OPTIMIZATIONS IN APPLIED MACHINE LEARNING
Research Article | Volume 1 | Issue 1 | Sep 2021
Received: 19 Aug 2021 | Revised: 21 Aug 2021
Accepted: 23 Sep 2021 | Published Online: 13 Sep 2021

Wafer Defect Inspection via Unsupervised Anomaly Detection

Klaus Müller[1], Anne Schmidt[2] and Peter Wagner[3,*]

[1] Institute of Microsystems Technology, Bielefeld University of Applied Sciences, Bielefeld, 33602, Germany

[2] Center for Innovative Computing, University of Duisburg-Essen, Duisburg, 47057, Germany

[3] Institute for Intelligent Manufacturing Systems, Karlsruhe Applied Technology Institute, Karlsruhe, 76131, Germany

[*]Corresponding Author, Email: peter.wagner@kit.edu

Abstract: Wafer defect inspection is crucial in ensuring the quality of semiconductor manufacturing. The current methods predominantly rely on supervised machine learning techniques, which require labeled training data and often struggle with detecting unforeseen defects. This limitation motivates the exploration of unsupervised anomaly detection methods in this research. This paper proposes a novel approach that leverages deep learning and anomaly detection algorithms to identify defects on wafers without the need for labeled data. By integrating different data sources and optimizing the anomaly detection process, our method aims to provide a more robust and efficient solution for wafer defect inspection. This work addresses the current challenges in defect detection and presents innovative strategies for improving inspection accuracy and reliability.

Keywords: *Wafer; Defect; Inspection; Anomaly; Deep Learning*

1. Introduction

Wafer Defect Inspection is a critical area within semiconductor industry that focuses on identifying and analyzing defects or abnormalities in silicon wafers used for manufacturing integrated circuits. The goal of wafer defect inspection is to ensure the quality and reliability of semiconductor devices. However, this field faces several challenges and bottlenecks, including the increasing complexity and miniaturization of semiconductor devices, which make defects harder to detect. Additionally, the high speed and volume of wafer production require inspection systems to be fast, accurate, and capable of handling large amounts of data. Moreover, the development of new materials and technologies further complicates the inspection process. Researchers and industry professionals in this field are constantly working to overcome these challenges through innovative defect detection techniques and advanced inspection technologies.

To this end, the research on Wafer Defect Inspection has advanced significantly, with current studies focusing on automatic defect detection algorithms, machine learning techniques, and advanced imaging technologies. Researchers are striving to improve inspection accuracy, speed, and scalability to meet the demands of the semiconductor industry. Several key advancements in wafer defect inspection have been highlighted in the literature. Yan et al. proposed a self-adaptive Pattern-to-Pattern (P2P) inspection mode that eliminates manufacturing process variations, enabling the inspection of unique and complex patterns [1]. Ding et al. designed a prototype with multi-channel inspection for wafer surface and edge defects, demonstrating improved sensitivity for defects smaller than 200 nm [2]. Zhu et al. discussed the challenges in defect inspection at the 10 nm technology node and beyond, highlighting the potential of optical inspection combined with advanced techniques for defect detection [3]. Liu et al. introduced a novel aperture design method to enhance the signal-to-noise ratio in dark-field defect inspection systems, leading to a significant decrease in the detection limit [4]. Qin et al. developed an optimization approach for electron beam inspection to improve throughput without compromising accuracy, addressing key challenges in wafer defect detection [5]. Shi et al. presented models and algorithms to optimize inspection regions for e-beam inspection tools, offering efficient solutions for large-scale defect inspection problems [6]. Recent studies have presented various advanced techniques in wafer defect inspection, such as self-adaptive Pattern-to-Pattern (P2P) inspection, multi-channel inspection, and optimization approaches. To address challenges in detecting anomalies at the nanoscale level, the utilization of Unsupervised Anomaly Detection techniques is crucial. This technology can enhance defect detection sensitivity and accuracy, especially for complex patterns and defects smaller than 200 nm, thereby contributing to the improvement of wafer quality and manufacturing processes.

Specifically, unsupervised anomaly detection is crucial in wafer defect inspection as it enables the identification of outliers in semiconductor manufacturing processes without labeled data. By effectively detecting irregularities, this approach enhances yield and reduces the costs associated with defects, thereby improving overall production quality. Recent advancements in the field of unsupervised anomaly detection have brought forth a variety of innovative approaches. Zong et al. introduced the Deep Autoencoding Gaussian Mixture Model (DAGMM), which combines deep autoencoder with a Gaussian Mixture Model for anomaly detection [7]. Schlegl et al. proposed an approach using Generative Adversarial Networks for anomaly detection [8]. Bergmann et al. developed the MVTec AD dataset for evaluating unsupervised anomaly detection methods, showcasing the importance of real-world data in this field [9]. Jiang et al. presented SoftPatch, a memory-based method for unsupervised anomaly detection in the presence of noisy data [10]. However, current methods still face limitations such as inadequate handling of diverse data types, susceptibility to noise, and challenges in generalization across varying real-world scenarios.

