

## **TOWARDS PROCESS OPTIMIZATION BY LEVERAGING RELATIONSHIPS BETWEEN ELECTRICAL WAFER SORTING AND COMPLETE-LINE STATISTICAL PROCESS CONTROL DATA**

Dmitrii Fomin<sup>1,3</sup>, Anastasiia Doynychko<sup>1</sup>, Andres Torres<sup>2</sup>, Valeria Borodin<sup>3</sup>, David Lemoine<sup>3</sup>, Agnès Roussy<sup>4</sup>, Daniele Pagano<sup>5</sup>, Marco Stefano Scropo<sup>5</sup>, Gabriele Tochino<sup>5</sup>, and Daniele Vinciguerra<sup>5</sup>

<sup>1</sup>Siemens, Electronic Design Automation, Grenoble, FRANCE

<sup>2</sup>Siemens, Electronic Design Automation, Wilsonville, USA

<sup>3</sup>IMT Atlantique, LS2N-CNRS, Nantes, FRANCE

<sup>4</sup>Mines Saint-Etienne, CNRS, UMR 6158 LIMOS, Gardanne, FRANCE

<sup>5</sup>STMicroelectronics, Catania, ITALY

### **ABSTRACT**

In semiconductor manufacturing, Statistical Process Control (SPC) ensures that products meet the Electrical Wafer Sort (EWS) tests performed at the end of the manufacturing flow. In this work, we model the EWS tests for several products using inline SPC data from the Front-End-Of-Line (FEOL) to the Back-End-Of-Line (BEOL). SPC data tend to be inherently sparse because measuring all wafers, lots, and products is both costly and can significantly impact the throughput. In contrast, EWS data is densely collected at the die level, offering high granularity. We propose to model the problem as a regression task to uncover interdependencies between SPC and EWS data at the lot level. By applying two learning strategies, mono- and multi-target, we demonstrate empirically that leveraging families of EWS tests enhances model performance. The performance and practical relevance of the approach are validated through numerical experiments on real-world industrial data.

### **1 CONTEXT, MOTIVATIONS, AND RELATED BACKGROUND**

Semiconductor manufacturing can be broadly divided into front-end processing (design and wafer fabrication) and back-end processing (assembly, packaging, and testing of chips). Modern wafer fabrication involves hundreds or even more than one thousand steps, many of which can be performed by multiple tools. The number and complexity of the product routes are increasing significantly with the advancement of semiconductor nodes. To prevent yield losses, these processes are statistically monitored and controlled through advanced process control frameworks (Moyné and Iskandar 2017). As illustrated in Figure 1, throughout the manufacturing flow, SPC techniques are employed to monitor and maintain process stability. Inline metrology tools measure critical process parameters and detect defects at various steps, while EWS performs electrical testing of chips on the wafer prior to packaging to guarantee the satisfaction of customer specifications, alongside other effectiveness-related evaluations (e.g., assembly and associated defectivity tests, final test) (Furnari et al. 2021). More specifically, EWS, also known as *probing*, consists of hundreds or more than a thousand electrical conformance tests, including short-circuit tests, leakage current detection, and parasitic capacitance measurements (Sarpietro et al. 2022; Rundo et al. 2023).

Despite tight process control, any excursion in front-end tools can still result in yield loss, caused by a variety of issues (Di Palma et al. 2005). Effectively linking electrical wafer sort data with process steps via associated parameters can contribute to yield enhancement and rapid root-cause identification (Di Palma et al. 2007). Visual analysis at the EWS stage is one of the most used approaches to characterize wafer defects and to detect early warning signs of issues originating in upstream manufacturing stages (Sarpietro et al. 2022; Rundo et al. 2024). Existing research in this area has primarily focused on the pattern recognition,

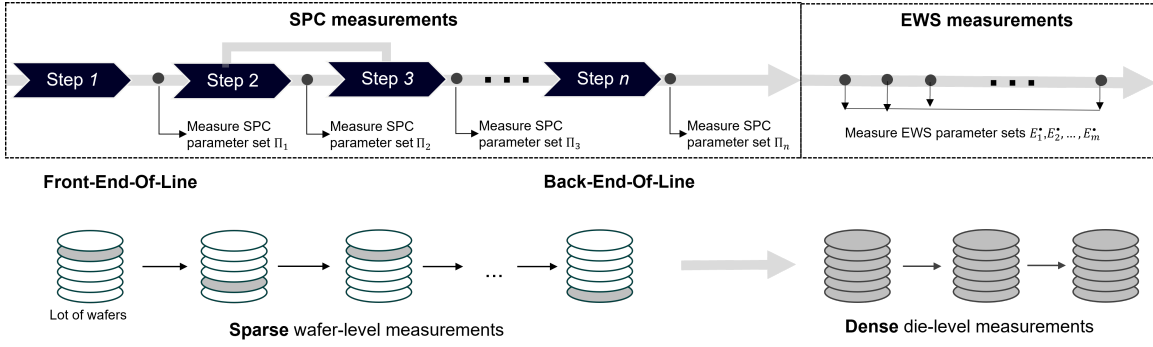


Figure 1: Statistical Process Control (SPC) and Electrical Wafer Sorting (EWS) measurement events.

semantic segmentation, and classification of EWS defect maps or defect patterns using approaches such as supervised learning approaches (Kim et al. 2020; Sarpietro et al. 2022), unsupervised deep learning approaches (Di Palma et al. 2005; Park et al. 2021), and hybrid-learning-based approaches (Yu et al. 2019). However, there is still limited research on leveraging EWS data to proactively support process optimization across the entire manufacturing flow, i.e., from FEOL to BEOL.

