Scientific Reports

https://doi.org/10.1038/s41598-026-46947-3

Article in Press

Self-supervised multi-resolution learning for label-agnostic morphology representation and clustering of semiconductor thin-film SEM defects

Received: 5 March 2026

Accepted: 28 March 2026

Published online: 03 April 2026

Cite this article as: Krishnamoorthy U., Chan C.K., Sonawane C. *et al.* Self-supervised multi-resolution learning for label-agnostic morphology representation and clustering of semiconductor thin-film SEM defects. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-46947-3

Umapathi Krishnamoorthy, Choon Kit Chan, Chandrakant Sonawane, Amol Vedpathak, Subhav Singh & Deekshant Varsheny

Self-supervised Multi-resolution Learning for Label-agnostic Morphology Representation and Clustering of Semiconductor Thin-film SEM Defects

Umapathi Krishnamoorthy*[1], Choon Kit Chan[2], Chandrakant Sonawane[3], Amol Vedpathak[3*], Subhav Singh[45], Deekshant Varsheny[67]

[1]Department of Electronics and Communication Engineering, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India.

[2] Department of Mechanical Engineering, INTI International University, Malaysia.

[3] Department of Mechanical Engineering, Symbiosis International University, Pune.

[4] Division of research and development, lovely professional University, Phagwara, Punjab, India

[5] Center for innovation and inclusive research, Sharda University, Greater Noida, Uttar Pradesh, India.

[6] Centre of Research Impact and Outcome, Chitkara University, Rajpura-140417, Punjab, India

[7]Noida Institute of Engineering and Technology, 19, Knowledge Park-II, Institutional Area, Greater Noida (UP) -201324, India.

*Corresponding authors: umapathi.uit@gmail, amol.vedpathak@scnn.edu.in

Abstract

Scanning Electron Microscope (SEM) image analysis plays a vital role in semiconductor thin-film characterisation. In particular, defect detection and classification are performed using SEM images. However, conventional methods rely on labelled datasets or handcrafted features for classification, which limits their generalisation in real-world industrial inspection settings. The present work proposes a self-supervised multi-resolution learning framework for label-agnostic morphology representation learning and clustering of semiconductor thin-film defects. It uses a SEM image dataset (4591 images) obtained from industrial wafer

inspection. The framework starts with image pre-processing to remove acquisition artifacts. It employs a multi-resolution image pyramid for capturing surface morphologies at fine, intermediate, and coarse spatial scales. A shared-weight convolutional encoder that ensures alignment between embeddings across the three resolutions is trained (on an unlabelled dataset) and utilised for unsupervised defect morphology grouping. The framework learns morphological representations without defect labels. However, defect labels are used only for post-hoc evaluation by normalized mutual information (NMI) and visualization. Intrinsic clustering metrics and low-dimensional visualization are used to assess the algorithm's efficacy. Experimental results reveal that the proposed method, gray level co-occurrence matrix (GLCM), local binary patterns (LBP), wavelet-based features, and principal component analysis (PCA) on raw pixels obtained a silhouette score of 0.50, 0.43, 0.31, 0.45, and 0.22, respectively. While normalized mutual information (NMI) values remained moderate across the models. These results reflect the label-agnostic nature of the proposed SSL framework. Further, UMAP and t-SNE visualizations confirm the coherent manifold structure and the effectiveness of morphology-driven grouping. These results demonstrate the robust, scale-invariant quality of the proposed self-supervised multi-resolution learning framework for defect clustering.

Keywords: defect clustering, label-agnostic learning, self-supervised learning, Process Innovation, SEM analysis.

1. Introduction

Advancements in semiconductor technologies have enabled the realisation of electronic devices at the nanoscale and whole electronic systems in thin films [1]. However, localised microstructural irregularities, surface roughness, voids, and grain boundaries affect the electrical characteristics, device stability, and its reliability [2]. Thus, surface morphology characterisations and defect detections play a critical role in yield monitoring and process control in semiconductor manufacturing. In this context, Scanning Electron Microscopy (SEM) images, which provide

rich structural information and high spatial resolution, are widely used for the identification of surface defects and process-induced variations [2]. But as data volumes increase, manual inspections have become impractical. This led to the search for automated SEM image interpretation. Image processing and data-driven approaches play a vital role in this context [3].