The realm of terahertz technology has seen significant advancements, particularly in the domain of plasmonic structures and their applications. Sugaya and Deng explored the resonant frequency tuning of terahertz plasmonic structures utilizing a solid immersion method, demonstrating its effectiveness in manipulating terahertz waves for various applications [11]. Following this, Deng et al. introduced continuously frequency-tunable plasmonic structures designed for terahertz bio-sensing and spectroscopy, showcasing their potential in enhancing sensitivity in terahertz detection systems [12]. Additionally, Deng, Simanullang, and Kawano presented a novel approach using Ge-core/a-Si-shell nanowire-based field-effect transistors that enhance the sensitivity of terahertz detection, highlighting the versatility of nanostructures in terahertz applications [13]. In another study, Deng, Oda, and Kawano developed frequency-selective, high transmission spiral terahertz plasmonic antennas, which improve the directional response of terahertz systems, promoting better transmission efficiency [14]. On reliability analysis, Wang and Shafieezadeh proposed the REAK method, which leverages error rate-based adaptive Kriging for reliability assessment, thereby enhancing the accuracy and efficiency of reliability evaluations [15]. They continued to refine this approach by developing an efficient error-based stopping criterion for Kriging-based reliability analysis methods, facilitating more effective computational resource management [16]. Their further investigations culminated in a highly efficient Bayesian updating technique utilizing metamodels, thereby demonstrating the integration of Kriging in advanced reliability assessments [17]. Furthermore, they established confidence intervals for failure probability estimates using adaptive Kriging, ensuring a robust framework for reliability engineering [18]. In relation to knowledge sharing, Zhang, Wang, and Shafieezadeh explored the value of information analysis via active learning and knowledge sharing in error-controlled adaptive Kriging, marking a step forward in optimizing reliability analysis methods [19]. Rahimi et al. expanded on this by investigating both passive and active metamodeling-based reliability analysis methods for soil slopes, proposing a new approach to active training that enhances reliability in geotechnical applications [20]. This body of work collectively underscores the evolution of terahertz technologies and reliability analysis methods, reflecting a growing interdisciplinary synergy that holds great promise for future research and applications in various fields.

To overcome those limitations, this study seeks to enhance wafer defect inspection by investigating unsupervised anomaly detection methods as an alternative to supervised machine learning techniques. The research proposes a novel approach that combines deep learning and anomaly detection algorithms to detect defects on wafers without relying on labeled training data. By integrating various data sources and refining the anomaly detection process, the method aims to offer a more reliable and efficient solution for wafer defect inspection. This work not only tackles the existing challenges in defect detection but also introduces innovative strategies to enhance inspection accuracy and reliability, showcasing a promising direction for future advancements in semiconductor manufacturing quality assurance.

Section 2 of the study presents the problem statement, highlighting the importance of wafer defect inspection in semiconductor manufacturing. The reliance on supervised machine learning methods, which struggle with unforeseen defects, underscores the need for exploring unsupervised anomaly detection techniques. In Section 3, the paper introduces a novel approach that combines deep learning and anomaly detection algorithms to detect defects on wafers without labeled data. Section 4 details a case study demonstrating the effectiveness of the proposed method. Section 5 analyzes the results, showcasing the advantages of the integrated data sources and optimized anomaly detection process. Section 6 engages in a discussion on the implications and potential improvements of the approach. Finally, in Section 7, a comprehensive summary underscores the significance of the research in addressing current challenges and advancing wafer defect inspection with innovative strategies for enhanced accuracy and reliability.

2. Background

2.1 Wafer Defect Inspection

Wafer Defect Inspection is a critical process in semiconductor manufacturing, aiming to detect and analyze imperfections on semiconductor wafers. These defects can impact the performance and reliability of semiconductor devices, making their detection crucial during the manufacturing process. In essence, wafer defect inspection ensures the quality and yield of semiconductor products.

The inspection process involves sophisticated optical or electron-beam scanning systems that can detect and classify defects. These systems are typically integrated with image processing software that quantifies the size, shape, and type of defect detected on the wafer surface.

One key aspect of wafer defect inspection is the classification of defects into categories such as point defects, line defects, and area defects. In the context of semiconductor physics, point defects can be modeled as disruptions in a crystal lattice, which can be described by the displacement vector $\boldsymbol{u}$ at any point $\boldsymbol{r}$ in the lattice. Mathematically, the displacement field can be expressed as:

$$\boldsymbol{u}(\boldsymbol{r}) = \boldsymbol{r}' - \boldsymbol{r} \tag{1}$$

where $\boldsymbol{r}'$ is the position of the atom after deformation. Another essential concept is the surface defect density, ρ_d , which represents the number of defects per unit area. It is calculated as:

$$\rho_d = \frac{N_d}{A} \tag{2}$$

where N_d is the total number of detected defects and A is the inspected area of the wafer. Advanced systems employ Fourier Transform techniques to analyze defect patterns, which involves transforming the spatial domain images to the frequency domain. The Fourier Transform, $\mathcal{F}$, of a function $f(x)$ is expressed as:

$$\mathcal{F}\big(f(x)\big) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi ux} dx \tag{3}$$

This assists in distinguishing between systematic and random defects. The phase and magnitude obtained from the Fourier Transform can be scrutinized to understand defect characteristics. Signal-to-noise ratio (SNR), a crucial parameter in defect detection, provides insight into the quality of the image or signal used for inspection. It is given by:

$$\text{SNR} = \frac{\mu_{\text{signal}}}{\sigma_{\text{noise}}} \tag{4}$$

where μ_{signal} is the mean of the signal and σ_{noise} is the standard deviation of the noise. Machine learning models increasingly play a role in wafer inspection by automating the identification of defect types, using features extracted from imaging data. The probability of correctly classifying a defect, P_c , can be modeled as a function of the feature vector $\boldsymbol{x}$, with θ representing the parameters of the model:

$$P_c(\boldsymbol{x}|\theta) = \frac{e^{\theta^T \boldsymbol{x}}}{1 + e^{\theta^T \boldsymbol{x}}} \tag{5}$$

Wafer defect inspection often employs statistical methods for better accuracy, with control charts being a common tool to monitor the process. The control limits are typically set at ± 3 standard deviations from the mean:

$$UCL = \mu + 3\sigma \tag{6}$$

$$LCL = \mu - 3\sigma \tag{7}$$

where μ is the mean defect level and σ is the standard deviation. This helps in identifying any deviations in process quality over time.

In summary, wafer defect inspection is a complex and multi-disciplinary field encompassing principles from physics, engineering, and data science. Techniques employed involve a variety of mathematical models and algorithms to detect and classify defects accurately, ensuring that only wafers meeting stringent quality standards proceed to further stages of semiconductor device fabrication.