The early identification of process steps responsible for electrical drifts, which are monitored by SPC, is critical for enhancing yield and accelerating root-cause analysis. Motivated by that, this paper investigates the following research questions: *Are there any relationships between EWS and SPC data? If yes, how and to what extent can these relationships be leveraged to optimize semiconductor manufacturing processes?* The main contributions of this work are as follows:

- Modeling the full production flow, from FEOL to BEOL, using SPC and aligning it with EWS data;
- Instead of relying on defect pattern-matching EWS maps to identify root causes, as done in the related literature, we apply a regression-based modeling approach to uncover interdependencies between SPC and EWS data at the lot level per product and without cross-product learning;
- Demonstrating how leveraging families of electrical wafer sorting tests during wafer sorting can enhance model performance;
- Conducting numerical experiments on real-world data and empirically evaluating the proposed approach;
- Providing a comprehensive overview of the manufacturing processes by connecting data from early stages (FEOL, tracked by SPC) to a later stage (BEOL, measured via EWS). This integrated view supports decision-makers in identifying process steps that may contribute to electrical performance issues, thereby guiding targeted process optimization.

The remainder of this paper is structured as follows. Section 2 defines the problem under study Section 3 presents the proposed modeling framework. Section 4 discusses the results of the numerical experiments conducted on real-world industrial data. Finally, Section 5 concludes the paper and outlines potential directions for future research.

## 2 PROBLEM STATEMENT

Let  $P_\bullet$  represent a given product. Each product follows a specific production route  $J_\bullet = \{o_1^\bullet, o_2^\bullet, \dots, o_n^\bullet\}$ , i.e., a fixed sequence of  $n$  process steps (i.e., manufacturing operations). The production sequence is illustrated in Figure 1. SPC data is collected at every process step in the product route  $P_\bullet$ .

After each operation  $o_i^\bullet$  in the production route, one of the qualified machines measures a lot-representative subset of randomly selected dies from one or several wafers in selected lots to collect SPC parameters. These lot-level SPC parameters, denoted as  $\Pi_i, \forall i \in \{1, 2, \dots, n\}$ , are obtained based on a predefined set of recipes  $R_j^i$  specific to each parameter  $\pi_j^i \in \Pi_i$ . To provide the conditions under which

a wafer is processed, each SPC entry is uniquely defined by the following triplet:

$$\langle o_i^\bullet, \pi_j^i, \rho_{k,j}^i \rangle$$

where:

- $o_i^\bullet \in J_\bullet$  represents the process step,
- $\pi_j^i \in \Pi_i$  denotes the measured parameter associated with process step  $o_i^\bullet$ , and
- $\rho_{k,j}^i \in R_j^i$  specifies the measurement recipe applied to parameter  $\pi_j^i$ .

With a higher number of process steps per product, SPC entries grow in a combinatorial way, leading to the so-called problem of the curse of dimensionality (Susto et al. 2015; Korabi et al. 2021).

Downstream of the manufacturing line, the electrical performance of every die (i.e., device) is measured in terms of  $m$  parameter families  $E_1^\bullet, E_2^\bullet, \dots, E_m^\bullet$ . In contrast to SPC measurements, EWS data are dense, providing comprehensive die-level information across the entire set of manufactured wafers (see Figure 1). To optimize the semiconductor manufacturing processes, this paper explores the relationships between EWS and complete-line SPC data based on a regression-based modeling approach. In what follows, parameters expressing electrical performance are referred to as *targets*, i.e., the variables of interest to be explained via SPC data.

### 3 MODELING EWS DATA VIA SPC DATA: REGRESSION MODELING APPROACH

Considering a given product, let us assume that there are interdependencies between the SPC and EWS data. To identify them, we propose two regression modeling strategies formalized in Algorithm 2 and Algorithm 3. The proposed approach includes two main stages:

1. *SPC and EWS data alignment at the lot level* (see Algorithm 1),
2. *Model fitting*: Two regression modeling strategies are applied in this paper, namely (i) **mono-target**, denoted by model\_mono (see Algorithm 2), and (ii) **multi-target**, denoted by model\_multi (see Algorithm 3).

**Algorithm 1** : Modeling SPC data via EWS data.

	Notation	Description
<b>Input:</b>	$\ell$	Number of lots
	$J_\bullet$	Product route
	$\Pi_i$	Sets of parameter associated with process steps $o_i^\bullet \in J_\bullet$
	$R_j^i$	Measurement recipe set applied to parameter $\pi_j^i \in \Pi_i, \forall i \in \{1, 2, \dots, n\}, \forall j \in \{1, 2, \dots,  \Pi_i \}$
	$E_1^\bullet, E_2^\bullet, \dots, E_m^\bullet$	EWS target families of product $P_\bullet$
	$s$	Number of SPC entries defined by $\langle o_i^\bullet, \pi_j^i, \rho_{k,j}^i \rangle, \forall o_i^\bullet \in J_\bullet, \forall \pi_j^i \in \Pi_i, \forall \rho_{k,j}^i \in R_j^i$
	$p$	Number of relevant tester-related parameters
	$t$	Total number of EWS targets $t = \sum_{i=1}^m  E_i^\bullet $

- 1: **Aggregate EWS targets** at the lot level  $y^{\ell \times t}$
- 2: **Extract matrix of SPC features**  $F^{\ell \times s}$
- 3: **Extract EWS tester-related categorical features**  $C^{\ell \times p}$
- 4: **Consolidate the set of baseline features**  $X^{\ell \times (s+p)} = [F^{\ell \times s} \ C^{\ell \times p}]$ , by concatenating SPC and EWS tester-related features

In addition to the mono-target baseline strategy, the multi-target strategy is designed to predict multiple targets concurrently to enable the revealing of latent relationships across EWS targets and SPC features during model training. This is achieved by identifying a partition of the EWS target set into nonempty

subsets that maximize the predictive performance of the regression model. It is important to note that the underlying set partitioning step is NP-hard (Balas and Padberg 1976). In this study, we address this combinatorial optimization challenge using a greedy heuristic that exploits target similarity. We assume that targets within the same electrical test family exhibit inherent similarity, and group them by EWS parameter family  $E_j^\bullet$ . In the case of multi-modality in the distribution of targets, the family is split per mode.

