text, speech, and vision domains, where RNN and Transformer architectures have demonstrated remarkable success [20], [21]. Moreover, in the VM setting, Transformer input 'tokens' are aggregated in some manner to predict the target variable, introducing an additional explainability challenge.

III. PROBLEM SETTING

This section outlines the machine learning tasks we address and the key modeling assumptions.

A. Attribution vs. Prediction

The attribution task considered in this paper is defined as follows:

Definition III.1 (In-line measurement attribution)**.** *Given a binary wafer quality label $y \in \{0, 1\}$ and a set of in-line measurements $\boldsymbol{x} \in \mathbb{R}^D$ (i.e., a D-dimensional real-valued vector), compute the attribution score $s_i(\boldsymbol{x})$ that quantifies the contribution of each measurement item $i = 1, \ldots, D$ to the outcome y, serving as a proxy for the underlying process.*

Here, the binary label is defined as 1 for bad wafers and 0 for good wafers. As illustrated in Fig. 1, each measurement tool is associated with a specific processing tool and acts as its proxy. Some measurements may be performed multiple times on the same wafer due to reentrant processes. In such cases, only the latest observations are retained, allowing the in-line measurement history to be represented as a fixed-dimensional vector $\boldsymbol{x} \in \mathbb{R}^D$ per wafer, although some measurement items may be missing. A single measurement tool can produce multiple measurement items. For example, a set of critical dimensions may be measured at multiple locations on a wafer, with their mean and standard deviation recorded. Consequently, D can be on the order of thousands.

Typically, solving the attribution task requires a predictive model for y as a function of $\boldsymbol{x}$, which we assume unknown in our problem setting. Since y is binary, we consider the following classification problem:

Definition III.2 (Wafer classifier training)**.** *Given a training dataset $\mathcal{D} \triangleq \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^N$, train a classifier $p(\boldsymbol{x})$ that predicts the probability of $y = 1$ (bad) given an input $\boldsymbol{x}$.*

Here, $\boldsymbol{x}^{(n)} \in \mathbb{R}^D$ denotes the set of observed in-line measurement items for the n-th wafer, and $y^{(n)}$ is its corresponding wafer quality label.

B. Key Data Characteristics

Although the tasks defined above resemble standard attribution and classification problems, semiconductor manufacturing introduces specific challenges that must be considered:

1) Wafers are not necessarily independent due to lot- and batch-based processing.
2) Wafers are processed sequentially, implying an inherent ordering among measurement tools.
3) Most in-line measurement values are unavailable due to wafer-, lot-, and batch-wise sampling policies.

The first and second points render many off-the-shelf machine learning tools ineffective, as most assume sample independence and permutation invariance among variables. Regarding the third point, since wafers are randomly selected for measurement at each tool, the resulting measurement trajectories are nearly unique for each wafer. Furthermore, in the yield ramp-up phase, where various processing recipes are tested simultaneously, the diversity of processing routes is potentially large. This further complicates model training, as the *effective*

sample size, defined as the number of wafers per unique processing or measurement condition, may be extremely small—possibly as low as one.

To illustrate the characteristics of a typical wafer dataset, we summarize key statistics from the dataset used in our empirical evaluation. As shown in Table I, the dimensionality D far exceeds the number of wafers, making the prediction task ill-posed. The extremely high missing rate exacerbates this challenge.

TABLE I
SUMMARY STATISTICS OF OUR DATASET COLLECTED FROM THE ALBANY NY CREATES FAB.

N	D	Average missing rate	$\text{Prob}(y = 1)$
572	23 498	94%	51%

IV. PRELIMINARY: SHAPLEY VALUES

Before getting into the details of the proposed approach, this section reviews existing Shapley value (SV)-based attribution methods.