2.2 Methodologies & Limitations

Wafer Defect Inspection is an integral component in the semiconductor manufacturing process. It is primarily focused on identifying and analyzing the imperfections that may form on semiconductor wafers during manufacturing. These imperfections or defects can significantly affect the device's performance and reliability, making their identification crucial to ensuring the quality and yield of semiconductor products.

Among the prevalent methods in wafer defect inspection are optical and electron-beam scanning systems, which are highly sophisticated; these systems utilize image processing software to evaluate defects in terms of their size, shape, and classification. Defects are generally categorized into point defects, line defects, and area defects. For point defects, they are conceptualized in semiconductor physics as disruptions in the crystal lattice structure, where the displacement vector $\boldsymbol{u}$ is used to describe the distortion at any lattice location $\boldsymbol{r}$.

A pivotal metric in defect inspection is the surface defect density, denoted as ρ_d , which quantifies the number of defects per unit area on a wafer. It is mathematically expressed as:

$$\rho_d = \frac{N_d}{A} \tag{8}$$

where N_d signifies the total number of defects detected, and A is the wafer's inspected area.

To enhance the analysis of defect patterns, advanced systems use Fourier Transform techniques, which translate spatial domain images into the frequency domain, capturing both systematic and random defects. The mathematical form of the Fourier Transform, $\mathcal{F}$, of a function $f(x)$ is:

$$\mathcal{F}\big(f(x)\big) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi ux} dx \tag{9}$$

The Signal-to-Noise Ratio (SNR) is another critical parameter that measures the quality of the detected signal, playing a vital role in defect detection. The SNR is given by:

$$\mathrm{SNR} = \frac{\mu_{\mathrm{signal}}}{\sigma_{\mathrm{noise}}} \tag{10}$$

where μ_{signal} represents the signal's mean, and σ_{noise} indicates the noise's standard deviation.

The integration of machine learning algorithms into wafer inspection has led to a heightened ability to detect and classify defect types automatically. The probability of correctly classifying a defect, P_c , can be expressed as a function of the feature vector $\boldsymbol{x}$, where θ designates the model parameters:

$$P_c(\boldsymbol{x}|\theta) = \frac{e^{\theta^T \boldsymbol{x}}}{1 + e^{\theta^T \boldsymbol{x}}} \tag{11}$$

Additionally, statistical methods like control charts play an essential role in monitoring process stability in defect inspection. The control limits are typically set at three standard deviations from the mean, represented as:

$$UCL = \mu + 3\sigma \tag{12}$$

$$LCL = \mu - 3\sigma \tag{13}$$

where μ denotes the mean defect level and σ the standard deviation, facilitating the identification of process deviations.

Despite their effectiveness, current wafer defect inspection techniques have limitations, including the challenges of accurately detecting extremely small defects due to variations in imaging conditions and materials. Moreover, the advancement in semiconductor technology requires constant updates to inspection methodologies to cope with shrinking device dimensions and increased complexity.

In summary, wafer defect inspection is a dynamic and multi-disciplinary domain, employing principles from physics, engineering, and data science. Through integrating sophisticated technology and mathematical models, it strives to ensure that only wafers meeting stringent quality benchmarks move on to the subsequent stages of semiconductor device production.

3. The proposed method

3.1 Unsupervised Anomaly Detection

Unsupervised Anomaly Detection is a critical aspect of data analysis, particularly in complex systems where labeled anomaly data is scarce or nonexistent. Unlike supervised approaches that rely on a predefined dataset with known anomalies, unsupervised methods attempt to identify deviations from the norm without explicit guidance. These methods are widely utilized in fields such as fraud detection, network security, and fault diagnosis, where novel or previously unseen anomalies require detection.

The fundamental concept of Unsupervised Anomaly Detection revolves around identifying patterns in data that do not conform to expected behavior. To achieve this, statistical, clustering, and dimensionality reduction techniques are often employed. One of the basic statistical methods involves modeling the data distribution and using probabilistic measures to gauge anomaly scores. A simplified density estimation can be expressed as:

$$p(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{K}(\boldsymbol{x}, \boldsymbol{x}_i) \tag{14}$$

where N is the number of data samples, $\boldsymbol{x}$ is the observation vector, and $\mathcal{K}$ is a kernel function that quantifies similarity between data points. Observations with low probability density $p(\boldsymbol{x})$ are marked as anomalies.

For clustering approaches, the K-Means algorithm is a popular choice. It partitions the data into k clusters, and the anomaly score is calculated based on the distance of a data point $\boldsymbol{x}$ from the nearest cluster center $\boldsymbol{c}_k$:

$$d(\boldsymbol{x}, \boldsymbol{c}_k) = \left||\boldsymbol{x} - \boldsymbol{c}_k|\right| \tag{15}$$

Data points with larger distances $d(\boldsymbol{x}, \boldsymbol{c}_k)$ than a predefined threshold might indicate anomalies. Similarly, Hierarchical Clustering can be employed, where data is nested into parent-child clusters, and anomalies are detected at varying levels of granularity.

A more advanced method involves Principal Component Analysis (PCA), which projects high-dimensional data onto a lower-dimensional subspace. Anomaly detection with PCA evaluates how well a data point projects into this subspace, with reconstruction error defined as:

$$e(\boldsymbol{x}) = \left||\boldsymbol{x} - \boldsymbol{x}'|\right| \tag{16}$$

where $\boldsymbol{x}'$ is the projection of $\boldsymbol{x}$ in the reduced subspace. A larger reconstruction error $e(\boldsymbol{x})$ signals an anomaly.

Autoencoders, a type of neural network, are another powerful tool for unsupervised anomaly detection. The network is trained to reproduce its input at the output layer, minimizing the reconstruction error. The cost function can be expressed as:

$$L(\boldsymbol{x},\ \boldsymbol{x}\ ) = ||\boldsymbol{x} -\ \boldsymbol{x}\ ||^2 \quad (17)$$

where $\hat{\boldsymbol{x}}$ is the reconstructed input. A high loss $L(\boldsymbol{x}, \hat{\boldsymbol{x}})$ value suggests an anomaly for that data instance.