---

**Algorithm 2** : Mono-target regression modeling.

---

1: **Fit  $t$  regression models:**  $\bar{y}_k^{\ell \times 1} = f(X^{\ell \times (s+p)}), \forall k = \{1, 2, \dots, t\}$

---

In this work, the greedy heuristic applied in **Step 1** of Algorithm 3 clusters targets as follows:

1. Group EWS targets by electrical test families  $E_1^\bullet, E_2^\bullet, \dots, E_m^\bullet$ ;
2. Split any  $E_i^\bullet$  where the associated data distribution exhibits multimodality,  $\forall i \in \{1, 2, \dots, m\}$ .

By design, the proposed greedy approach is intended to support cross-target learning within each electrical test family. Further generalizations can be explored and are left for future research.

---

**Algorithm 3** : Multi-target regression modeling.

---

- 1: **Cluster targets** into  $g$  groups  $G_q, q = \{1, 2, \dots, g\}$
- 2: **Fit  $g$  regression models**
- 3: Let  $T^{(\ell \times u) \times 1}$  be categorical features related to test-specific characteristics

$$\bar{y}_q^{(\ell \times u) \times 1} = f([X^{(\ell \times u) \times s} \ T^{(\ell \times u) \times 1}]),$$

where  $u = \sum_{j=1}^g |G_q|, q = \{1, 2, \dots, g\}$

---

The proposed approach confronts and addresses several intrinsic complexities of semiconductor manufacturing, as follows:

- *Redundant machines:* SPC and EWS parameters of interest can potentially be processed/measured by different qualified metrology or tester machines. To address this, the EWS tester-related context information has been explicitly considered in this paper.
- *Low-Volume:* Little historical data may be compatible and applicable to support approaches dedicated to process optimization. In response to low-volume regimes, a multi-target strategy is proposed.
- *Data characteristics:* Operating in sensor-rich environments, semiconductor facilities generate vast amounts of data at every stage of manufacturing. However, despite this abundance, applying machine learning in the semiconductor manufacturing domain remains highly challenging due to multiple inherent data characteristics (Migueláñez et al. 2005; Clain et al. 2021), including: **(i)** Data fragmentation, high-dimensionality: To confront this, high-dimensional SPC data with EWS outputs are aggregated at the lot level **(ii)** Non-Gaussian distributions and multi-modal data, and temporal dependencies: These distribution characteristics are explicitly handled by the design of the Algorithm 3 and EWS categorical features.

## 4 COMPUTATIONAL EXPERIMENTS

This section presents an empirical evaluation of the proposed modeling approach under both mono-target and multi-target strategies. The industrial relevance of the approach is assessed through numerical experiments conducted on representative industrial data, as described in Section 4.1. The experimental settings and design are detailed in Section 4.2. A comparative performance analysis of the two strategies, model\_mono

and `model_multi`, is provided in Section 4.3. Finally, Section 4.4 discusses the industrial implications of the identified relationships between EWS and SPC data.

#### 4.1 Dataset Description

Numerical experiments have been carried out on three products  $\{P_1, P_2, P_3\}$  and described in Table 1. The SPC data includes more than 8,708,300 observations and has been collected over two years. Given the severe sparsity of the SPC data, columns with more than 60% missing values have been removed. As a result, data related to over 50% (80% in the case of product  $P_3$ ) of the process steps in the product routes (see the fifth column in Table 1) have been excluded. Regarding the EWS data, numerical experiments have been conducted on three products, each associated with tens to hundreds of EWS targets, organized into between 6 and 36 families.

Table 1: Description of the considered dataset.

Product	# EWS targets	# EWS groups	# Lots	# Kept (# Total)	process steps	# Baseline features
$P_1$	25	6	490		42 (306)	228
$P_3$	70	16	135		119 (306)	1,200
$P_2$	223	36	164		149 (306)	1,444

#### 4.2 Experiments: Settings and Design of Experiments

Numerical experiments were conducted in Python using the CatBoost library<sup>1</sup>, an advanced gradient boosting framework optimized for handling categorical features (Prokhorenkova et al. 2018). CatBoost employs *ordered boosting* to prevent target leakage and *ordered target statistics* to encode categorical variables through unbiased estimates of conditional target probabilities. These mechanisms mitigate prediction shift and substantially reduce overfitting.

To train and validate the regression models in Algorithm 1, a 10-fold cross-validation approach has been used to ensure robust performance evaluation. To do this, data are divided into ten equal-sized parts, referred to as *folds*. Training and evaluating models across these ten different data splits allow us to capture any variations in the model performance that may arise due to differences in training and test data distribution. This is particularly important when data distribution affects model quality, as each split can expose the model to different characteristics of the target variable. At the end of the cross-validation process, the models are compared, and the best one is selected.

#### 4.3 Analysis of the Performance of the Proposed Approach

Let us analyze the performance of the proposed approach under both mono-target and multi-target strategies. For validation, we use multiple evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), and the coefficient of determination ( $R^2$ ). Most of these metrics are presented in Figure 4. Among these metrics used for validation, we focus mainly on the  $R^2$  and MAPE, due to their consistent behavior relative to others across experiments. While  $R^2$  focuses on capturing general patterns by measuring how well the model explains the variability in the actual data, MAPE quantifies the average deviation of the predictions as a percentage, providing an intuitive measure of precision. Furthermore,  $R^2$  and MAPE are often easier to interpret, as they are scale-free and unit-independent.

Figure 2 compares the performance of `model_mono` and `model_multi` strategies in terms of  $R^2$  and MAPE. The X-axis (resp., Y-axis) represents the performance of the mono-target (resp., multi-target) strategy for a given evaluation metric. The diagonal line indicates parity between models for the selected metric. Figure 2 reveals that the multi-target model generally demonstrates a slight advantage in accuracy,