Conventional automated SEM image analysis relies on handcrafted texture descriptors for identifying defects in semiconductor thin films [4]. Specifically, (i) Gray-Level Co-occurrence Matrix (GLCM) employs statistical texture features, (ii) Local Binary Patterns (LBP) encode intensity transitions, and (iii) Wavelet-based methods use frequency domain features for detecting surface defects [5][6]. Handcrafted features are sensitive to imaging conditions and magnification and cannot generalise across materials and process variations. And thus, they become insufficient to capture complex, non-linear surface morphologies present in modern semiconductor thin films.

Artificial intelligence (AI)- based deep learning methods, such as convolutional neural networks (CNNs), are a boon for image analysis. CNNs can learn hierarchical representations directly from microscopy images. Supervised CNN algorithms have been shown to be efficient for semiconductor surface quality assessment and defect classification [7]. However, these methods rely on labelled datasets for learning. Obtaining high-quality defect labels is often costly and challenging in semiconductor thin-film defect detection [8]. This challenge roots from the manufacturing process complexity and variations. Specifically, (i) highly imbalanced defect distributions and (ii) varying defect taxonomies across fabrication nodes and processes hinder model performance [9]. These factors limit the scalability and robustness of supervised learning algorithms, which are demanding requirements for real-world industrial SEM analysis.

Self-supervised learning (SSL) methods can address these challenges [10]. SSLs have defined tasks, otherwise called surrogate objectives, that enable them to capture ground truth from the given inputs. They use this learned knowledge to capture features from unlabelled data and perform

clustering [11]. SSL models are highly beneficial when obtaining a labelled dataset is a constraint. However, existing SSL approaches often use single-scale image representations [12]. To harvest the benefits of SSL and improve clustering quality, this research proposes an SSL method that employs multi-scale image resolutions. This method employs nanoscale, grain-scale, and mesoscale features to cluster defects from SEM images. These references reveal two critical research gaps, and they are,

1. The limitations existing with the use of conventional handcrafted features and labelled datasets for the semiconductor thin film defect classification task.
2. The lack of scale-aware, self-supervised representation learning for the SEM image defect clustering task.

These limitations are addressed in this research work, and its objectives are,

- To propose a self-supervised multi-resolution learning framework for semiconductor defect clustering from nanoscale morphologies present in SEM images. It employs a domain-aware preprocessing pipeline and a multi-resolution image pyramid to capture morphology across spatial scales. By defining an encoder with cross-scale consistency loss and stop-gradient regularization, the SSL framework enables a scale-invariant morphology learning from unlabelled training data.
- To categorise the learned features employing the K-means clustering algorithm and evaluate the performance of the proposed SSL learning method in comparison with classical texture-based and principal component analysis (PCA)-based methods.
- To visualize the clustering results in terms of Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE) to validate the morphology-aware grouping of the proposed method.

With this brief introduction, the rest of the manuscript is organised as follows: Section 2 presents a review of the state-of-the-art literature on semiconductor defect classification and clustering, while Section 3

presents the details of the dataset and the theory behind the algorithmic methods used in the proposed framework. Section 4 presents the proposed framework, and Section 5 details the experimental setup, results obtained, and discussion of interpretations. Section 6 concludes the article.

2. Related Work

Wafer surface defect detection methods are broadly classified into three classes: image processing, machine learning, and deep learning [13]. Image processing-based methods employ handcrafted features for detecting defect classes, while learning-based methods learn from the available dataset. The efficacy of the learning-based method roots from its adaptability to the intrinsic features present in the input dataset. CNN-based deep learning models are efficient in classifying semiconductor wafer defects. For instance, an Xception model was employed for wafer characterisation, and defect size was extracted by the class activation mapping method in an experiment. And results demonstrated a classification accuracy of 96.9% [14]. Though CNNs are effective at classifying semiconductor wafer defects, they are computationally intensive and thus effective alternatives are in high demand. For instance, a two-step approach combining light-weight SqueezeNet CNN and classical computer vision techniques is studied. And such a model achieved a classification efficiency of 99.356%, which is on par with a high-complexity ResNet-50 model that achieved 99.44%. However, this model consumed 80% less time compared to ResNet-50 [15]. Another issue frequently occurring in wafer defect detection is the missing detections and multiple box detections. These issues are addressed by employing a faster recurrent-CNN (RCNN) in an experiment. Such an experiment achieved an average precision of 87.5% and a detection speed of 0.26 seconds per image [16]. However, introducing background subtraction together with Faster R-CNN increased the mean average precision by 5.2% [17]. Further, to achieve faster defect predictions, Inception modules and skip-connection-based learning algorithms were employed. This method achieved 59% faster inference than the baseline [18]. In another experiment, a combined CNN-DNN (deep neural network) model achieved