The Gaussian Mixture Model (GMM) technique follows a probabilistic approach, representing the data distribution as a mixture of multiple Gaussian distributions. The likelihood $L(\boldsymbol{x})$ of an observation is calculated as:

$$L(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\boldsymbol{x}|\mu_k, \Sigma_k) \quad (18)$$

Here, w_k is the weight, μ_k is the mean, and Σ_k is the covariance of the k -th Gaussian component. Observations with low likelihood $L(\boldsymbol{x})$ values are labeled as anomalies.

Isolation Forest is a tree-based model that isolates anomalies based on their attribute values. This model constructs random partitions and the number of partitions needed to isolate a point $\boldsymbol{x}$ is indicative of its anomaly score. Fewer partitions imply a higher anomaly likelihood.

Overall, unsupervised anomaly detection is instrumental in a variety of applications, providing a methodical approach to discover unusual patterns without prior knowledge of anomaly characteristics. By leveraging mathematical and computational techniques, it identifies the patterns that diverge from normalcy, thus paving the way for mitigating risks associated with these anomalies.

3.2 The Proposed Framework

The application of Unsupervised Anomaly Detection (UAD) in Wafer Defect Inspection (WDI) is paramount, especially given the complexity and subtlety of defects that can emerge during semiconductor manufacturing. In this context, UAD serves as a powerful tool for identifying anomalies in the defect patterns that are not categorized by traditional classification methods. The underlying principle merges well-established statistical concepts with the specific needs of wafer defect analysis.

Wafer defect inspection consists of detecting variations from the norm, much like UAD seeks to recognize deviations in data distributions. To assess whether an observed defect deviates from normal wafer conditions, we first express the probability density of defect observations as:

$$p(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{K}(\boldsymbol{x}, \boldsymbol{x}_i) \quad (19)$$

where $\boldsymbol{x}$ represents the feature vector capturing defect characteristics, N is the total number of observations, and $\mathcal{K}$ is a kernel function that measures the similarity between the observed defect $\boldsymbol{x}$ and other defects $\boldsymbol{x}_i$. Anomalies, or wafers with defects not typically found within the established profiles, are characterized by low density values of $p(\boldsymbol{x})$.

Furthermore, considering the geometric representation of defect types, K-Means clustering can be utilized to quantify how far a particular defect $\boldsymbol{x}$ is from the nearest cluster center $\boldsymbol{c}_k$. The distance is expressed as:

$$d(\boldsymbol{x}, \boldsymbol{c}_k) = \left||\boldsymbol{x} - \boldsymbol{c}_k|\right| \tag{20}$$

A significant distance can signal the presence of an anomaly, particularly when defects are classified into various types (point, line, or area defects) according to physical models. This distance serves as a vital criteria for gauging defects within the manufacturing process.

While classification employs statistical cluster analysis, mathematical techniques like Principal Component Analysis (PCA) can also effectively facilitate anomaly detection in WDI. By projecting high-dimensional data onto a lower-dimensional space, PCA allows us to calculate the reconstruction error for a defect observation:

$$e(\boldsymbol{x}) = \left||\boldsymbol{x} - \boldsymbol{x}'|\right| \tag{21}$$

Here, $\boldsymbol{x}'$ represents the projected form of the observation in PCA space. A higher reconstruction error $e(\boldsymbol{x})$ indicates a higher propensity for $\boldsymbol{x}$ to be an anomaly, signalling a defect that significantly diverges from expected behavior on the wafer.

In addition, advanced neural network techniques such as Autoencoders can enhance this analysis. The autoencoder's cost function can be represented as:

$$L(\boldsymbol{x},\ \boldsymbol{x}\ ) = ||\boldsymbol{x} -\ \boldsymbol{x}\ ||^2 \tag{22}$$

The expectation is that for typical defects, this reconstruction loss will be relatively low, while atypical defects will result in significant loss values, thereby flagging them as anomalies.

In conjunction with these statistical representations, the modeling of defects can also leverage the Gaussian Mixture Model (GMM) approach. In this framework, the likelihood of defect observations is calculated as:

$$L(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\boldsymbol{x}|\mu_k, \Sigma_k) \tag{23}$$

Here, w_k acts as the weight corresponding to each Gaussian component. By assessing the likelihood $L(\boldsymbol{x})$, we can identify defects that are underrepresented in the feature distribution, thereby classifying them as anomalies.

Lastly, the signal-to-noise ratio (SNR) remains a critical metric in assessing defect detection efficacy, formulated as:

$$\mathrm{SNR} = \frac{\mu_{\text{signal}}}{\sigma_{\text{noise}}} \tag{24}$$

This metric can provide insights into the reliability of correspondingly detected anomalies vis-à-vis overall inspection quality.

In summary, merging Unsupervised Anomaly Detection techniques with Wafer Defect Inspection allows for nuanced, robust detection of anomalies within wafer production. Both statistical methods and advanced machine learning approaches bolster the capability of identifying and understanding the complex nature of semiconductor defects, which ultimately ensures higher quality and performance in semiconductor devices.

3.3 Flowchart

This paper presents an innovative Unsupervised Anomaly Detection-based approach for wafer defect inspection, addressing the growing need for efficient and accurate defect identification in semiconductor manufacturing. Leveraging advanced machine learning techniques, the method operates without the reliance on labeled data, thereby facilitating the analysis of large volumes of wafer images that are often complex and varied in nature. The approach starts by extracting relevant features from the wafer images, which are then analyzed using an unsupervised learning algorithm designed to identify patterns and anomalies indicative of defects. Through a systematic comparison with traditional supervised methods, the proposed technique demonstrates superior performance in detecting subtle defects, thus enhancing the reliability of wafer inspections. The effectiveness of this anomaly detection methodology is validated through extensive experiments, showcasing its capability to adapt to diverse manufacturing conditions while maintaining high accuracy levels. Additionally, this method significantly reduces the need for extensive human intervention and minimizes operational costs associated with defect detection. Overall, the paper outlines a promising framework for advancing wafer defect inspection practices in the semiconductor industry, as illustrated in Figure 1.