A. Definition of SV

The goal of Shapley Value (SV) is to evaluate the contribution of each input variable to a model's prediction. Suppose there are D input variables, denoted as $\boldsymbol{x} \triangleq (x_1, \ldots, x_D)^\top$, and a prediction model $y = f(\boldsymbol{x})$, which can produce the probability $p(\boldsymbol{x})$ in the classification task defined in Definition III.2. Formally, the SV for the i-th variable is defined as [22], [23]

$$\text{SV}_i(\boldsymbol{x}^t) = \frac{1}{D} \sum_{k=1}^{D} \sum_{\mathcal{S} \in \mathcal{S}_k^{(i)}} \binom{D-1}{k-1}^{-1} [v(\mathcal{S}) - v(\mathcal{S} - \{i\})],$$
(1)

where $\boldsymbol{x}^t$ is a test sample at which SV is calculated, and $\mathcal{S}_k^{(i)}$ denotes the collection of *all* index sets of size k (out of the D members) that include i (Examples are given below). The notation $\mathcal{S} - \{i\}$ represents the index set obtained by removing i from $\mathcal{S}$. The dependency on $\boldsymbol{x}^t$ indicates that SV offers a *local* explanation, meaning that the explanation is about the specific test sample $\boldsymbol{x}^t$. This contrasts with traditional correlation analysis, where the relevance to a target variable y is quantified through a statistic defined on a population of samples.

In the original game-theoretic definition [9], the term $v(\mathcal{S}) - v(\mathcal{S} - \{i\})$ quantifies the influence of the i-th player within the coalition $\mathcal{S}$, assuming that $v(\mathcal{S})$, known as the *value function*, represents the outcome achieved by the coalition. Since the number of such possible sets is $\binom{D-1}{k-1}$, SV_i computes the average contribution of the i-th variable across all possible variable combinations. While multiple definitions of $v(\cdot)$ are possible in XAI applications [22], [23], it is typically chosen as the prediction function itself, as discussed in the next subsection.

B. Handling Non-Participating Variables

A critical question when applying SV in XAI tasks is how to reconcile $v(\mathcal{S})$, which is defined over a subset of variables, with $f(\boldsymbol{x})$, which requires a full D-dimensional input. To be specific, consider the case where $(i, D) = (i, 3)$ and $\mathcal{S} = \{1, 2\}$. In SHAP's default setting, the value function $v(\mathcal{S})$ evaluated at $\boldsymbol{x} = \boldsymbol{x}^t$ is defined as

$$v(\{1, 2\}) = f(x_1^t, x_2^t, \bar{x}_3), \quad \text{(baseline Shapley)} \quad (2)$$

where $\bar{x}_3$ is the population mean of x_3. In this setting, the selected variables in $\mathcal{S}$ retain their actual values from $\boldsymbol{x}^t$, while non-selected variables are replaced with default values, which in this case is the population mean. In other application domains, alternative default values are used; for example, in computer vision, pixel values may be set to zero. These variants are collectively referred to as *baseline Shapley values*.

Another widely used approach marginalizes the non-selected variables using training data:

$$v(\{1, 2\}) = \frac{1}{N} \sum_{n=1}^{N} f(x_1^t, x_2^t, x_3^{(n)}), \quad \text{(CE Shapley)} \quad (3)$$

which is known as the *conditional expectation (CE) Shapley value* [22].

C. Limitations of SV in Semiconductor Applications

There are two major limitations when applying SV-based attribution using Eq. (1) to semiconductor manufacturing data.

First, while SV provides a theoretically fair method for evaluating variable dependencies, it does not account for the sequential nature of semiconductor processes. In the present context, a set $\mathcal{S}$ can be viewed as a hypothetical processing route, represented by in-line measurements as proxies. However, many variable combinations are physically inadmissible—for instance, pattern formation without etching. It is unclear whether attribution scores derived from such inadmissible combinations carry meaningful information or should be regarded as artifacts.

Second, SV relies on an arbitrarily chosen baseline point. The use of the population mean as a reference can be problematic, as it disregards the wafer-lot hierarchy. A similar issue arises with CE Shapley, where averaging over training samples does not necessarily preserve the structure of semiconductor processing dependencies.

Beyond these issues, SV is also known to exhibit *deviation insensitivity* [24] when the prediction function itself is used as the value function. This means that while SV is effective for characterizing the general behavior of $f(\cdot)$, it is not well-suited for explaining deviations of the form $f(\boldsymbol{x}^t) - y^t$, which is often of primary interest in defect diagnosis tasks.

V. PROPOSED APPROACH

Given the limitations of standard SVs, we propose a novel SV-based framework for measurement attribution. This section introduces two key components of the trajectory Shapley attribution (TSA) framework: lot-aware kernel imputation and trajectory Shapley values.