an accuracy of 99.45% in classifying semiconductor wafer defects [19]. Despite the improved detection performance of learning-based algorithms, their use for industrial-scale defect detection is often challenging. This is because of (i) variability of defect scales and types, (ii) variations in background, and (iii) imbalanced training data. These issues were addressed by employing an augmentation technique named SegMix in an experiment. These augmentations were employed to diversify the training data, with the objective of improving model robustness. Further, this method employed an attention-based classifier for improving classification performance [9]. To address input data variability and leverage synergies across algorithms, ensemble-based classifiers were employed for defect detection. For instance, a soft voting classifier combining SVM, logistic regression, and random forests achieved 98% classification accuracy when classifying defects into four classes [20]. Further, interpretable models were developed for classifying wafer defects. For instance, local interpretable model-agnostic explanations (LIME) based interpretations were provided for a CNN-based wafer defect classification algorithm [21]. To filter out new defect categories before reaching the classifier, a semiautomatic algorithm was proposed in the literature. This framework quantifies wafer map uncertainty and identifies divert images with higher uncertainty for manual inspection. This avoids missing defects [22]. In addition, semi-supervised learning methods that employ dual-head CNNs were investigated to address class imbalance. Such a method achieved 98.2% accuracy with a ResNet-50 backbone [23]. The present study employs a self-supervised learning method for improving model prediction robustness and generalisation.

3. Materials and Methods

3.1. Dataset Description

This study utilises a publicly available dataset, the Carinthia SEM Defect Dataset [24]. This dataset contains SEM images acquired from a real industrial semiconductor wafer inspection process. The dataset is available on Zenodo, an open-access research data repository operated by CERN under the OpenAIRE program. The dataset was released (February

27, 2024) by researchers affiliated with Infineon Technologies Austria AG, Infineon Technologies Dresden GmbH & Co. KG, and KAI GmbH, as part of the European research project AIMS5.0, supported by the Chips Joint Undertaking. The dataset does not contain personal, medical, or sensitive information and is available for research and academic use under the licensing terms specified on the Zenodo platform.

The SEM images in the dataset correspond to defects identified on one production layer of an unstructured semiconductor wafer. The dataset preserves realistic variabilities in surface morphologies, imaging conditions, and defect appearances and thus represents the semiconductor manufacturing environment. It includes a total of 4591 grayscale SEM images in standard jpg format. It comes with structured metadata provided in a CSV file that contains image path, file name, and defect labels. The dataset directory structure follows a clear, reproducible organization, separating image files from metadata. The dataset labels are encoded into numerical values from 1 through 6. The majority of the images were non-defective and fell under class '6'. The dataset is imbalanced across the classes. This reflects the real-world industrial scenario and poses challenges for the supervised learning approach. The present study utilises a self-supervised framework for label-agnostic morphology representation learning and clustering and thus, defect labels are strictly not used for model training. However, numerical labels are employed for post-hoc visualization and quantitative evaluation of the proposed method.

3.2 Image pre-processing

As the images are captured from SEM under automated inspection settings, defects occupy the centre of the image. However, images contain instrument-induced artifacts like (i) narrow black borders on one side of the image frame and (ii) occasional framing patterns from electron beam alignment procedures. The presence of such artifacts makes data-driven models learn spurious patterns and thus, data pre-processing becomes necessary. The image pre-processing stage is designed with two objectives, and they are (i) removing instrument-induced artifacts, (ii)

standardising spatial resolution and intensity characteristics for fair and stable representation learning. This facilitates an artifact-aware pre-processing in the context of semiconductor defect detection from SEM images.

