Synthetic Wafer Test Data Generation – Principles, Methods, and Validation

Tarun Parmar

Independent Researcher
Austin, TX
ptarun@ieee.org

Abstract

Wafer test data plays a crucial role in semiconductor manufacturing, enabling defect identification, process optimization, and yield improvement. However, acquiring real-world data presents challenges, such as data scarcity, privacy concerns, and high costs. Synthetic data generation has emerged as a promising solution that offers increased data availability, privacy preservation, cost-effectiveness, and flexibility. This study explores the principles, methods, and validation techniques for generating synthetic wafer test data. Key techniques include randomized sampling with variability modeling to introduce controlled randomness, spatial modeling using Gaussian processes and Markov random fields for realistic defect map generation, and physics-based simulations incorporating semiconductor physics principles. Generative AI techniques such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are discussed, highlighting their suitability for different wafer test data types. GANs excel in visual inspection tasks, whereas VAEs are well suited for parametric testing and anomaly detection. The validation and evaluation of synthetic data quality are crucial, emphasizing the importance of preserving statistical similarity, correlations, and improving downstream tasks. The metrics and methods for assessing data quality, including statistical tests, visual inspections, and domain-specific metrics, are discussed. The potential for synthetic data to revolutionize semiconductor manufacturing by enhancing decision making, optimizing yields, and driving innovation. Future research directions include refining generative models, developing sophisticated validation techniques, and exploring hybrid modeling approaches that integrate synthetic and real-world data.

Keywords: Synthetic Data Generation, Wafer Test Data, Semiconductor Manufacturing, Generative AI Models, Data Validation

I. INTRODUCTION

Wafer test data plays a crucial role in semiconductor manufacturing, providing essential insights into the quality and performance of integrated circuits. These data are collected during various stages of production, including in-line testing and final electrical testing, and are used to identify defects, optimize processes, and improve the overall yield [1]. The importance of wafer test data lies in their ability to enable manufacturers to make data-driven decisions, enhance product reliability, and reduce costs associated with defective chips.

However, acquiring real-world wafer-test data presents several challenges. First, the sheer volume of data generated during semiconductor manufacturing can be overwhelming and requires sophisticated data management and analysis systems. Second, the sensitive nature of proprietary manufacturing processes often limits data sharing between companies, hindering collaborative research efforts. Third, the high cost and

time-consuming nature of data collection, particularly for rare defect cases, can impede comprehensive analysis and model development.

Synthetic data generation has emerged as a promising solution to address these challenges. Synthetic data refer to artificially created datasets that mimic the statistical properties and characteristics of real-world data [2]. In wafer testing, synthetic data can be generated to represent various defect patterns, process variations, and test results. This approach offers several advantages:

1. Increased data availability: Synthetic data can be generated in large quantities, overcoming the limitations of real-world data scarcity, particularly in rare defect cases.
2. Privacy preservation: Using synthetic data, manufacturers can share information without compromising sensitive proprietary information.
3. Cost-effectiveness: Generating synthetic data is often more cost-effective than collecting extensive real-world data, particularly for exploring various scenarios and edge cases.
4. Flexibility: Synthetic data generation allows the creation of diverse datasets representing different manufacturing conditions, enabling more robust model development and testing.

II. BACKGROUND

Wafer testing is a critical step in semiconductor manufacturing and involves evaluating the electrical performance and functionality of integrated circuits on silicon wafers. The process typically includes parametric testing to measure the electrical characteristics and functional testing to verify logical operations. However, real-world wafer test data often have limitations that can affect machine learning model development for defect prediction and process optimization.

Some key limitations of real-world wafer test data include the following.

1. Limited sample sizes for rare defect types
2. Imbalanced datasets with few defective samples compared to good samples.
3. Incomplete or missing data due to test failures or equipment issues.
4. Noise and measurement variations from test equipment
5. Difficulty capturing all possible defect scenarios and process variations.

These limitations can lead to challenges in developing robust machine learning models for applications like defect classification, yield prediction, and process control.

Synthetic data generation offers several benefits for enhancing machine-learning model training in wafer testing.

1. Augmenting limited real-world datasets with additional synthetic samples.
2. Creating balanced datasets by generating more samples of rare defect types.
3. Simulating a wider range of defect scenarios and process variations.
4. Producing noise-free data to isolate key features and patterns.
5. Enabling privacy-preserving model development by avoiding use of sensitive production data

By leveraging synthetic data, semiconductor manufacturers can build more accurate and generalizable machine learning models for defect prediction and process optimization. Synthetic data allow for larger training datasets, exploration of edge cases, and simulation of future scenarios. This can lead to improved

model performance, particularly for rare defect types, and enable more proactive process control and yield management.