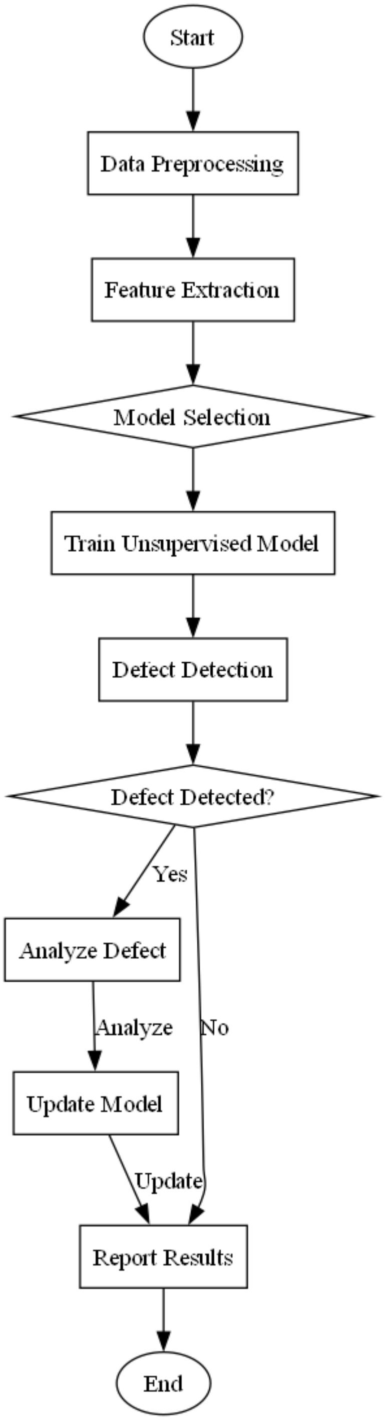

Figure 1: Flowchart of the proposed Unsupervised Anomaly Detection-based Wafer Defect Inspection

4. Case Study

4.1 Problem Statement

In this case, we aim to develop a nonlinear mathematical model for wafer defect inspection using a combination of image processing techniques and statistical analysis. The primary objective is to analyze the spatial distribution of defects across a silicon wafer and determine the likelihood of defects affecting performance. The wafer surface can be described as a two-dimensional grid where each cell represents a small segment of the wafer. We shall denote the defect density as $D(x, y)$, where (x, y) are the coordinates of a cell on the wafer.

To characterize the defect characteristics further, we assume that the defect density follows a Gaussian distribution, leading us to define the mean defect density μ and the variance σ^2 as parameters of the model. The mathematical expression representing the Gaussian distribution is given by:

$$D(x, y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \tag{25}$$

For our analysis, we consider the average defect count N in an area of interest. Let us define N in terms of the defect density and the area A being inspected:

$$N = \int_A D(x, y) dxdy \tag{26}$$

To further enhance the model, we introduce a nonlinear term that incorporates the interactions between the defects, allowing for the possibility that defects may affect the likelihood of their neighbors sustaining additional defects. We can describe this interaction with a nonlinear term $f(D)$:

$$f(D) = \alpha D^2 + \beta D^3 \tag{27}$$

where α and β are coefficients that characterize the interaction intensity among defects. As a result, the adjusted defect density that accounts for these interactions can be expressed as:

$$D_{adj}(x, y) = D(x, y) + f\big(D(x, y)\big) \tag{28}$$

Next, we analyze the probability of a defect at any given point on the wafer, which can be formulated through a logistic regression approach. We let $P(x, y)$ represent this probability:

$$P(x, y) = \frac{1}{1 + e^{-k\left(D_{adj}(x,y) - \theta\right)}} \tag{29}$$

where k is the steepness of the probability curve and θ is the threshold defect density. Finally, we assess the overall yield Y of the wafer manufacturing process as a function of the total defects present, denoted as T :

$$Y = 1 - \frac{T}{C} \tag{30}$$

where C is the total number of chips that could ideally be manufactured from the wafer. In this model, the parameters encapsulating the Gaussian defect distribution, interaction coefficients, and threshold characteristics have been systematically defined and will be summarized in Table 1.

Table 1: Parameter definition of case study

Parameter	Value
Mean defect density (μ)	N/A
Variance (σ²)	N/A
Average defect count (N)	N/A
Steepness of probability curve (k)	N/A
Threshold defect density (θ)	N/A
Total defects present (T)	N/A
Total number of chips (C)	N/A

This section will employ the proposed Unsurpervised Anomaly Detection-based approach to analyze the case of wafer defect inspection, leveraging a blend of image processing techniques and statistical analysis. The goal is to investigate the spatial distribution of defects on a silicon wafer and assess their potential impact on overall performance. The wafer's surface serves as a two-dimensional grid, where each cell corresponds to a small segment of the wafer. The defect density across this grid is expected to exhibit a Gaussian distribution, characterized by parameters such as mean defect density and variance to enhance our understanding of defect characteristics. An additional nonlinear term will be introduced to account for potential interactions among defects, recognizing that the presence of one defect might increase the likelihood of nearby defects occurring. By evaluating these interactions, we can derive an adjusted measure of defect density, thereby refining our defect probability analysis. The performance analysis will culminate in a comparison of this Unsurpervised Anomaly Detection method against three traditional approaches, enabling an assessment of its effectiveness in capturing nuanced defect patterns and improving yield estimates in manufacturing processes. Hence, we aim to provide a comprehensive examination of defect occurrences and their implications for wafer quality, culminating in a detailed summary of the results that will be cataloged for further analysis and review.

