Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P0884-cd

SHAP Analysis for Diagnosing Anomalies in Semiconductor Manufacturing

Joaquín Figueroa

Department of Energy, Politecnico di Milano, Italy. E-mail: joaquineduardo.figueroa@polimi.it

Ibrahim Ahmed

Department of Energy, Politecnico di Milano, Italy. E-mail: ibrahim.ahmed@polimi.it

Piero Baraldi

Department of Energy, Politecnico di Milano, Italy. E-mail: piero.baraldi@polimi.it

Enrico Zio

Department of Energy, Politecnico di Milano, Italy. E-mail: enrico.zio@polimi.it
MINES Paris-PSL, Centre de Recherche sur les Risques et les Crises (CRC), Sophia Antipolis, France. E-mail: enrico.zio@minesparis.psl.eu

Abstract: We consider the problem of predicting the quality of semiconductor devices and, in case of low quality, diagnosing the anomaly occurred during production. A multi-branch neural network is developed for quality prediction based on multimodal data. Specifically, a dedicated autoencoder is trained for each data modality; then, the latent representations provided by the encoders are concatenated and a regression layer is added for quality prediction. Shapley Additive exPlanation (SHAP) is used to quantify the contribution of each data modality to the quality outcome. Since different data modalities contain information about different production stages, the causes of the production anomaly can be identified. The developed method is demonstrated using a synthetic case study, which mimics the complexity of semiconductor manufacturing. Wafer map (images) and signal measurements (time series) from a production machine are the two considered data modalities. The method is shown able to effectively predict the quality of semiconductor devices and diagnose anomalies occurred at different stages of production.

Keywords: Semiconductor manufacturing, quality, multimodality data, multi-branch neural network, SHAP.

1. Introduction

Semiconductor manufacturing must meet high levels of quality to satisfy industry standards (May & Spanos, 2006). To do this, burn-in (BI) testing puts devices under stress conditions to identify early life failures which decrease the quality of the lot. Thus, discarding failed devices increases lot quality.

Recent advancements in sensing and data acquisition have made available large amounts of heterogeneous information at various stages of the manufacturing production pipeline, which can be used for developing data-driven quality control approaches. However, effectively leveraging these multisource and multimodality data is a challenging task (Chen, 2022).

Traditional quality control methods rely on statistical approaches (Dahari et al., 2025; Ooi et al., 2007). More recent approaches based on machine learning use signals measured during production to infer the production quality (Ahmed et al., 2023, 2024; Wang & Chen, 2024). In (Ahmed et al., 2023), the authors used Probabilistic Support Vector Regression (PSVR) for predicting the number of defects in a production lot based on data collected during the production process. In (Wang & Chen, 2024), XGBoost and Particle Swarm Optimization (PSO) were used to predict the yield of each wafer from the results of wafer acceptance tests (WATs). (Lundberg & Lee, 2017). In (Ahmed et al., 2024), the authors considered three sources of data: *i)* signals from machines used for semiconductor production, *ii)* wafer map images from probe tests on dies, and *iii)* results of

electrical tests performed before burn-in (BI) testing to predict the number of BI-relevant failures, i.e. the number devices that will not pass the BI test.

In (Figueroa et al., 2024), the authors developed a two-branches neural network for predicting the quality of semiconductor production lots, where one branch is dedicated to process wafer map images and the other to process signals from production machines. The proposed approach was applied to a synthetic case study characterized by an imbalance between high quality (HQ) and low quality (LQ) data.

Multi-branch neural networks remain black boxes, and the causes of low-quality production remain, thus, unknown, which hinder the identification of effective countermeasures.

Explainable artificial intelligence (XAI) has been recently introduced as a paradigm concerned with explaining black box models (Patel et al., 2024; Wang & Chen, 2024). In the context of wafer map classification, (Junayed et al., 2024) used Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) to highlight the part of the input image which the model focuses on for classifying the wafer map defect. In (Patel et al., 2024), the authors used Grad-CAM and Linear Interpretable Model-Agnostic Explanation (LIME) (Ribeiro & Guestrin, 2016) techniques for explaining the outcomes of a convolutional neural network (CNN) trained to identify anomalies in wafer maps.

In this paper, we propose to couple a multi-branch neural network with SHAP analysis for predicting production quality and identifying the causes of production anomaly. The method is applied to a multimodal synthetic dataset built in such a way that different anomalies are visible in different data modalities.

The remainder of the paper is organized as follows. Section 2 briefly introduces SHAP analysis. Section 3 introduces the proposed method for quality prediction and identification of the causes of the production anomaly. Section 4 describes the dataset used for evaluating the proposed method. Section 5 discusses the results, and finally Section 6 gives concluding remarks about the work done.