However, care must be taken to ensure that the synthetic data accurately represent real-world wafer characteristics and defect patterns. A combination of real and synthetic data, along with domain expertise, is often ideal for developing robust machine-learning solutions for wafer testing and semiconductor manufacturing.

III. Data Generation Techniques

Randomized Sampling with Variability Modeling is a crucial technique in semiconductor manufacturing for introducing controlled randomness into the electrical and physical parameters. This method allows for the simulation of real-world variability in wafer production, enabling more accurate predictions of the device performance and yield. Engineers can generate realistic test data that reflect the inherent variability in manufacturing processes by incorporating statistical distributions and correlations between different parameters. This approach is particularly useful for modeling process variations, such as fluctuations in dopant concentrations, gate oxide thickness, or channel length, which can significantly affect device characteristics.

Spatial Modeling techniques, including Gaussian processes and Markov random fields, play a vital role in generating realistic defect maps for wafer testing [3]. These methods account for the spatial correlations between defects on a wafer, reflecting the fact that defects are often clustered or follow specific patterns owing to manufacturing processes. Gaussian processes provide a flexible framework for modeling spatial dependencies, allowing the incorporation of prior knowledge of defect distributions. Markov random fields, on the other hand, are particularly useful for modeling discrete spatial patterns and can capture complex interactions between neighboring regions on a wafer. These spatial modeling techniques enable more accurate predictions of defect locations and densities, thereby improving the efficiency of wafer testing and failure analysis processes.

Physics-Based Simulations are essential for understanding and predicting the key physical processes that affect the wafer test results. These simulations incorporate fundamental principles of semiconductor physics, such as carrier transport, electrostatics, and quantum mechanics, to model device behavior under various conditions [1]. Process models play a crucial role in these simulations by providing a link between the manufacturing parameters and device characteristics. For example, models of ion implantation, diffusion, and oxidation processes can be used to predict dopant profiles and junction depths, which in turn affect device performance. By integrating these physics-based models with statistical techniques, engineers can create more comprehensive and accurate simulations of wafer-level variability, enabling better optimization of manufacturing processes and improved yield prediction.

IV. Generative-AI Techniques

Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and other AI-based generative methods have emerged as powerful tools in various domains including semiconductor manufacturing and wafer testing [4]. These techniques offer unique approaches for data generation, feature extraction, and anomaly detection, making them valuable for different types of wafer tests.

GANs and VAEs are two types of generative AI models used in wafer testing. GANs use competing neural networks to generate realistic synthetic data, useful for augmenting datasets and improving defect detection. VAEs encode and decode data to learn its distribution, excelling in dimensionality reduction and anomaly detection. Other methods like autoregressive and flow-based models offer alternative approaches for data generation and analysis in wafer testing [4,5,6].

Several factors should be considered when comparing the suitability of these methods for different types of wafer tests.

1. Visual inspection: GANs are particularly well suited for generating realistic wafer images and detecting visual defects. They can augment datasets for training defect-detection algorithms and help identify rare defect types.
2. Parametric testing: VAEs excel in modeling the distribution of electrical parameters and detecting anomalies. They can capture complex relationships between parameters and identify subtle deviations from normal behavior.
3. Temporal analysis: Autoregressive models are more appropriate for analyzing time-series data in wafer testing, such as tracking parameter changes over time, or predicting future test results based on historical data.
4. Density estimation: Flow-based models offer advantages in estimating the probability densities of test data, making them useful for identifying outliers and anomalies in high-dimensional parameter spaces.
5. Data augmentation: Both GANs and VAEs can be used to generate synthetic wafer test data, helping to balance datasets and improve the performance of the machine learning models used in wafer testing.
6. Interpretability: VAEs generally offer better interpretability of the learned latent space than GANs, which can be beneficial for understanding the underlying factors affecting wafer test results.
7. Training stability: VAEs tend to have more stable training processes than GANs, which can be advantageous when dealing with complex wafer-test data.

In practice, the choice of the generative method for wafer testing depends on the specific requirements of the test, the nature of the data, and the desired outcomes. Hybrid approaches combining multiple generative methods may also be employed to leverage the strengths of each technique and address diverse challenges in wafer testing.