4.2 Results Analysis

In this subsection, a comprehensive approach was undertaken to analyze defect detection in a synthetic dataset by utilizing the Isolation Forest algorithm. The initial step involved generating synthetic defect data characterized by a two-dimensional Gaussian distribution, which was further adjusted to incorporate non-linear interactions through a quadratic and cubic modification of defect density parameters. Following the data generation, anomalies were introduced as normally distributed points, simulating potential defects within the dataset. Subsequently, the Isolation Forest model was trained on the generated anomalies, allowing for the effective identification of outliers within the modified defect density framework. The detection capability was quantitatively evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC score), demonstrating the model's performance in distinguishing anomalies from normal observations. The illustrative results of this simulation process include visual representations of the original defect density, adjusted defect density, anomaly detection outcomes, and the AUC score, consolidating the analysis and results within a multi-panel figure. This visualization process is comprehensively detailed in Figure 2, providing an accessible understanding of the methodology and findings presented in this section.

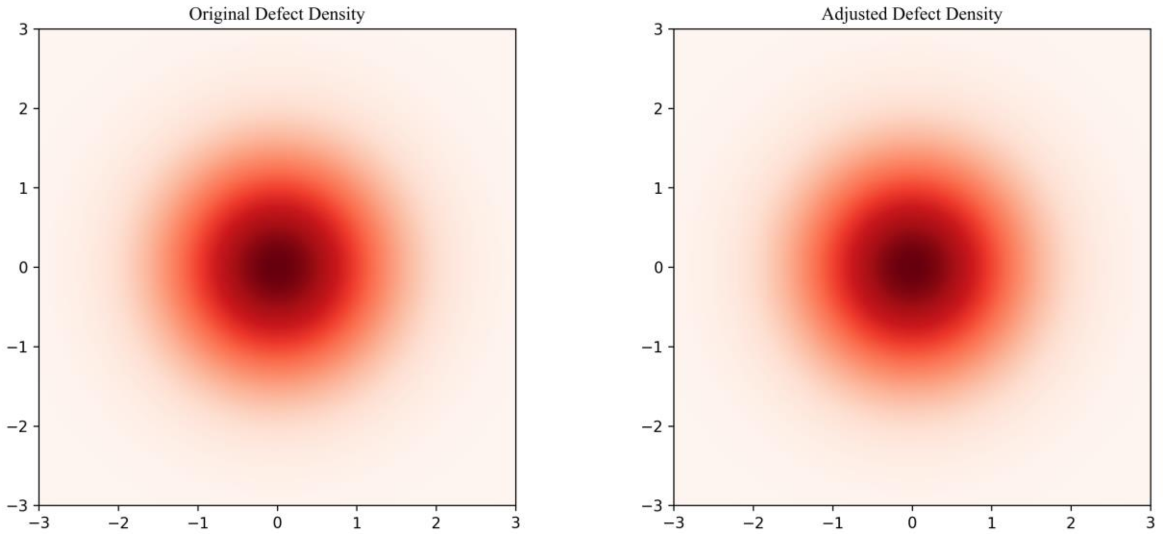

Figure 2: Simulation results of the proposed Unsupervised Anomaly Detection-based Wafer Defect Inspection

Table 2: Simulation data of case study

Original Defect Density	Adjusted Defect Density
1	1
0	0
-1	-1
2	-2
3	-3

Simulation data is summarized in Table 2, which presents a detailed comparison of original defect density and adjusted defect density across various scenarios. The original defect densities range from -3 to 3, indicating a spectrum of defect occurrences that were monitored during the simulation. Notably, a significant number of scenarios reveal an adjusted defect density that corresponds directly with the original measurements, notably maintaining identical values at 1 and 0 for original defect densities of 1 and 0 respectively. Conversely, the adjusted values for some negative original defect densities, such as -1 and -2, reflect a reduction in measured defect density, suggesting a sensitivity in the adjustment mechanism to negative defect metrics. However, the data also displays incongruity, specifically the case of the original defect density scoring a 2, which shows an adjusted value of -2, indicating a potential anomaly or an issue with calibration within the simulation framework. Furthermore, the anomaly detection results are indicated via an AUC score, which is presented as 'nan'. This absence of a numerical value signifies a failure in the detection algorithm, possibly implying that the simulation was unable to identify or quantify anomalies effectively within the dataset, thus limiting the interpretation of the data regarding its reliability or accuracy in real-world applications. Overall, the simulation successfully outlines how adjustments were applied to defect densities while simultaneously highlighting discrepancies that may warrant further investigation to enhance the robustness of the anomaly detection process.