2. SHapley Additive exPlanation (SHAP)

Shapley Additive Explanation (SHAP) is a technique used for explaining predictions of black box models. It is based on Shapley values, which were introduced in game theory with the aim of fairly distributing a payout among different players in a cooperative game (Shapley, 1953). In the context of machine learning, the payout is the prediction of the model, and the players are its input features. Thus, the idea of SHAP is to calculate the contribution of an input feature to the model prediction using its Shapley value. Specifically, the Shapley value of input feature i is:

$$\phi_i = \sum_{S \subseteq \{1,\dots,M\} \setminus \{i\}} \frac{|S|!\,(M-|S|-1)!}{M!} \left(f(S \cup \{i\} - f(S))\right) \quad (1)$$

where M is the total number of features, S is a subset of features that does not include feature i and $f(\cdot)$ is the black-box model. The contribution of a subset of features to the model prediction can be computed as the sum of the Shapley values of the individual features in the subset, since SHAP considers the interactions between features when computing the importance of an individual feature (Molnar, 2020).

3. Method

Without loss of generality, we consider two sources of data: signal measurements from a machine of the semiconductors production process and wafer map images from electrical tests performed during production to identify defects in the products. We introduce the following mathematical formalism for the data of the generic l-th production lot:

- the vector $S(l)$ contains signal measurements collected from a production machine;
- $\{W(j)\}_{j=1,\dots,J}$ are J wafer map images collected during the production of the J wafers of the lot; these images are aggregated to obtain a single composite image per lot $\widetilde{W}(l)$, following the procedure illustrated in (Figueroa et al., 2024).

The training of the model is made using data from L production lots $D = \{S(l), y(l)\}_{l=1,..,L}$, where $y(l)$ is the number of BI-relevant failures.

The proposed model consists of five modules (Fig. 1):

I. An autoencoder for image reconstruction (AE_I).

II. An autoencoder for signal reconstruction (AE_S).

III. A neural network that takes the latent representations z_S provided by AE_S and z_I provided by AE_I, concatenates them, and uses them as input for predicting the number y of BI-relevant failures in the lot. Thus, the input of this model is $z = [z_S, z_I] \in \mathbb{R}^{d_S+d_I}$, with d_S and d_I being the dimensions of the latent representation provided by the encoders AE_S and AE_I, respectively.

IV. A module that determines the quality of the lot by applying the Clopper-Pearson estimator (Clopper & Pearson, 1934). This estimator uses y and the number of samples of the lot to compute the $(1-\alpha)$-quantile of a Beta distribution. The computed value is taken to represent the early life failure probability (ELFP) of the lot and is used to assess the quality of production: if it exceeds a preset threshold, the lot is considered as LQ, otherwise as HQ.

V. A module that computes the importance of the different data modalities considered by applying SHAP to the neural network of module III. Thus, for each feature, SHAP computes a contribution value ϕ_i. Then, the contributions, C_S and C_I, of the signal (S) and the image (I) modalities are obtained as the sum of the contributions of the corresponding individual features:

$$C_S = \left| \sum_{i=1}^{d_S} \phi_i \right| \quad (2)$$

$$C_I = \left| \sum_{i=1}^{d_I} \phi_i \right| \quad (3)$$

and the corresponding relative contribution as:

$$C_S^{\%} = \frac{C_S}{C_I + C_S} \quad (4)$$

$$C_I^{\%} = \frac{C_I}{C_I + C_S} = 1 - C_S^{\%} \quad (5)$$

During the training phase, autoencoders AE_I and AE_S are separately trained in an unsupervised way on the corresponding data modality. As a result, the encoders generate representations of the data that summarize their contents in a limited number of features. The neural network described in III is trained using the concatenated latent representations extracted from AE_I and AE_S as inputs and the number of BI-relevant failures as output.

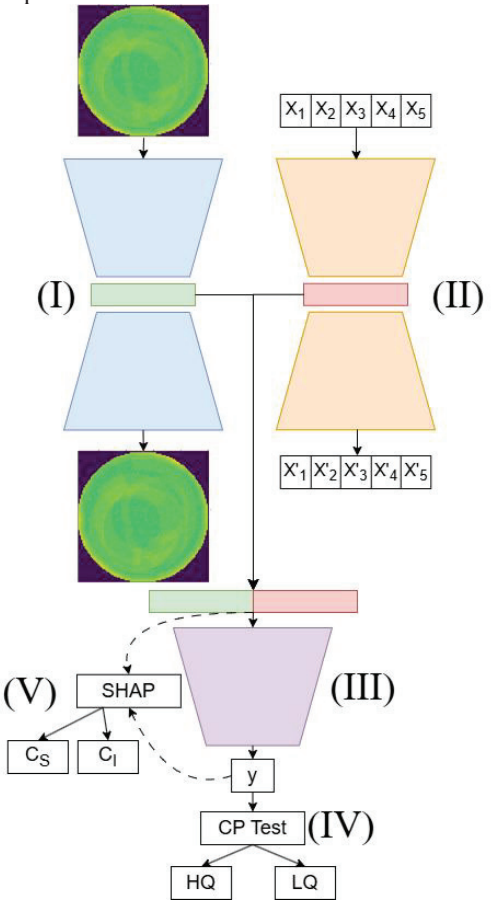

Fig. 1. Representation of the methodology for predicting the number of BI-relevant failures, classifying the lot quality, and estimating the

contribution of each data modality to the quality prediction outcome.