<sup>1</sup><https://catboost.ai>

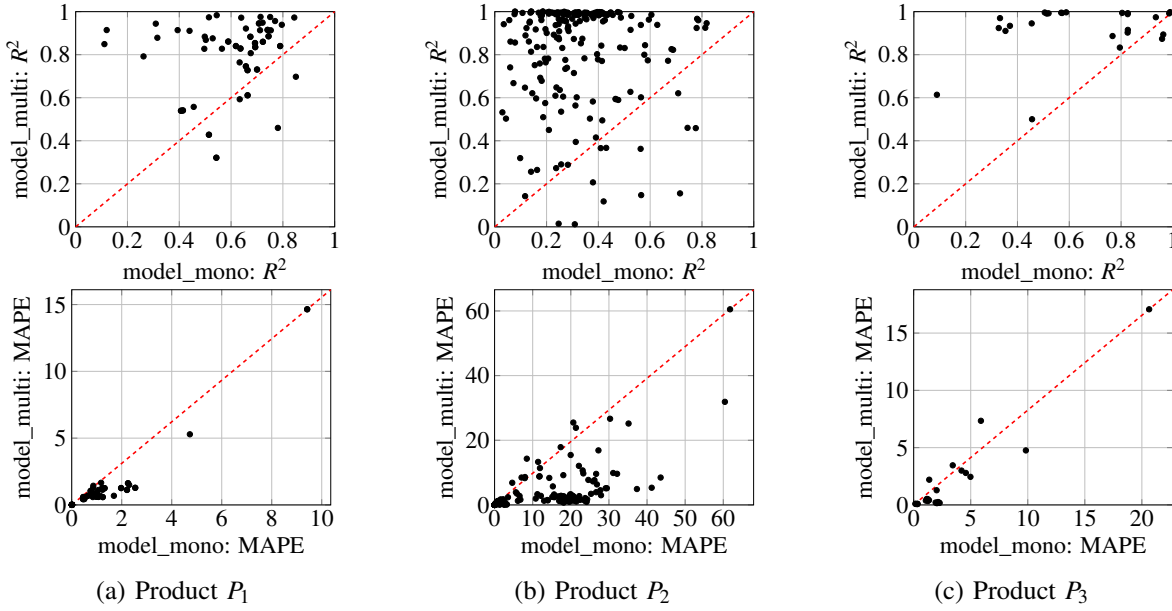


Figure 2: Performance of mono-target *versus* multi-target models:  $R^2$  and MAPE.

as measured by MAPE. While this difference may seem insufficient to justify adopting the multi-target approach, a closer look at the  $R^2$  plots reveals a more pronounced improvement in favor of the multi-target model. Several factors contribute to this performance behavior: (i) The ability of the multi-target to learn from multiple targets simultaneously enables it to capture interactions between targets and better understand the underlying processes. This results in a more comprehensive representation of the data and, consequently, a better ability to generalize. (ii) In some cases, the distribution of a target may cause the associated mono-target model to converge toward predicting average values. While this strategy can yield reasonable predictive accuracy in terms of MAPE, it often reflects a limited capacity to explain the data and uncover the underlying patterns, as indicated by lower  $R^2$  scores. The multi-target model avoids this pitfall by leveraging information from the joint distribution of multiple targets, enhancing both its explanatory power and predictive performance. At the same time, it is worth mentioning that the multi-target model can still benefit from improving the quality of the EWS target grouping.

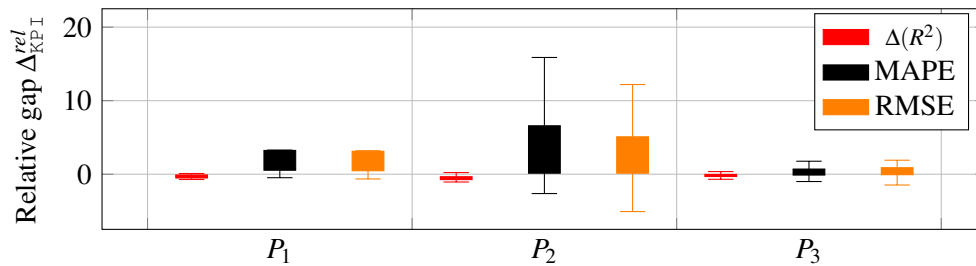


Figure 3: Performance of mono-target model *versus* multi-target model:  $R^2$ , MAPE, and RMSE.

To measure how close a strategy is to the best one in relative terms, the performance of strategies has also been compared in terms of relative gaps, as follows:

$$\Delta_{KPI}^{rel} = \frac{KPI(model\_mono) - KPI(model\_multi)}{\text{best}\{KPI(model\_mono), KPI(model\_multi)\}}$$

A consistent trend emerges across different products and EWS targets, as illustrated in Figure 3. The negative (resp., positive) part of boxplots represents that the multi-target model outperforms the mono-target one for  $R^2$  (resp., MAPE, RMSE). Despite minor differences in precision, the multi-target model consistently demonstrates superior explanatory power. It reveals patterns that provide deeper insight into the relationships between features and the physical phenomena driving the manufacturing processes. This makes it not only a robust predictive tool but also a valuable asset for understanding and interpreting the system's inner mechanisms.

Let us examine the difference in the ability of models to predict data for the training and test sets, illustrated in Figure 4a and Figure 4b. In the training set, the performance of both models appears similar, indicating that they are equally capable of learning the underlying patterns. However, when extending the analysis to the test set, differences in their prediction capacity become apparent. While both models maintain high predictive accuracy as measured by the MAPE, a closer inspection of the residuals indicates that the mono-target model tends to regress toward the mean. This behavior is reflected in a lower  $R^2$  score, indicating that, while the model is effective in reducing overall error, it does not capture the full variance of the data as well as the multi-target model. Such a behavior to predict average values suggests that the mono-target model may be less responsive to the nuances present in the test data, potentially resulting in underfitting for complex patterns.