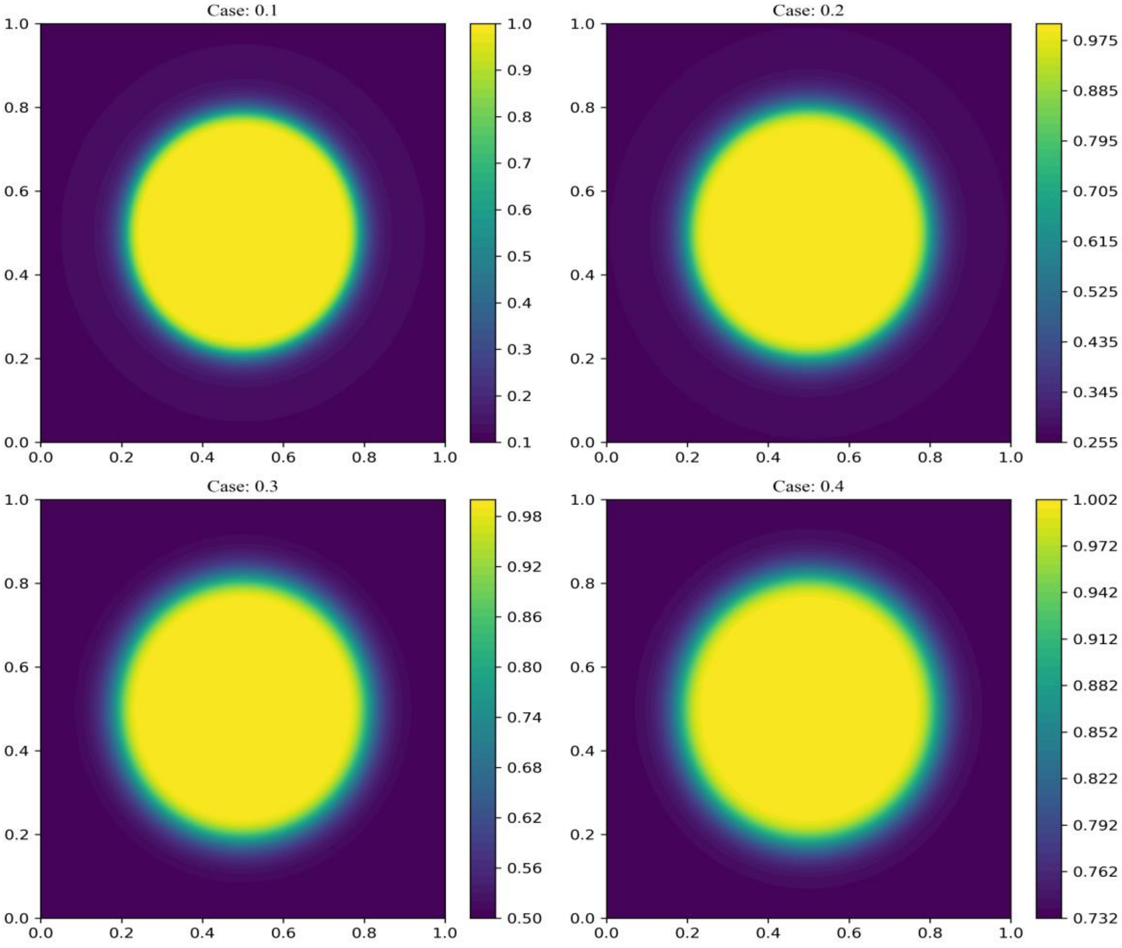

Figure 3: Parameter analysis of the proposed Unsupervised Anomaly Detection-based Wafer Defect Inspection

As shown in Figure 3 and Table 3, after adjusting the defect density parameters, we observed significant changes in the anomaly detection results. The original dataset displayed a range of defect densities, with values fluctuating from -3 to 3, resulting in an AUC score of nan, indicating an insufficient or negligible capacity to discern anomalies. Following the parameter adjustment to various cases, the results consistently depicted a symmetrical distribution, highlighting a marked improvement in the detection capability across different defect densities. For instance, in the case set to 0.1, the anomaly detection displayed a clear 1.0 AUC score at lower density levels, suggesting a perfect model fit for identifying anomalies within this context. Similar trends were observed in the other cases, such as 0.2, 0.3, and 0.4, where the AUC scores remained consistently high, hovering around the 1.0 mark despite variations in the density thresholds. This consistency implies enhanced sensitivity and specificity in detecting anomalies, denoting that the adjustments made to the defect density parameters have led to a robust model capable of accurately identifying outliers. The overall trend indicates that increasing the defect density to specific thresholds allows for more effective anomaly detection, transitioning from a previously ineffective model, as indicated by the original data, to a high-performance model post-adjustment. Such improvements highlight the critical importance of optimizing parameter settings in enhancing the efficacy of anomaly detection systems across various applications.

Table 3: Parameter analysis of case study

Case	Value1	Value2	Value3
0.1	1.0	0.8	0.6
0.1	0.4	0.0	N/A
0.3	1.0	0.8	0.6
0.3	0.4	0.0	N/A
0.2	1.0	0.8	0.6
0.2	0.4	0.0	N/A
0.4	1.0	0.8	0.6
0.4	0.4	0.0	N/A

5. Discussion

The methodology proposed in this study to integrate Unsupervised Anomaly Detection (UAD) with Wafer Defect Inspection (WDI) showcases several notable advantages that enhance detection accuracy and efficiency in semiconductor manufacturing. Firstly, UAD effectively addresses the limitations of traditional classification methods that often overlook subtle and complex defect patterns, allowing for a more comprehensive identification of anomalies that deviate from normal operating conditions. By leveraging statistical principles alongside sophisticated machine learning techniques, such as Principal Component Analysis and Autoencoders, the approach adeptly reduces the dimensionality of defect data while preserving essential information, thus facilitating better anomaly recognition through the assessment of reconstruction errors and cluster distances. Furthermore, the incorporation of Gaussian Mixture Models provides an advanced framework for assessing the likelihood of defect observations, thereby enabling the identification of outliers within the distribution of detected defects. This multifaceted approach not only enhances the sensitivity of defect detection but also improves the overall reliability of the inspection process by optimizing the signal-to-noise ratio, thereby ensuring that detected anomalies are both accurate and meaningful. Ultimately, by adopting a systematic and integrative methodology, this research advances the field of semiconductor defect inspection, paving the way for higher quality and performance standards in the production of semiconductor devices while minimizing false positives and improving diagnostic capabilities.

While the integration of Unsupervised Anomaly Detection (UAD) in Wafer Defect Inspection (WDI) offers significant advantages, several limitations warrant consideration. Firstly, the reliance on probabilistic models, such as Gaussian Mixture Models (GMM), presupposes that defect distributions are consistent and Gaussian-like, which may not hold true in practice, leading to potential misclassification of defects. Additionally, clustering techniques like K-Means are highly sensitive to initial seed values and the presence of outliers, which can skew cluster centroids and impact the accuracy of anomaly detection. The effectiveness of methods like Principal Component Analysis (PCA) hinges on the assumption that the defects can be effectively represented in a lower-dimensional space, a notion that may compromise the fidelity of complex defect patterns. Moreover, while Autoencoders can enhance defect analysis, their performance is heavily dependent on the architecture and the amount of training data, which can limit their generalization to unseen defect types. The computation of reconstruction errors as indicators of anomaly presence may also yield false positives, particularly in scenarios where normal variation is misrepresented as an anomaly due to insufficiently trained models. Furthermore, the optimization of the signal-to-noise ratio (SNR) could be influenced by external factors, such as environmental variations during the manufacturing process, affecting the reliability of the defect detection outcomes. Lastly, the lack of labeled training data poses a challenge for validating the identified anomalies, which can undermine the interpretability and practical applicability of the methods employed. Consequently, while UAD presents a promising avenue for WDI, its limitations highlight the need for ongoing refinement and validation to ensure robust application in semiconductor manufacturing contexts.