A. Lot-Aware Kernel Imputation

As discussed earlier, in-line measurement data contain many missing entries due to lot-, wafer-, and item-specific sampling policies. A key challenge is that missing patterns in in-line measurements are neither purely random nor fully systematic. In the TSA framework, we adopt a relatively simple yet effective data imputation approach based on the lot-wafer hierarchy. This imputation method is crucial, as it defines the concept of 'non-participation' for variables.

Let $W_{n,m}$ denote the kernel function between wafers n and m. To compute W, we first define the lot-sharing count matrix C, where each element $C_{n,m}$ represents the number of operations shared between wafers n and m within the same lot, up to the point where either $y^{(n)}$ or $y^{(m)}$ is measured. $C_{n,n}$ corresponds to the total number of operations that wafer n underwent before the measurement of $y^{(n)}$. Using C, we define the similarity matrix W based on the Jaccard index:

$$W_{n,m} \triangleq \frac{C_{n,m}}{C_{m,n} + C_{n,n} - C_{n,m}}, \quad n \neq m. \quad (4)$$

Each element of W lies within the unit interval $[0, 1]$, with diagonal elements $W_{n,n}$ set to zero.

Finally, LAKI imputes a missing element, denoted as $x_k^{(n)} = \bot$, using a weighted average based on W:

$$x_k^{(n)} \leftarrow \frac{1}{\sum_j W_{n,j}} \sum_m W_{n,m} x_k^{(m)}, \quad (5)$$

where terms with $x_k^{(m)} = \bot$ are excluded from the summation. If no peers with observed values for item k exist, we set $x_k^{(n)}$ to $\bar{x}_k$, the mean of item k.

B. Trajectory Shapley Values

Suppose we have a VM model $f(\boldsymbol{x})$ that predicts an outcome variable y given a measurement trajectory $\boldsymbol{x}$. We generalize the standard SV definition (1) as follows:

$$s_i(\boldsymbol{x}^t) = \sum_{\mathcal{S} \in \mathcal{T}^{(i)}} \mu(\mathcal{S}) \left[v(\mathcal{S}) - v(\mathcal{S} - \{i\}) \right], \quad (6)$$

where $\boldsymbol{x}^t$ denotes an input trajectory of interest, and $\mathcal{T}^{(i)}$ is the set of all *physically admissible* process trajectories that include the i-th measurement item. In semiconductor manufacturing, physically admissible trajectories are those that progress sequentially from the first process without skipping intermediate steps. For example, for $(i, D) = (2, 3)$, the admissible trajectories are (x_1, x_2) and (x_1, x_2, x_3). The prefactor $\mu(\mathcal{S})$ is a weighting function that averages over all admissible configurations. To properly determine $\mu(\mathcal{S})$, we impose the sum rule on the trajectory Shapley values:

$$\sum_{i=1}^{D} s_i(\boldsymbol{x}^t) = v(\mathbf{1}) - v(\emptyset), \quad (7)$$

where $v(\mathbf{1})$ and $v(\emptyset)$ denote the value function under full variable participation and no participation, respectively. Equations (6) and (7) together define the trajectory Shapley value within our framework.
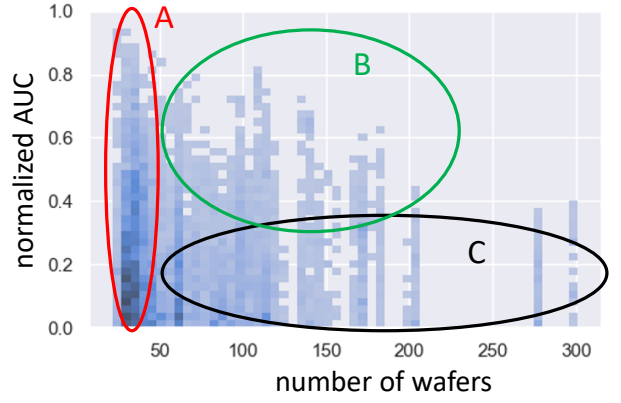

Fig. 2. Results of univariate correlation analysis. Most measurement items either provide weak signals or have too few samples, highlighting the need for trajectory-based analysis.