3.3 Self-Supervised Learning Mathematical modelling

Self-supervised learning enables a model to make predictions and clustering of an unlabelled dataset. This can be explained mathematically as follows. Let, $X = \{x_i\}_{i=1}^{N}$ denote a set of unlabeled SEM images of semiconductor wafer thin films. The objective of this work is to learn a meaningful representation function $f_\theta: X \rightarrow R^d$, that maps each SEM image to a low-dimensional embedding capturing the intrinsic surface morphology. This enables training the model without using defect labels. Here, the hierarchical nature of surface morphology is considered, and each SEM image x_i is represented by a set of multi-resolution views, $\{x_i^{(0)}, x_i^{(1)}, x_i^{(2)}\}$, corresponding to fine-, intermediate-, and coarse-scale representations. These views are derived from the same physical surface region and are used for training the algorithm. This enables two-fold benefits, (i) facilitates a label-free learning and (ii) enables the model to be robust to different resolutions.

3.4 Model evaluation & Quantitative assessment

Being an SSL framework, the evaluation strategy focuses on the intrinsic structure and separability of the learned embeddings rather than model accuracy. To assess the separability of the learned representations, K-means clustering is applied to the extracted embeddings. The number of clusters is made equal to the number of defect categories available in the dataset. Clustering is performed without labels, and resulting cluster assignments are used to compute intrinsic clustering quality metrics and to assess post-hoc alignment with defect labels. Two complementary metrics were used to evaluate clustering performance: the silhouette score and normalised mutual information (NMI).

Silhouette score: Silhouette score measures the compactness and separation of clusters based on Euclidean distance in the embedding space (as standard). This parameter helps assess clustering quality without

relying on ground-truth labels. Higher scores indicate better-defined cluster structures.

Normalized Mutual Information (NMI): NMI quantifies the agreement between known defect labels and cluster assignments. This enables understanding how learned morphology representations relate to existing defect categorisations. However, labels are not used for training. Hence, NMI serves only as a post-hoc analysis tool and does not reflect a supervised learning performance.

3.5 Model Interpretability & Qualitative Assessment

Model interpretability enables understanding the reason behind an AI prediction. Specifically, they reveal the structure of the learned feature space, or the significant features on which the model's predictions are based. In this work, the structure of the learned feature space is visualised by applying dimensionality reduction techniques to the extracted embeddings. Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE) are used for visualisation; however, they do not affect the quantitative evaluation metrics. But they enable a qualitative assessment of the model results.

- UMAP is used for visualising the global structure of the embedding space while preserving neighborhood relationships.
- t-SNE is used as a complementary visualisation method to examine local clustering behavior.

4. Proposed Methodology

The flow diagram representation of the proposed SSL framework is shown in Figure 1. In Figure 1, the block (a) represents the overall block diagram, while blocks labelled (b) to (e) represent the detailed operations/functions/layers that constitute the abstract blocks in (a). It could be inferred from Figure 1 that, once the grayscale SEM images are loaded, the first step in the SSL pipeline is pre-processing. This framework employs artifact-aware preprocessing before model training. In the semiconductor thin-film defect-detection scenario, SEM tool alignment and acquisition framing may introduce black borders along the image edges. These borders are not related to semiconductor defects, and their

presence may misguide training. Thus, a fixed-width border region is cropped from each image. Cropping eliminates acquisition-related artifacts while preserving the central region of the image, where the defect predominantly resides. Cropped images were normalised to a fixed base resolution of 512 x 512 using bicubic interpolation. This step helps preserve nanoscale textures and ensures consistent spatial representation during multi-resolution construction. Pixel intensities are normalised to a common range in order to reduce variability caused by contrast differences and detector response.