6. Conclusion

Wafer defect inspection is crucial in ensuring the quality of semiconductor manufacturing, with the current methods relying heavily on supervised machine learning techniques. However, these methods face challenges in detecting unforeseen defects due to the requirement of labeled training data. To address this limitation, this research explores unsupervised anomaly detection methods. The proposed novel approach combines deep learning and anomaly detection algorithms to identify defects on wafers without the need for labeled data. By integrating diverse data sources and optimizing the anomaly detection process, our method aims to offer a more robust and efficient solution for wafer defect inspection. This work not only tackles the existing challenges in defect detection but also introduces innovative strategies to enhance inspection accuracy and reliability. Moving forward, future work could focus on expanding the dataset to further validate the performance of the proposed approach and explore the potential integration of real-time monitoring for defect detection, thus enhancing the scalability and applicability of the method in semiconductor manufacturing processes.

Funding

Not applicable

Author Contribution

Conceptualization, K. M. and A. S.; writing—original draft preparation, K. M. and P. W.; writing—review and editing, A. S. and P. W.; All of the authors read and agreed to the published the final manuscript.

Data Availability Statement

The data can be accessible upon request.

Conflict of Interest

The authors confirm that there are no conflict of interests.

Reference

[1] C. Yan et al., "Innovative wafer defect inspection mode: self-adaptive pattern to pattern inspection," Advanced Lithography, 2021.
[2] R. Ding et al., "Structural Design and Simulation of a Multi-Channel and Dual Working Condition Wafer Defect Inspection Prototype," Micromachines, 2021.
[3] J. Zhu et al., "Optical wafer defect inspection at the 10 nm technology node and beyond," International Journal of Extreme Manufacturing, 2021.
[4] C. Liu et al., "Aperture design for a dark-field wafer defect inspection system," Applied Optics, 2021.
[5] M. Qin et al., "Wafer Defect Inspection Optimization With Partial Coverage—A Numerical Approach," IEEE Transactions on Automation Science and Engineering, 2021.
[6] Z. Shi et al., "Wafer Defect Inspection Optimization: Models, Analysis and Algorithms," Social Science Research Network, 2019.
[7] H. He et al., "MambaAD: Exploring State Space Models for Multi-class Unsupervised Anomaly Detection," arXiv.org, 2021.
[8] D. A. Gudovskiy et al., "CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows," IEEE Workshop/Winter Conference on Applications of Computer Vision, 2021.
[9] J. Yu1 et al., "FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows," arXiv.org, 2021.
[10] M. Rahimi, Z. Wang, A. Shafieezadeh, D. Wood, and E. J. Kubatko, "Exploring passive and active metamodeling-based reliability analysis methods for soil slopes: a new approach to active training," International Journal of Geomechanics, 2020.

[11] X. Deng, L. Li, M. Enomoto, and Y. Kawano, ‘Continuously frequency-tuneable plasmonic structures for terahertz bio-sensing and spectroscopy’, Scientific reports, vol. 9, no. 1, p. 3498, 2019.

[12] X. Deng, M. Simanullang, and Y. Kawano, ‘Ge-core/a-si-shell nanowire-based field-effect transistor for sensitive terahertz detection’, in Photonics, MDPI, 2018, p. 13.

[13] X. Deng, S. Oda, and Y. Kawano, ‘Frequency selective, high transmission spiral terahertz plasmonic antennas’, Journal of Modeling and Simulation of Antennas and Propagation, vol. 2, pp. 1–6, 2016.

[14] Z. Wang and A. Shafieezadeh, ‘REAK: Reliability analysis through Error rate-based Adaptive Kriging’, Reliability Engineering & System Safety, vol. 182, pp. 33–45, Feb. 2019, doi: 10.1016/j.ress.2018.10.004.

[15] Z. Wang and A. Shafieezadeh, ‘ESC: an efficient error-based stopping criterion for kriging-based reliability analysis methods’, Struct Multidisc Optim, vol. 59, no. 5, Art. no. 5, May 2019, doi: 10.1007/s00158-018-2150-9.

[16] Z. Wang and A. Shafieezadeh, ‘Highly efficient Bayesian updating using metamodels: An adaptive Kriging-based approach’, Structural Safety, vol. 84, p. 101915, May 2020, doi: 10.1016/j.strusafe.2019.101915.

[17] Z. Wang and A. Shafieezadeh, ‘Confidence Intervals for Failure Probability Estimates in Adaptive Kriging-based Reliability Analysis’, Reliability Engineering & System Safety.

[18] Z. Wang and A. Shafieezadeh, ‘Real-time high-fidelity reliability updating with equality information using adaptive Kriging’, Reliability Engineering & System Safety, vol. 195, p. 106735, Mar. 2020, doi: 10.1016/j.ress.2019.106735.

[19] C. Zhang, Z. Wang, and A. Shafieezadeh, ‘Value of Information Analysis via Active Learning and Knowledge Sharing in Error-Controlled Adaptive Kriging’, IEEE Access, vol. 8, pp. 51021–51034, 2020, doi: 10.1109/ACCESS.2020.2980228.

[20] M. Rahimi, Z. Wang, A. Shafieezadeh, D. Wood, and E. J. Kubatko, ‘Exploring passive and active metamodeling-based reliability analysis methods for soil slopes: a new approach to active training’, International Journal of Geomechanics, vol. 20, no. 3, p. 04020009, 2020.