V. VALIDATION AND EVALUATION

To evaluate synthetic data quality effectively, researchers must employ a combination of statistical measures and domain-specific metrics. These assessments should focus on preserving key statistical properties, such as distributions, correlations, and relationships between variables, to ensure that the synthetic data closely mimic the original dataset. Statistical tests, such as the Kolmogorov-Smirnov test for distribution similarity and the Pearson correlation coefficient for relationship preservation [7], can be utilized to quantify the similarity between synthetic and real data [8]. Moreover, visual inspection techniques such as Q-Q plots and histograms can provide intuitive insights into the quality of the generated data.

Additionally, it is crucial to evaluate the performance of machine learning models trained on synthetic data compared with those trained on real data, as this provides insights into the utility and reliability of the generated data for downstream tasks. This evaluation can involve training models on both synthetic and real datasets and then comparing their performance on a held-out test set of real data. Metrics such as accuracy, precision, recall, and F1-score can be used to assess the model performance across various tasks. Furthermore, researchers should consider the generalization capabilities of models trained on synthetic data and examine their ability to capture nuanced patterns and edge cases present in real-world data.

Domain-specific metrics are equally important for assessing synthetic data quality because they account for the unique characteristics and requirements of specific fields. For instance, in healthcare, synthetic patient records should maintain realistic relationships among symptoms, diagnoses, and treatments. In financial data,

synthetic datasets should preserve the temporal patterns and complex interdependencies between economic indicators. By incorporating domain expertise into the evaluation process, researchers can ensure that synthetic data not only meet statistical criteria but also remain practically useful and meaningful within its intended context.

VI. CONCLUSION

Synthetic wafer test data generation offers a powerful solution for addressing the challenges in semiconductor manufacturing data acquisition. This study explored foundational concepts, methods, and validation techniques, highlighting their potential to revolutionize defect prediction, process optimization, and machine learning model development. Key aspects include controlled randomness through variability modeling, spatial modeling techniques, and physics-based simulations. Generative AI techniques, such as GANs and VAEs, show promise, with strengths suited to different wafer testing aspects. Rigorous validation using statistical measures, domain-specific metrics, and performance assessments is crucial to ensure data quality and reliability. As the semiconductor industry faces data scarcity and privacy concerns, synthetic data generation provides a valuable tool for enhancing decision making, optimizing yields, and driving innovation. Future research should focus on refining generative models, developing sophisticated validation techniques, and exploring hybrid modeling approaches that integrate synthetic and real-world data.

REFERENCES

[1] K. R. Skinner *et al.*, "Multivariate statistical methods for modeling and analysis of wafer probe test data," *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, no. 4, pp. 523–530, Nov. 2002, doi: 10.1109/tsm.2002.804901.

[2] D. Jiang, W. Lin, and N. Raghavan, "A Novel Framework for Semiconductor Manufacturing Final Test Yield Classification Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 197885–197895, Jan. 2020, doi: 10.1109/access.2020.3034680.

[3] M. Kim, J. Shin, and J. Tak, "A Deep Learning Model for Wafer Defect Map Classification: Perspective on Classification Performance and Computational Volume," *physica status solidi (b)*, vol. 261, no. 1, Nov. 2023, doi: 10.1002/pssb.202300113.

[4] S.-Y. Lee, J.-H. Kim, D. Kim, Y.-W. Lee, T. P. Connerton, and D. Kim, "Semi-GAN: An Improved GAN-Based Missing Data Imputation Method for the Semiconductor Industry," *IEEE Access*, vol. 10, pp. 72328–72338, Jan. 2022, doi: 10.1109/access.2022.3188871.

[5] S.-K. S. Fan, D.-M. Tsai, and P.-C. Yeh, "Effective Variational-Autoencoder-Based Generative Models for Highly Imbalanced Fault Detection Data in Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 36, no. 2, pp. 205–214, May 2023, doi: 10.1109/tsm.2023.3238555.

[6] I. Cho and Y. Ju, "Text mining method to identify artificial intelligence technologies for the semiconductor industry in Korea," *World Patent Information*, vol. 74, p. 102212, Jul. 2023, doi: 10.1016/j.wpi.2023.102212.

[7] B. Espinar *et al.*, "Analysis of different comparison parameters applied to solar radiation data from satellite and German radiometric stations," *Solar Energy*, vol. 83, no. 1, pp. 118–125, Aug. 2008, doi: 10.1016/j.solener.2008.07.009.

[8] D. O. Cardoso and T. D. Galeno, "Online evaluation of the Kolmogorov–Smirnov test on arbitrarily large samples," *Journal of Computational Science*, vol. 67, p. 101959, Feb. 2023, doi: 10.1016/j.jocs.2023.101959.