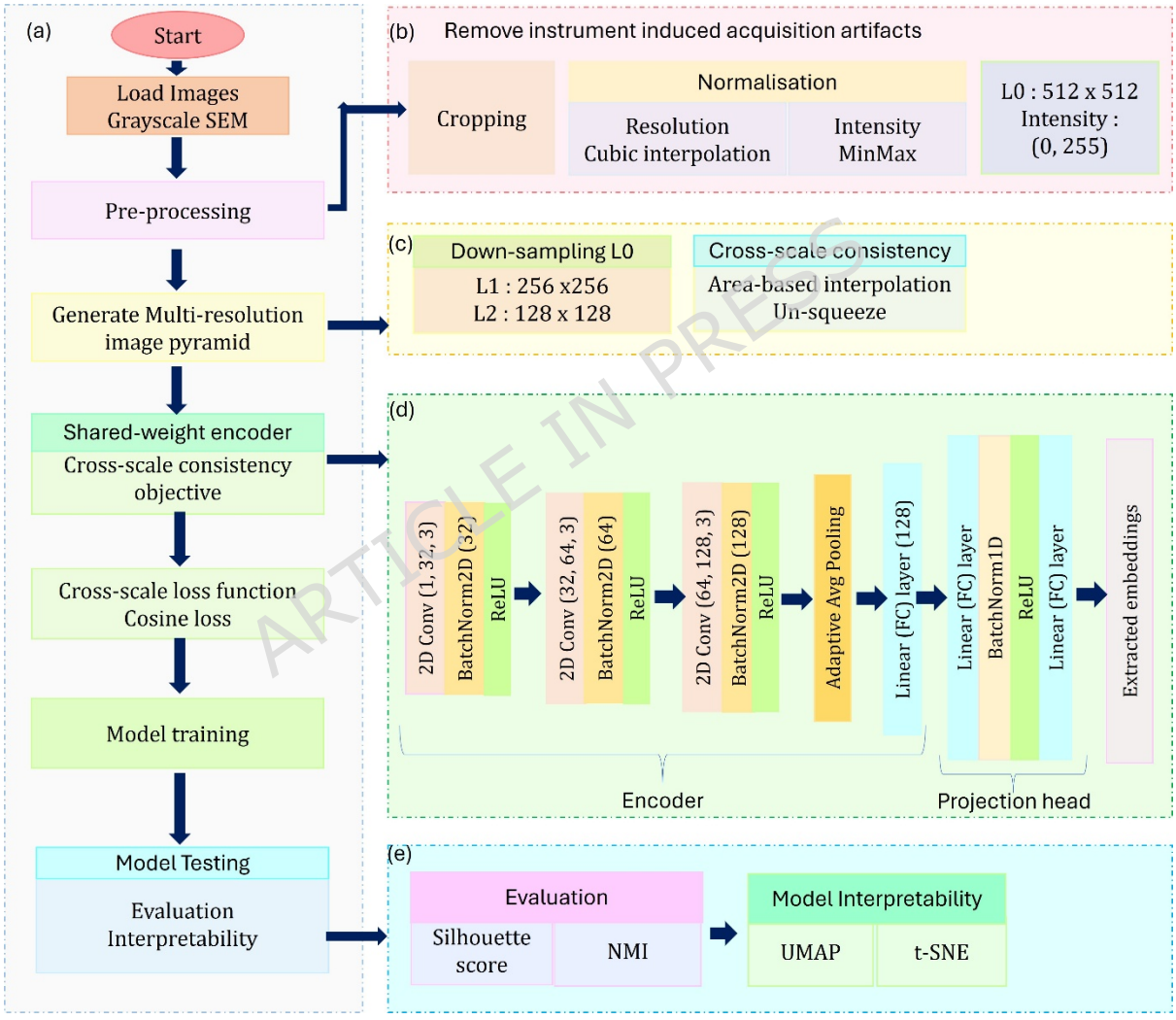

Figure 1. Architecture of the proposed self-supervised learning framework, (a) blocks in the SSL framework, (b) pre-processing operations, (c) multi-resolution image pyramid generation functions, (d) encoder architecture, (e) model evaluation and interpretation functions.

Multi-Resolution Image Pyramid Construction: Semiconductor thin films exhibit a hierarchical surface morphology spanning multiple spatial scales. This makes single-resolution representations insufficient while capturing defects that span this hierarchical morphology. Thus, a multi-resolution image pyramid is constructed for each SEM image to explicitly encode morphologies at different spatial scales. Specifically, a multi-resolution image pyramid is generated with three scales. This includes a fine-scale nano texture that captures local surface variations, an intermediate-scale that reveals structural organisations, and a coarse-scale that represents surface patterns. The pyramid levels are named L0, L1, and L2 and correspond to 512x512, 256x256, and 128x128 resolutions, respectively. This image pyramid is constructed by downsampling with area-based interpolation to preserve cross-scale structural consistency while reducing spatial resolution. The pre-processing and multi-resolution representation strategies ensure that learned representations reflect intrinsic surface morphologies, free of resolution and artifact-induced biases.