To define the value function, we relabel the index set $\mathcal{S}$ and its complement $-\mathcal{S}$ so that $\boldsymbol{x} = (\boldsymbol{x}_\mathcal{S}, \boldsymbol{x}_{-\mathcal{S}})$, where the subvector $\boldsymbol{x}_\mathcal{S}$ collects all x_i's for $i \in \mathcal{S}$. The value function in TSA, evaluated at $\boldsymbol{x} = \boldsymbol{x}^t$, is then given by:

$$v(\mathcal{S}) = f(\boldsymbol{x}_\mathcal{S} = \boldsymbol{x}^t, \boldsymbol{x}_{-\mathcal{S}} = \bot), \quad \text{(TSA)} \quad (8)$$

where $\boldsymbol{x}_{-\mathcal{S}} = \bot$ symbolically represents that the non-participating variables are treated as missing values. Unlike standard SV, where missing variables are replaced with a fixed baseline value, TSA employs LAKI to adaptively impute $\bot$ based on the wafer and lot IDs.

Our theoretical analysis has established the following result:

Theorem V.1 (Trajectory Shapley Value). *The weighting function $\mu(\mathcal{S})$ is uniquely determined by the condition (7). With this $\mu(\mathcal{S})$, the trajectory Shapley value is given by:*

$$s_i(\boldsymbol{x}^t) = f(\boldsymbol{x}_{1:i}^t, \boldsymbol{x}_{(i+1):D}^0) - f(\boldsymbol{x}_{1:(i-1)}^t, \boldsymbol{x}_{i:D}^0), \quad (9)$$

where $\boldsymbol{x}_{s:e} \triangleq (x_s, x_{s+1}, \ldots, x_e)$ for $s \leq e$, and $\emptyset$ (the empty set) otherwise. $\boldsymbol{x}_{(i+1):D}^0$ and $\boldsymbol{x}_{i:D}^0$ are baseline values computed using the LAKI algorithm (5).

The proof is omitted due to space constraints. In the next section, we apply this attribution score to a good-bad wafer characterization scenario.

VI. EMPIRICAL EVALUATION

We applied the Trajectory Shapley Attribution (TSA) approach to a process-limited yield (PLY) root-cause analysis task for state-of-the-art front-end-of-line (FEOL) processing routes in the Albany NanoTech fab. In this empirical evaluation, we focused on a specific scenario involving a burst of a defect type over a certain period. Specifically, we collected data from $N = 572$ wafers, approximately half of which exhibited the defect. A binary target variable $y \in \{0, 1\}$ was obtained at a PLY measurement point, where $y = 1$ indicates the presence of a specific defect type, and $y = 0$ otherwise. The input variables $\boldsymbol{x}$ consisted of upstream in-line measurements from hundreds of measurement processes.

As shown in Table I, the missing value rate is extremely high—only about 6% of the entries contain actual measurements. Additionally, the dimensionality of the input space (D) is much larger than the number of wafers (N), making the prediction task highly ill-posed. In such settings, data-hungry deep learning models are prone to overfitting. Our goal is to identify which measurement items are most relevant to the defect burst by computing attribution scores.

A. Baseline: Conventional Univariate Analysis

As a baseline approach, we conducted a conventional univariate correlation analysis. Using the training dataset $\mathcal{D}$, we trained a separate univariate logistic regression classifier for each of the $D = 23\,498$ measurement items to predict the good-bad wafer flag y. No LAKI imputation was used in this step. Due to missing values, different measurement items had varying sample sizes, which impacted prediction accuracy. For each measurement item, we computed the area under the curve (AUC) score of the receiver operating characteristic (ROC) curve.

Figure 2 presents the results, where the vertical axis represents the normalized AUC, defined as $|2 \times \text{AUC} - 1|$, which takes a value of 0 for the random classifier and 1 for the perfect classifier. As shown, most measurement items produced unreliable results due to insufficient sample sizes (Region A). Generally, measurement items with larger sample sizes exhibited weaker predictive performance (Region C). Although a few items displayed moderate predictive power, no clear dominant predictor emerged (Region B), providing limited actionable insights.

B. Training a Good-Bad Classifier

As the first step in the TSA analysis, we trained a binary classifier in the multivariate setting. To mitigate potential overfitting issues (see Sec. III-B), we employed ℓ_2-regularized logistic regression to model $p(\boldsymbol{x})$, the probability of observing $y = 1$ (bad wafer). The regularization strength was selected via cross-validation. With LAKI imputation, the true positive rate was 0.80, compared to 0.75 without LAKI, demonstrating the effectiveness of the LAKI strategy.