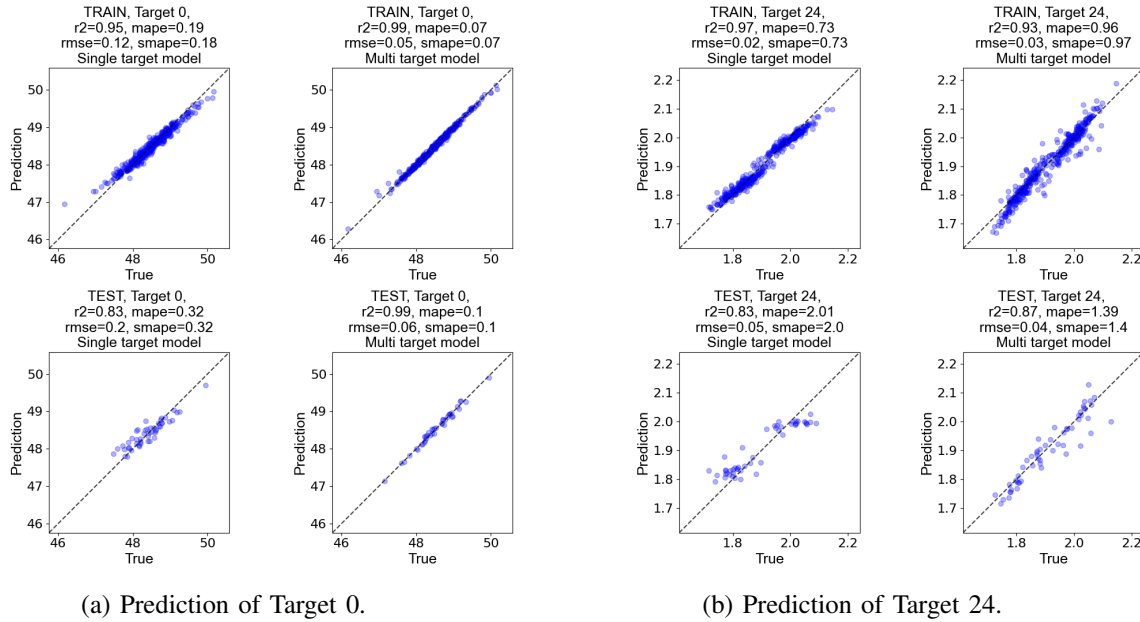


Figure 4: Prediction results for both mono- and multi-target models for train and test sets.

Moreover, the differences in performance between the two models become more significant when considering the practical implications of these metrics. A high MAPE value may indicate that the model performs well in terms of relative error. However, the lower  $R^2$  of the mono-target model implies that it fails to account for the variability inherent in the data. This could be critical in applications where understanding the spread and distribution of predictions is essential. This discrepancy underscores the importance of employing multiple complementary evaluation metrics when assessing model performance and highlights the potential benefits of using a multi-target approach in scenarios where data complexity is high.

#### 4.4 Industrial Implications

Until now, the focus has been on applying predictive models across the entire manufacturing process flow, from FEOL to BEOL, to evaluate the relationships between SPC and EWS data quantitatively. In what follows, let us discuss how the proposed approach can support decision-makers in identifying the process steps responsible for electrical drifts.



To analyze the relationship between SPC and EWS stages, we use SHapley Additive exPlanations (SHAP) values to interpret the model's predictions (Lundberg and Lee 2017). This approach quantifies the contribution of each input feature for every sample to the predicted EWS values. Given sufficient predictive performance, SHAP offers a robust and interpretable way to reveal the interdependencies between SPC parameters and EWS outcomes.

SHAP beeswarm plots in Figure 5 and Figure 6 illustrate the impact of each feature on the output of the models across all instances. Each point represents a single instance (i.e., row in the dataset) for a given feature. Features are sorted on the Y-axis by their overall importance. The X-axis displays SHAP values, indicating whether the feature increased (positive) or decreased (negative) the prediction. The color of each dot reflects the actual feature value (blue for low, red for high). A wide horizontal spread means the feature strongly affects predictions. The joint analysis of color and SHAP value reveals how the magnitude of a feature relates to its effect on predictions.

Figure 5 and Figure 6 illustrate the ten most important features for both models concerning the two targets discussed previously. Notably, the most important features include both SPC-related variables and EWS target-specific features, highlighting the key factors influencing the testing process.

Let us focus on the features presented in Figure 5. Some features correspond to the same process step but involve different measured parameters or even the same parameter measured under different conditions. For the mono-target model, one of the most influential features belongs to process step process31, along with parameters 108 and 308. These parameters appear across different recipes and process steps, with process31 recurring twice as a critical procedure. A similar pattern is observed in the multi-target model, where the most important process steps are process31 and process0. The fact that process31 is important for both models, despite their structural differences, further reinforces its significance. Comparing multiple models and identifying common influential features is essential for pinpointing important predictors for a given electrical target, and thus enhancing the model's explanation robustness.

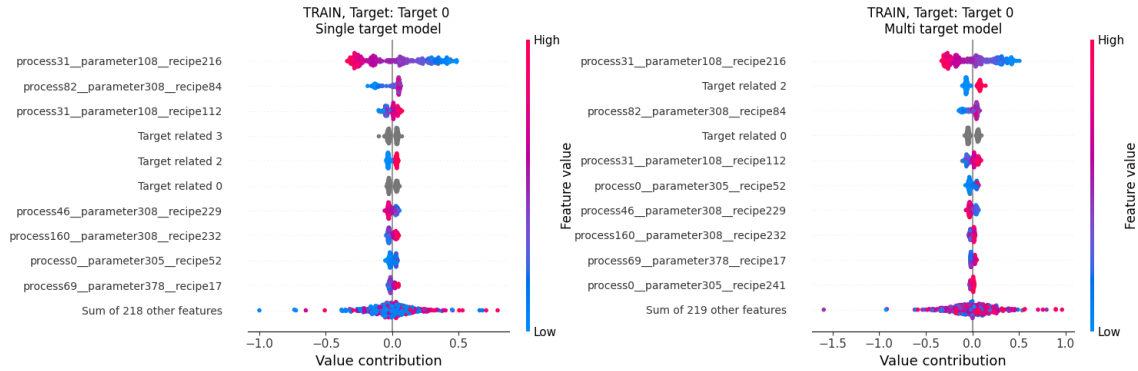


Figure 5: Shap values of Target 0 for mono-target and multi-target strategies.