Multi-Resolution Self-Supervised Learning Framework

The proposed SSL framework enforces cross-scale consistency between embeddings obtained from different resolutions of the same SEM image. It processes multi-resolution (k) representations $x_i^{(k)}$, by a shared-weight encoder network. And enforces cross-scale consistency during training by mapping all resolution levels into a common latent space. Such a design helps preserve resolution-dependent morphological information despite explicitly defining scale invariance.

Encoder Architecture: The encoder is a convolutional neural network (CNN) composed of sequential convolutional blocks, each consisting of a convolution layer, batch normalization, and non-linear activation. Spatial dimensionality is reduced progressively by strided convolutions and adaptive average pooling. The encoder output is a 128-dimensional embedding vector representing the base morphological representation. The same encoder parameters are maintained across all resolution levels to ensure consistent feature extraction across scales. Further, shared-

weight encoder design promotes model robustness by learning meaningful, scale-consistent morphologies and preventing the model from learning resolution-specific shortcuts. The encoder architecture consists of three 2D CNN blocks, followed by an adaptive average pooling layer and a linear fully connected layer. Each 2D CNN block consists of a CNN layer, a batch normalisation layer, and a ReLU activation. The CNN layers employ a 3x3 kernel, with a stride of 2 and a padding of 1. The input CNN layer receives a grayscale image and produces a 32-channel feature map. The subsequent CNN blocks increase the feature depth by a factor of 2, producing a 128-channel output feature map at the third block.

Projection Head for Representation Learning: A lightweight projection head follows the encoder output. It consists of a small multilayer perceptron layer with batch normalisation and non-linear activation. This projection head serves two purposes: (i) it decouples the representation space from the encoder embeddings, (ii) it stabilises training and improves representation quality. The projection head is used only for training. The projection head has 128 dimensions and consists of a single block with a linear fully connected layer, batch normalisation layer, ReLU activation, and an output layer.

Cross-Scale Consistency Objective: To learn scale-invariant morphological representations, the model is trained using a cross-scale consistency loss function. This function aligns embeddings obtained from different resolution views of the same SEM image. Here, if $z_i^{(k)} = f_\theta(x_i^{(k)})$, denotes the encoder output for the resolution level k, and $p_i^{(k)} = g(z_i^{(k)})$ denotes the corresponding projection head output, then a cosine similarity-based loss function is employed to enforce consistency between embeddings across resolution pairs. Further, a stop-gradient operation is applied asymmetrically to prevent representational collapse. This function ensures one branch of the network provides a stable target while the other branch is optimized. The overall loss is computed by averaging pair-wise, cross-scale losses between selected resolution combinations. This encourages embeddings from different scales to converge to a common morphology-aware representation.

Avoidance of representational collapse: A challenging task in SSL is the risk of representational collapse when all inputs are mapped to identical embeddings. However, this challenge is handled in this framework by an asymmetric stop-gradient operation. This operation involves decoupling the projection head from the encoder and using a cosine-similarity-based consistency objective. Further, the proposed design's stable training dynamics and meaningful feature learning are confirmed by the non-zero, smoothly converging training loss.

5. Experimental Results and Discussion

Experiments were conducted using a Python-based deep learning framework. Model implementation and training were performed in a cloud-based environment with GPU acceleration when available. Image processing, model training, and evaluation were performed using standard scientific computing and machine learning tools. All experiments used fixed random seeds where applicable, and a deterministic preprocessing step was applied consistently across the dataset to ensure reproducibility.

Training strategy: A self-supervised learning framework was trained using the unlabeled SEM images described in section 3. There are no data labels, defect annotations, or prior knowledge of defect categories used at any stage of training. Each training sample has three resolution views derived from the same SEM image. These views are derived from the multi-resolution pyramid described in section 4. Every iteration processes all resolution views of an image, and the cross-scale consistency loss is optimized using the Adam optimizer. A fixed learning rate of ‘0.001’ was used throughout training. The batch size was selected based on available computational resources, and a moderate value of ‘32’ was used to balance convergence stability and memory constraints. Training is performed over a limited number of epochs (15), sufficient to achieve convergence of the self-supervised objective without overfitting. No early stopping or label-based validation was employed as the learning process is entirely self-supervised.