C. Results of Trajectory Shapley Attribution

Using the trained classifier $p(\boldsymbol{x})$ and setting $f(\boldsymbol{x}) \triangleq p(\boldsymbol{x})$, we computed TSA attribution scores for each wafer outside the training dataset with Eq. (9):

$$s_i(\boldsymbol{x}^t) = p(\boldsymbol{x}_{1:i}^t, \boldsymbol{x}_{(i+1):D}^0) - p(\boldsymbol{x}_{1:(i-1)}^t, \boldsymbol{x}_{i:D}^0). \qquad (10)$$

An effective way to assess the *relative* importance of each measurement item is to visualize the scores using a cumulative attribution plot. Figure 3 shows two examples of such cross-process TSA visualizations. By leveraging the sum rule (7), we can visualize how 'badness' accumulates as a wafer progresses through the processing route.
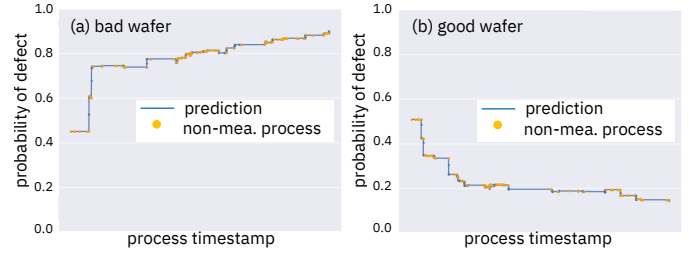

Fig. 3. Cumulative attribution score, $\beta(\tau, \boldsymbol{x}^t)$, defined in Eq. (11), plotted against process timestamp τ. $\boldsymbol{x}^t$ is the vector of inline measurements of the specified wafer. The plot shows how the baseline prediction $p(\boldsymbol{x}^0)$ evolves into the final prediction $p(\boldsymbol{x}^t)$ as time progress.

Specifically, we plot the *cumulative attribution score* up to time τ, defined as:

$$\beta(\tau, \boldsymbol{x}^t) \triangleq v(\emptyset) + \sum_{k=1}^{D} s_k(\boldsymbol{x}^t)\mathbb{I}(t_k \leq \tau), \qquad (11)$$

where t_k is the timestamp of the k-th measurement item and $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the condition in the parenthesis is satisfied, and 0 otherwise. For the specific expression in Eq. (10), the sum rule leads to

$$\beta(0, \boldsymbol{x}^t) = p(\boldsymbol{x}^0), \quad \beta(\infty, \boldsymbol{x}^t) = p(\boldsymbol{x}^t), \qquad (12)$$

assuming that the process starts after $\tau = 0$. These equations determine the start and end points of the cumulative attribution plot in Fig. 3.

Notably, certain time points exhibit significant jumps, suggesting potential root causes for favorable or unfavorable outcomes. In particular, Figure 3(a) indicates that our TSA approach could have facilitated earlier process termination for defective wafers, potentially months before reaching the PLY measurement point.

VII. Conclusion

We have proposed the Trajectory Shapley Attribution (TSA) method as a model-agnostic and interpretable framework for root-cause analysis in sequential manufacturing processes.

The proposed method addresses two major limitations of standard Shapley value (SV)-based attribution: reliance on unrealistic process routes and sensitivity to an arbitrary baseline point. TSA resolves these issues by introducing two novel components—lot-aware kernel imputation and trajectory Shapley values. We mathematically demonstrated that the trajectory Shapley value reduces to a closed-form expression without combinatorial complexity when the sum rule is imposed.

We demonstrated the practical utility of TSA through a case study in a good-bad wafer classification task, using real-world data from the Albany NanoTech fab.

While TSA was developed as a variant of the Shapley value, other algorithmic approaches may also be effective for attribution in semiconductor manufacturing. Exploring alternative methods, such as partial trajectory regression [25], and establishing a unified framework for defect root-cause analysis represent promising directions for future research.

References

[1] G. A. Susto, S. Pampuri, A. Schirru, A. Beghi, and G. De Nicolao, "Multi-step virtual metrology for semiconductor manufacturing: A multilevel and regularization methods-based approach," *Computers & Operations Research*, vol. 53, pp. 328–337, 2015.