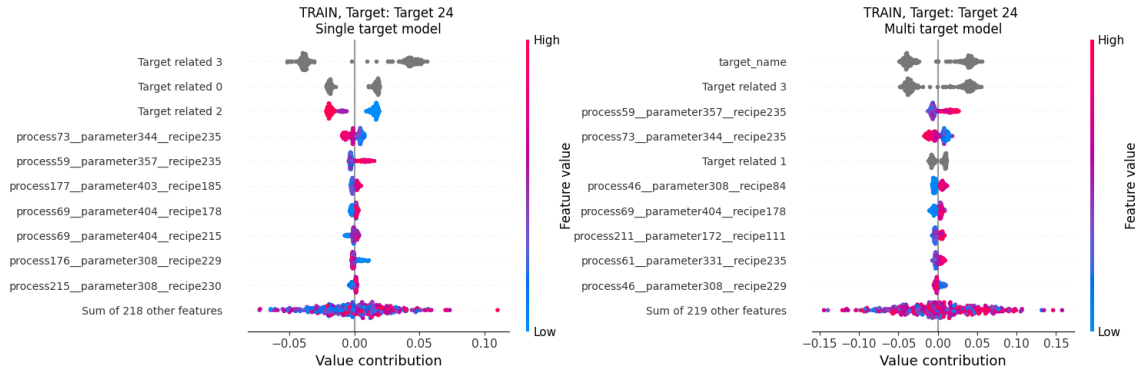


Figure 6: Shap values of Target 24 for mono\_target and multi\_target strategies.



Figure 6 reveals a significant influence of the categorical EWS-related features. A comparison of the SHAP values shows that these parameters largely determine the main characteristics of the two different distributions presented for the respective target. In addition to the categorical features related to test types, the target name is also observed. This feature is unique to the multi-target model and helps to group features and highlight their main differences, such as the average value of the distribution.

Categorical features are prioritized over SPC parameters in both strategies. During training, the associated model first establishes a general, rough representation of the distribution, with categorical features defining the number of modes and other characteristics. Subsequently, SPC data refines these predictions to provide more accurate results. Certain process steps are repeated with different parameters, emphasizing their importance. For instance, process73 is consistently among the most significant in both models, and parameters such as 404 and 308 appear multiple times across the models.

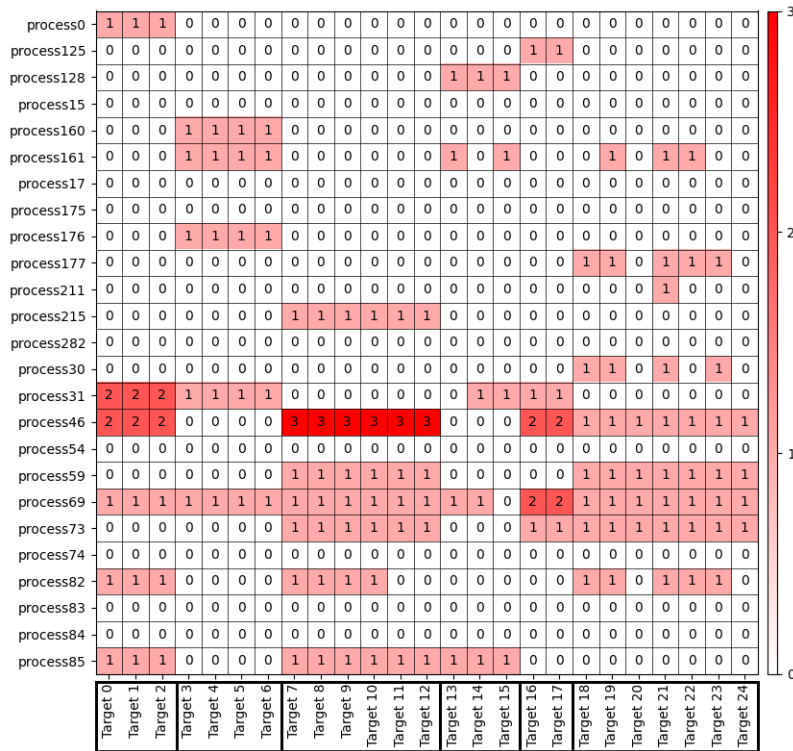


Figure 7: Process steps of interest and associated numbers of relevant SPC parameters to model EWS data for a given EWS family via multi-target strategy: Product  $P_1$ . Boxes in the X-axis represent EWS groups.

There are 218 remaining features for the mono-target model and 219 for the multi-target model, ranked by explanatory power. The process steps responsible for electrical drifts become apparent only when the models demonstrate sufficient accuracy.

The analysis of the top ten most important features is summarized in Figure 7 for the multi-target model. Figure 7 reports how often SCP features associated with a given process step appear among the top ten most important features for each EWS target. Let us start by examining how to identify the most important process steps for each group (i.e., family) of EWS targets. For the first group of targets, the most important processes (31, 46, 69, etc.) are easily identifiable because they are the same for each target in the group. This also indicates the similarity of the targets that constitute the group. However, this does not hold for all groups, as can be seen with the 4th and 6th groups.

Some process steps, such as 69 and 46, appear as important across multiple families. When a process step significantly impacts more than one group, it further emphasizes its critical role in the manufacturing process. The analysis of Figure 7 enables the identification of relevant SPC entries from among thousands of possibilities,

specifying both the SPC parameters and the process steps in which they occur. As a result, it gives the ability to optimize specific manufacturing processes, thereby improving both control and quality.

The results of the proposed regression-based modeling approach have been discussed with process engineers for one product and found industrially relevant. Further in this direction, Fault Detection and Classification (FDC) will be shared to enable wafer-level alignment with SPC data and will help us identify which process parameters need to be adjusted to achieve efficient quality production settings.

## 5 CONCLUSIONS AND PERSPECTIVES

This paper investigates the interdependencies between EWS and SPC data across the entire semiconductor manufacturing flow, from FEOL to BEOL, with the aim of optimizing production processes. To this end, a regression-based modeling approach is proposed, incorporating two strategies: mono-target and multi-target learning. Experimental results on industrial data confirm the existence of interdependencies between EWS and SPC measurements. Building on these findings, the paper discusses several industrial implications to support decision-makers in identifying the main process steps that contribute to electrical deviations.