Output representation: After training convergence, the projection head is removed, and the 128-dimensional morphology embeddings from SEM

images are extracted. The extracted embedding serves as the basis for all subsequent analysis, including visualisation, clustering, and comparison with classical feature-extraction methods. To provide consistent evaluation and capture fine-scale surface morphologies, embeddings were extracted from the highest-resolution input (Level L0).

The performance of the proposed method was evaluated through both quantitative clustering metrics and qualitative visualisations. The SSL model is analysed, focusing on (i) intrinsic structure of learned feature space, (ii) cluster separability and compactness, (iii) post-hoc alignment with defect categories, and (iv) comparison with classical methods.

5.1 Quantitative Clustering Performance

Two quantitative metrics, namely, silhouette score (label-independent) and NMI (post-hoc with labels), are used to assess the proposed model's performance relative to other classical feature extraction models such as GLCM, LBP, wavelet decomposition, and PCA, as shown in Table 1. Specifically, (i) GLCM is extracted based on five features - contrast, dissimilarity, homogeneity, energy and correlation, (ii) LBP features are extracted with a radius of '1' and '8' neighbours, (iii) two-level wavelet decomposition is employed, and (iv) PCA that reduces raw 512×512 pixels to 128 dimensions were chosen for a fair comparison. The same input images with clustering parameters were chosen across these models to ensure a fair and consistent evaluation.

Table 1. Clustering performance comparison for a cluster size of six in the embedding space

Si. No.	Method	Silhouette score	NMI
1	GLCM	0.428901	0.287785
2	LBP	0.312685	0.289118
3	Wavelet method	0.449926	0.210032
4	PCA (raw pixels)	0.222131	0.281617
5	Proposed SSL	0.501581	0.203185

Table 1 reveals that the proposed SSL framework achieves the highest silhouette score of 0.50 when compared to other classical feature-based defect clustering methods. This directly depicts the intrinsic compactness of the cluster and the separation of classes in the learned feature space. Classical texture-based methods like GLCM and wavelet extraction exhibit moderate performance, suggesting that their performance is comparatively lower due to the use of handcrafted features. Further, the poor performance of the linear variance-based PCA method reflects its inability to capture complex SEM morphologies. The improved performance with the proposed SSL method reveals the benefit of learning hierarchical, non-linear features directly from data. Further, enforcing cross-scale consistency enabled the model to capture scale-invariant morphological patterns more effectively than single-scale and handcrafted approaches.

NMI values obtained are modest across all methods, including the proposed method. However, slightly lower NMI of the proposed method (0.203) versus classical texture-based baseline methods (0.21–0.29) is consistent with its label-agnostic objective, prioritising morphological coherence over exact label replication. This outcome coincides with the theoretical expected value and does not indicate poor performance. Significant points that justify considering morphologies instead of direct defect labels are (i) defect labels are discrete and are process dependent, using them for classification often leads to reduced model generalisation capability, (ii) classical texture features align strongly with defect annotations as they are sensitive to local intensity patterns and edge artifacts. Thus, the lower NMI value is a direct reflection of the label-agnostic nature of the learned representations and supports the claim that the model captures broader surface characteristics beyond predefined defect classes.

5.2 Embedding Visualization by UMAP

The UMAP visualisation of the clustering results of the proposed SSL method is shown in Figure 2. Figure 2 shows that UMAP projections reveal a well-structured, continuous manifold. It could be observed from the

UMAP that there exists (i) a clear separation between several morphological regimes (indicated by different colors), (ii) a smooth transition between clusters rather than abrupt boundaries, (iii) the absence of random scattering, indicating non-degenerated embeddings. In addition, more samples of a particular color reveal class imbalances. However, the slight overlap between different-colored samples highlights the model's efficacy in handling class imbalance without bias. Further, the organisation of samples into coherent regions indicates the model's ability to capture meaningful low-dimensional representations from SEM surface morphology rather than memorising noises or acquisition artifacts.