[2] M. A. Jebri, E. El Adel, G. Graton, M. Ouladsine, and J. Pinaton, "Virtual metrology on semiconductor manufacturing based on just-in-time learning," *IFAC-PapersOnLine*, vol. 49, no. 12, pp. 89–94, 2016.

[3] K.-J. Kim, K.-J. Kim, C.-H. Jun, I.-G. Chong, and G.-Y. Song, "Variable selection under missing values and unlabeled data in semiconductor processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 32, no. 1, pp. 121–128, 2018.

[4] J. Yella, C. Zhang, S. Petrov, Y. Huang, X. Qian, A. A. Minai, and S. Bom, "Soft-sensing conformer: A curriculum learning-based convolutional transformer," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 1990–1998, IEEE, 2021.

[5] S. Han, et al., "Deep learning-based virtual metrology in multivariate time series," in *2023 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pp. 30–37, IEEE, 2023.

[6] F. Dalla Zuanna, N. Gentner, and G. A. Susto, "Deep learning-based sequence modeling for advanced process control in semiconductor manufacturing," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 8744–8751, 2023.

[7] K. B. Lee and C. O. Kim, "Recurrent feature-incorporated convolutional neural network for virtual metrology of the chemical mechanical planarization process," *Journal of Intelligent Manufacturing*, vol. 31, no. 1, pp. 73–86, 2020.

[8] C.-Y. Hsu and Y.-W. Lu, "Virtual metrology of material removal rate using a one-dimensional convolutional neural network-based bidirectional long short-term memory network with attention," *Computers & Industrial Engineering*, vol. 186, p. 109701, 2023.

[9] A. E. Roth, *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[10] E. Strumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," *The Journal of Machine Learning Research*, vol. 11, pp. 1–18, 2010.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 4768–4777, 2017.

[12] J. A. Torres, I. Kissiov, M. Essam, C. Hartig, R. Gardner, K. Jantzen, S. Schueler, and M. Niehoff, "Machine learning assisted new product setup," in *2020 31st Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, pp. 1–5, IEEE, 2020.

[13] J. Senoner, T. Netland, and S. Feuerriegel, "Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing," *Management Science*, vol. 68, no. 8, pp. 5704–5723, 2022.

[14] 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.

[15] P. Guo and Y. Chen, "Enhanced yield prediction in semiconductor manufacturing: Innovative strategies for imbalanced sample management and root cause analysis," in *2024 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*, pp. 1–6, IEEE, 2024.

[16] V. Maitra, Y. Su, and J. Shi, "Virtual metrology in semiconductor manufacturing: Current status and future prospects," *Expert Systems with Applications*, p. 123559, 2024.

[17] H.-W. Xu, Q.-H. Zhang, Y.-N. Sun, Q.-L. Chen, W. Qin, Y.-L. Lv, and J. Zhang, "A fast ramp-up framework for wafer yield improvement in semiconductor manufacturing systems," *Journal of Manufacturing Systems*, vol. 76, pp. 222–233, 2024.

[18] C. Molnar, *Interpretable machine learning*. Lulu.com, 2020.

[19] J. Labaien, T. Idé, P.-Y. Chen, E. Zugasti, and X. De Carlos, "Diagnostic spatio-temporal transformer with faithful encoding," *Knowledge-Based Systems*, vol. 274, p. 110639, 2023.

[20] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[21] H. Zheng, L. Shen, A. Tang, Y. Luo, H. Hu, B. Du, Y. Wen, and D. Tao, "Learning from models beyond fine-tuning," *Nature Machine Intelligence*, pp. 1–12, 2025.

[22] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *International conference on machine learning*, pp. 9269–9278, PMLR, 2020.

[23] E. Borgonovo, E. Plischke, and G. Rabitti, "The many shapley values for explainable artificial intelligence: A sensitivity analysis perspective," *European Journal of Operational Research*, 2024.

[24] T. Idé and N. Abe, "Generative perturbation analysis for probabilistic black-box anomaly attribution," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 845–856, 2023.

[25] K. Miyaguchi, M. Joko, R. Sheraw, and T. Idé, "Wafer defect root cause analysis with partial trajectory regression," in *2025 SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, p. TBD, IEEE, 2025.