Following our first promising findings, future research efforts will be dedicated to improving the proposed approach in several directions:

- *Partitioning of the set of EWS targets:* Target grouping is performed using a greedy heuristic that exploits the similarity of EWS parameters at the family level. However, it was observed that not all EWS measurements are compatible within the considered families using the current strategy, and in some cases, combining them can degrade model performance. We have not yet investigated whether targets with different characteristics could positively influence each other when modeled together. Future work will explore methods for identifying efficient EWS target groupings by analyzing the characteristics and interactions between targets.
- *Imputation:* Our findings indicate that denser data representations lead to a more reliable model. This, in turn, improves the consistency in identifying the most influential process steps for a given electrical test and product. We aim to include an imputation process by incorporating and modeling FDC and SPC data. This would enable wafer-level modeling and support product-specific process optimization.
- *Outlier removal:* Outliers can be broadly categorized into two types: (i) those that may result from measurement errors, and (ii) those that contain valuable information by reflecting meaningful departures from standard process control. The latter are informative outliers that can provide deeper insights into the manufacturing process. One of our perspectives is to distinguish between these two types of outliers, which, in contexts with small sample sizes and fragmented data, can enhance model quality. In addition, we aim to develop approaches that mitigate model degradation when outliers are identified as true, non-systematic errors.
- *Cross-product learning and transferability:* Each product has different electrical performance requirements, and these requirements are affected differently by each process step. This makes learning across multiple products challenging when there are no intersecting EWS characteristics. While our current study focuses on within-product learning, exploring cross-product generalization and transferability remains a promising direction for future work.
- *Validation of the relevance of the results in real-life settings:* We aim to extend the analysis to a larger set of products and share the results with relevant process owners to assess the validity and applicability of the proposed approach under real-life high-mix manufacturing conditions. Note that SPC data includes not only the products of interest but also other products that share the same process steps. This broader coverage is valuable for identifying and investigating real-world interactions between products (e.g., machine memory effects, cross-product contamination, or batch-related influences) that may impact electrical performance.
- *Development of decision-support dashboards:* The analysis conducted in this paper will be extended to provide decision-makers with informative dashboards, such as matrices of critical process steps, machines, or other context-related settings.

## 6 ACKNOWLEDGMENT

The work presented in this paper has received funding from the Chips Joint Undertaking (JU) under grant agreement No 101097296. The JU receives support from the European Union's Horizon EU research and innovation programs and Austria, Belgium, Denmark, Finland, France, Germany, Greece, Hungary, Israel, Italy, Netherlands, Romania, Sweden, Switzerland and Turkey.

## REFERENCES

- Balas, E., and M. W. Padberg. 1976. "Set Partitioning: A Survey". *SIAM Review* 18(4):710–760.
- Clain, R., V. Borodin, M. Juge, and A. Roussy. 2021. "Virtual Metrology for Semiconductor Manufacturing: Focus on Transfer Learning". In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, 1621–1626. Lyon, France, August 23<sup>rd</sup>–27<sup>th</sup>, 2021.
- Di Palma, F., G. De Nicolao, O. M. Donzelli, and G. Miraglia. 2005. "Unsupervised Algorithms for the Automatic Classification of EWS Maps: A Comparison". In *ISSM 2005, IEEE International Symposium on Semiconductor Manufacturing*, 253–256. San Jose, California, USA, September 13<sup>th</sup>–15<sup>th</sup>, 2005.
- Di Palma, F., G. De Nicolao, G. Miraglia, and O. Donzelli. 2007. "ACID: Automatic Sort-Map Classification for Interactive Process Diagnosis". *IEEE Design & Test of Computers* 24(4):352–361.
- Di Palma, F., G. De Nicolao, G. Miraglia, and O. M. Donzelli. 2005. "Process Diagnosis Via Electrical-Wafer-Sorting Maps Classification". In *Fifth IEEE International Conference on Data Mining (ICDM'05)*. Houston, Texas, USA, November 27<sup>th</sup>–30<sup>th</sup>, 2005.
- Furnari, G., F. Vattiato, D. Allegra, F. L. M. Milotta, A. Orofino, R. Rizzo, *et al.* 2021, Aug. "An Ensembled Anomaly Detector for Wafer Fault Detection". *Sensors* 21(16).
- Kim, Y., D. Cho, and J.-H. Lee. 2020. "Wafer Map Classifier Using Deep Learning for Detecting Out-of-Distribution Failure Patterns". In *2020 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*, 1–5. Singapore, July 20<sup>th</sup>–23<sup>rd</sup>, 2020.
- Korabi, T. E., V. Borodin, A. Roussy, and M. Juge. 2021. "A Hybrid Feature Selection Approach for Virtual Metrology: Application to CMP Process". In *2021 32nd Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 1–5. Milpitas, California, USA, May 10<sup>th</sup>–12<sup>th</sup>, 2021.
- Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions". *Advances in Neural Information Processing Systems* 30.
- Migueláñez, E., A. M. S. Zalazala, and P. Buxton. 2005. "Swarm Intelligence in Automated Electrical Wafer Sort Classification". In *2005 IEEE Congress on Evolutionary Computation, Vols 1–3, Proceedings*, IEEE Congress on Evolutionary Computation, 1597–1604. Edinburgh, UK, September 2<sup>nd</sup>–5<sup>th</sup>, 2005.
- Moyne, J., and J. Iskandar. 2017. "Big Data Analytics for Smart Manufacturing: Case Studies in Semiconductor Manufacturing". *Processes* 5(3):39.
- Park, S., J. Jang, and C. O. Kim. 2021. "Discriminative Feature Learning and Cluster-Based Defect Label Reconstruction for Reducing Uncertainty in Wafer Bin Map Labels". *Journal of Intelligent Manufacturing* 32:251–263.
- Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. 2018. "CatBoost: Unbiased Boosting With Categorical Features". *Advances in Neural Information Processing Systems* 31.
- Rundo, F., M. Calabretta, C. Pino, S. Coffa, A. Messina, C. Spampinato *et al.* 2023. "Deep Learning for Automatic Wafer Monitoring System". In *2023 Smart Systems Integration Conference and Exhibition (SSI)*, 1–6. Bruges, Belgium, March 28<sup>th</sup>–30<sup>th</sup>, 2023.
- Rundo, F., M. Calabretta, M. S. Rundo, G. Castagnolo, C. Pino, S. Battiato *et al.* 2024. "Intelligent Electrical Assessment of Silicon and Silicon Carbide Wafers for Power Applications in Automotive Field". In *2024 IEEE International Workshop on Metrology for Automotive (MetroAutomotive)*, 18–23. Bologna, Italy, June 26<sup>th</sup>–28<sup>th</sup>, 2024.
- Sarpietro, R. E., C. Pino, S. Coffa, A. Messina, S. Palazzo, S. Battiato, *et al.* 2022. "Explainable Deep Learning System for Advanced Silicon and Silicon Carbide Electrical Wafer Defect Map Assessment". *IEEE Access* 10:99102–99128.
- Susto, G. A., S. Pampuri, A. Schirru, A. Beghi, and G. De Nicolao. 2015. "Multi-Step Virtual Metrology for Semiconductor Manufacturing: A Multilevel and Regularization Methods-Based Approach". *Computers & Operations Research* 53:328–337.
- Yu, J., X. Zheng, and J. Liu. 2019. "Stacked Convolutional Sparse Denoising Auto-Encoder for Identification of Defect Patterns in Semiconductor Wafer Map". *Computers in Industry* 109:121–133.