4. Case Study

We consider a synthetic dataset containing simulated signal values and wafer map images. We assume that the two data modalities refer to different production stages. Each pattern in the dataset refers to a production lot and is comprised of a vector of signals and an aggregated image, whose pixel values are obtained as the sum of the corresponding pixels of 25 individual wafer maps of the lot. The simulation of both data modalities (signals and images) is performed by random sampling an indicator of the quality of the production stage to which the data modality refers. Then, the procedure described in (Ahmed et al., 2023) for the generation of the synthetic signals and that in (Maksim et al., 2019) for the generation of the synthetic images are applied. This latter considers the four classes of wafer maps shown in Fig. 2. Class None refers to wafers without defects, whereas the remaining 3 classes refer to three different types of defects. The Donut pattern indicates annular clusters, the Edge-Ring pattern represents ring-shaped clusters around the edge of the wafer, and the Edge-Loc represents a more localized version of the Edge-Ring, with only clusters of defects around the edge of the wafer.

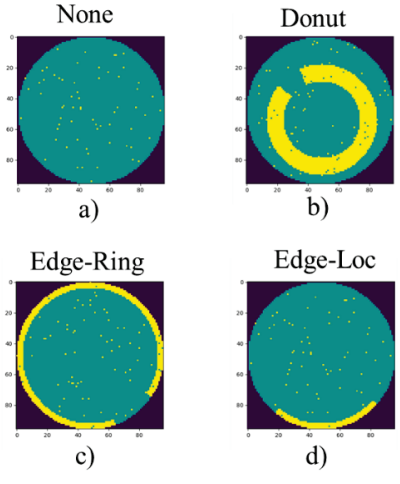

Fig. 2. Wafer map images of wafer produced by machines in normal (a) and abnormal (b, c and d) operating conditions.

Two examples of aggregated wafer map images are shown in Fig. 3.

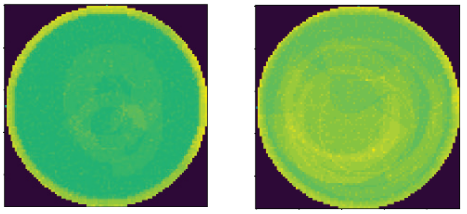

Fig. 3. Examples of aggregated wafer map images.

The dataset is comprised of 2000 multimodal patterns, 1800 of which are used for training and 200 for testing. The 1800 training patterns contain 900 HQ and 900 LQ lots. Among the LQ patterns, 3 different situations are possible: *i*) the anomaly occurs during the production stage monitored by the signals (referred to as LQ1) *ii*) the anomaly occurs during the production stage monitored by the images modality (LQ2), and *iii*) anomalies occur in both production stages (LQ0). Specifically, there are 180 LQ1 patterns, 180 LQ2 patterns, and 540 LQ0 patterns.

5. Results

Table 1 and 2 report the performance in the classification of the lot quality (output of module IV of Section 3), obtained performing a 5-fold cross-validation procedure. While the model achieves a satisfactory overall performance with accuracy and F1-score above 90%, it is more effective at identifying LQ lots (98.7% sensitivity) than HQ lots (83.4% specificity), indicating a tendency to err on the conservative side of classifying HQ as LQ.

Table 1 Performance of the model.

Metric	Value
Accuracy	91.05% ± 3.19%
Sensitivity	98.70% ± 0.75%
Specificity	83.40% ± 7.02%
F1-Score	91.78% ± 2.54%

Table 2 reports the accuracy of the classifier for lots of classes LQ1, LQ2 and LQ0. When the production anomaly occurs during the production stage monitored by the images (LQ2), the

performance is slightly less satisfactory than that obtained in the other cases.

Table 2 Accuracy in the classification of LQ0, LQ1 and LQ2 patterns.