## AUTHOR BIOGRAPHIES

**DMITRII FOMIN** is PhD student at IMT Atlantique, France, completing his thesis in industrial engineering, automation, and computer science through working at Siemens EDA in Grenoble, France. His research interests include predictive analytics,

data science, and optimization. His e-mail address is [dmitrii.fomin@siemens.com](mailto:dmitrii.fomin@siemens.com).

**ANASTASIIA DOINYCHKO** is a Software Engineer at Siemens EDA in Grenoble, France. She holds a Ph.D. in Mathematics and Computer Science from the Université Grenoble Alpes. Her professional interests include Machine Learning and Advanced Process Control. She is actively involved in applied research and innovation at the intersection of software engineering and manufacturing technologies. Her email address is [anastasiia.doinychko@siemens.com](mailto:anastasiia.doinychko@siemens.com).

**J. ANDRES TORRES** holds a B.S. in Chemical Engineering from the National Autonomous University of Mexico, a M.S. in Chemical Engineering from UW-Madison, and a PhD degree in Electrical Engineering from the Oregon Health and Science University. He has been investigating the interactions between manufacturing process and electronic design flows to exploit areas of design and process co-optimization that provide more predictable and manufacturable designs. He is currently a distinguished Engineer for Siemens EDA, focusing on applying machine learning techniques in semiconductor manufacturing processes to improve yield and productivity. His e-mail address is [andres.torres@siemens.com](mailto:andres.torres@siemens.com).

**VALERIA BORODIN** is Associate Professor at IMT Atlantique, France, where she is actively involved in academic-industrial collaborations and knowledge transfer initiatives. She received her Ph.D. in Optimization and Systems Safety from the University of Technology of Troyes, France. Her research focuses on quantitative operations management in manufacturing and logistics systems, including mathematical modeling, optimization, simulation, and data analytics. Her email address is [valeria.borodin@imt-atlantique.fr](mailto:valeria.borodin@imt-atlantique.fr).

**DAVID LEMOINE** is Full Professor at IMT Atlantique (Nantes, France) in the Automation, Production, and Computer Science department. He received a Ph.D. in Computer Science (2008) from Blaise Pascal University (France) and defended his habilitation thesis (2020) in Nantes University (France). His research focuses on decision-making integration in production planning (lot-sizing models), particularly by incorporating financial criteria and considering maintenance decisions to enhance the robustness of production plans. His e-mail address is [david.lemoine@imt-atlantique.fr](mailto:david.lemoine@imt-atlantique.fr).

**AGNÈS ROUSSY** is Professor at the CMP of Mines de Saint Etienne. She received the Ph.D degree from INP-Grenoble, France. She was a Postdoctoral Fellow at the University of Twente, NL. She has been Associate Professor and since 2020, she is Professor at Mines de Saint Etienne in France. She works in the field of Advanced Process Control applied to semiconductor manufacturing. Her e-mail address is [roussy@emse.fr](mailto:roussy@emse.fr).

**DANIELE PAGANO** is Funding Project Manager at STMicroelectronics. He has covered various positions and responsibility in Catania Wafer Fab Operations (Lithography, Dry Etching, APC & SPC, Epitaxy, Quality & Process Control), past experiences in collaborative projects like IMPROVE (2012), INTEGRATE (2015), MADEin4 (2022), SATURN (2023) and nowadays HiCONNECTS and IPCEI. He is author and co-author of several publications on journals and international conferences. His e-mail address is [daniele.pagano@st.com](mailto:daniele.pagano@st.com).

**MARCO STEFANO SCROPPA** received the B.S. degree in computer engineering from the University of Catania, Italy, in 2013, and the M.S. degree in computer engineering from the University of Catania, Italy, in 2015, where he is currently pursuing the Ph.D. degree in systems, energy, computer and telecommunications engineering. His research interest focuses on the enhancement of interoperability in Industry 4.0, IoT, and IIoT. The research aims to study and define interoperability solutions that make smooth integration of IIoT applications and increasing security at the same time. His e-mail address is [marco-stefano.scroppo@st.com](mailto:marco-stefano.scroppo@st.com).

**GABRIELE TOCHINO** has been an IT Operations Engineer at STMicroelectronics since 2016. He earned his B.S. in Computer Science from the University of Catania, Italy, in 2012, and his M.S. degree in Computer Science from the same university in 2016. His professional interests include IT Manufacturing Site Solutions and user technical support, with a particular focus on Engineering Data Analysis and Fault Detection and Classification fields. His email address is [gabriele.tochino@st.com](mailto:gabriele.tochino@st.com).

**DANIELE VINCIGUERRA** works at STMicroelectronics, focusing on advanced process control, metrology, and data analytics within semiconductor manufacturing. His professional interests span process control, operations management, and yield optimization. His email address is [daniele.vinciguerra@st.com](mailto:daniele.vinciguerra@st.com).