Subset	Accuracy
LQ1	100.00% ± 0.00%
LQ2	94.00% ± 3.39%
LQ0	99.83% ± 0.33%

To investigate the capability of SHAP of identifying the production stage responsible of the anomaly, Table 3 reports the average contribution of each modality for patterns of classes LQ1, LQ2 and LQ0 correctly classified as LQ. As expected, the data modalities with the largest average contributions are the signals for patterns of class LQ1 and the images for patterns of class LQ2. These results confirm that SHAP can be used to assist the operators in the identification of the type of anomaly occurred during production. A limitation of the method is that it cannot be used to identify the problems affecting both data modalities (LQ0 class) for which the contribution $C_S^{\%}$ is larger than $C_I^{\%}$.

Table 3 Average contribution of each data modality for patterns of classes LQ1, LQ2 and LQ0.

Subset	$C_S^{\%}$ (Contribution of Signals)	$C_I^{\%}$ (Contribution of Wafer Map Images)
LQ1	69.3% ± 9.7%	30.7% ± 9.7%
LQ2	25.8% ± 5.0%	74.2% ± 5.0%
LQ0	77.9% ± 6.3%	22.1% ± 6.3%

5. Conclusions

In this paper, a multi-branch neural network for predicting the quality of production lots in the semiconductor industry from multimodality data has been combined with a module of SHAP analysis for identifying the type of anomaly responsible of LQ production. The results obtained on a synthetic case study have shown that: *i*) the multi-branch neural network is able to effectively classify the lot quality, with most error regarding HQ lots erroneously classified as LQ; *ii*) anomalies of different types, occurring in different production stages, can be distinguished by considering the SHAP values associated to signals and images. Future work will include the validation of the proposed method with real data from semiconductor production and its extension to allow distinguishing anomalies occurred in multiple production stages.

Acknowledgement

The participation of Joaquín Figueroa, Ibrahim Ahmed and Enrico Zio to this work is supported by the SAFEPOWER project under the HORIZON-CL5-2024-D3-01 call of the European Union, Grant Agreement 101172940. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them. The participation of Piero Baraldi to this work is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

References

Ahmed, I., Baraldi, P., Zio, E., Lewitschnig, H., & others. (2023). Prediction of the Number of Defectives in a Production Batch of Semiconductor Devices. *Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)*, 2615–2620.

Ahmed, I., Hosseinpour, F., Baraldi, P., Zio, E., & Lewitschnig, H. (2024). An artificial intelligence-based framework for burn-in reduction in the semiconductor manufacturing industry. In *Recent advances in microelectronics reliability: contributions from the European ECSEL JU project iRel40* (pp. 117–133). Springer.

Chen, T.-C. T. (2022). Big data analytics for semiconductor manufacturing. In *Production planning and control in semiconductor manufacturing: Big data analytics and Industry 4.0 applications* (pp. 1–19). Springer.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413.

Dahari, S., Talib, M. A., & Ghapor, A. A. (2025). Robust Control Chart Application in Semiconductor Manufacturing Process. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *43*(2), 203–219.

Figueroa, J., Ahmed, I., Baraldi, P., & Zio, E. (2024). Multibranch Neural Network For Predicting

Production Lot Quality In Semiconductor Industry. *ESREL 2024 Monograph Book Series*. https://esrel2024.com/part-9-crisis-management-support-systems-monitoring-and-early-warning-systems/

Junayed, M., Reza, T. T., & Islam, M. S. (2024). Enhancing Defect Recognition: Convolutional Neural Networks for Silicon Wafer Map Analysis. *2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, 1–6.

Maksim, K., Kirill, B., Eduard, Z., Nikita, G., Aleksandr, B., Arina, L., Vladislav, S., Daniil, M., & Nikolay, K. (2019). Classification of Wafer Maps Defect Based on Deep Learning Methods With Small Amount of Data. *2019 International Conference on Engineering and Telecommunication (EnT)*, 1–5. https://doi.org/10.1109/EnT47717.2019.9030550

May, G. S., & Spanos, C. J. (2006). *Fundamentals of semiconductor manufacturing and process control*. John Wiley & Sons.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Ooi, M. P.-L., Kassim, Z. A., & Demidenko, S. N. (2007). Shortening burn-in test: Application of HVST and Weibull statistical analysis. *IEEE Transactions on Instrumentation and Measurement*, *56*(3), 990–999.

Patel, T., Murugan, R., Yenduri, G., Jhaveri, R., Snoussi, H., & Gaber, T. (2024). Demystifying Defects: Federated Learning and Explainable AI for Semiconductor Fault Detection. *IEEE Access*.

Ribeiro, M. T., & Guestrin, C. (2016). " Why Should I Trust You ?" Explaining the Predictions of Any Classifier. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Shapley, L. S. (1953). A value for n-person games. *Contribution to the Theory of Games*, *2*.

Wang, S., & Chen, Y. (2024). Improved Yield Prediction and Failure Analysis in Semiconductor Manufacturing with XGBoost and Shapley Additive exPlanations Models. *2024 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*, 1–8.