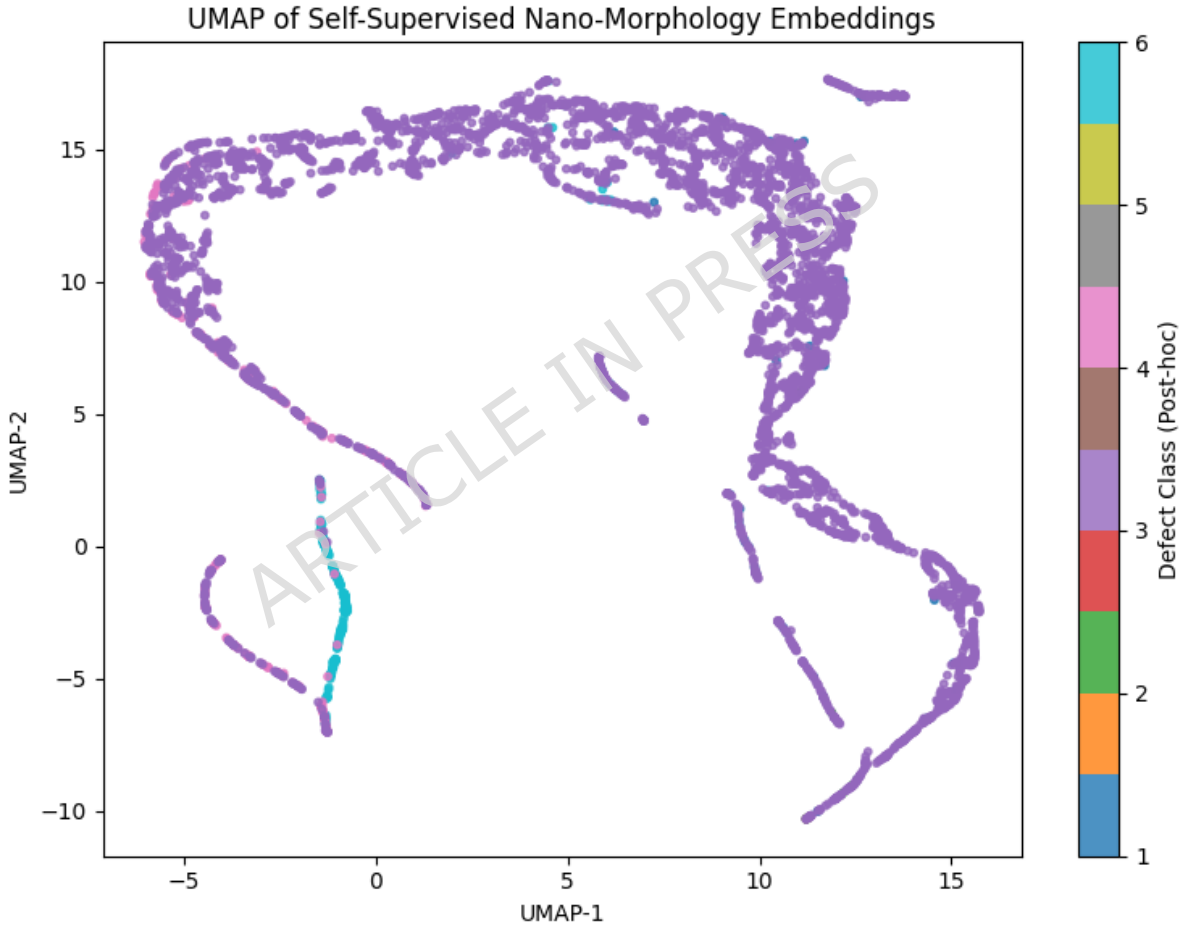

Figure 2. Illustrates the UMAP projection of the self-supervised nano-morphology embeddings learned from SEM images.

5.3 t-SNE Visualization

The t-SNE plots obtained for the proposed SSL framework are shown in Figure 3. t-SNE visualisations highlight local neighborhood consistency within the learned embedding space. Samples with visual similarity tend

to cluster together in a low-dimensional projection. It could be inferred from Figure 3 that the proposed method (i) improved local compactness, (ii) shows elongated and curved regions. These observations coincide with the objective of morphology-aware representations without supervision.

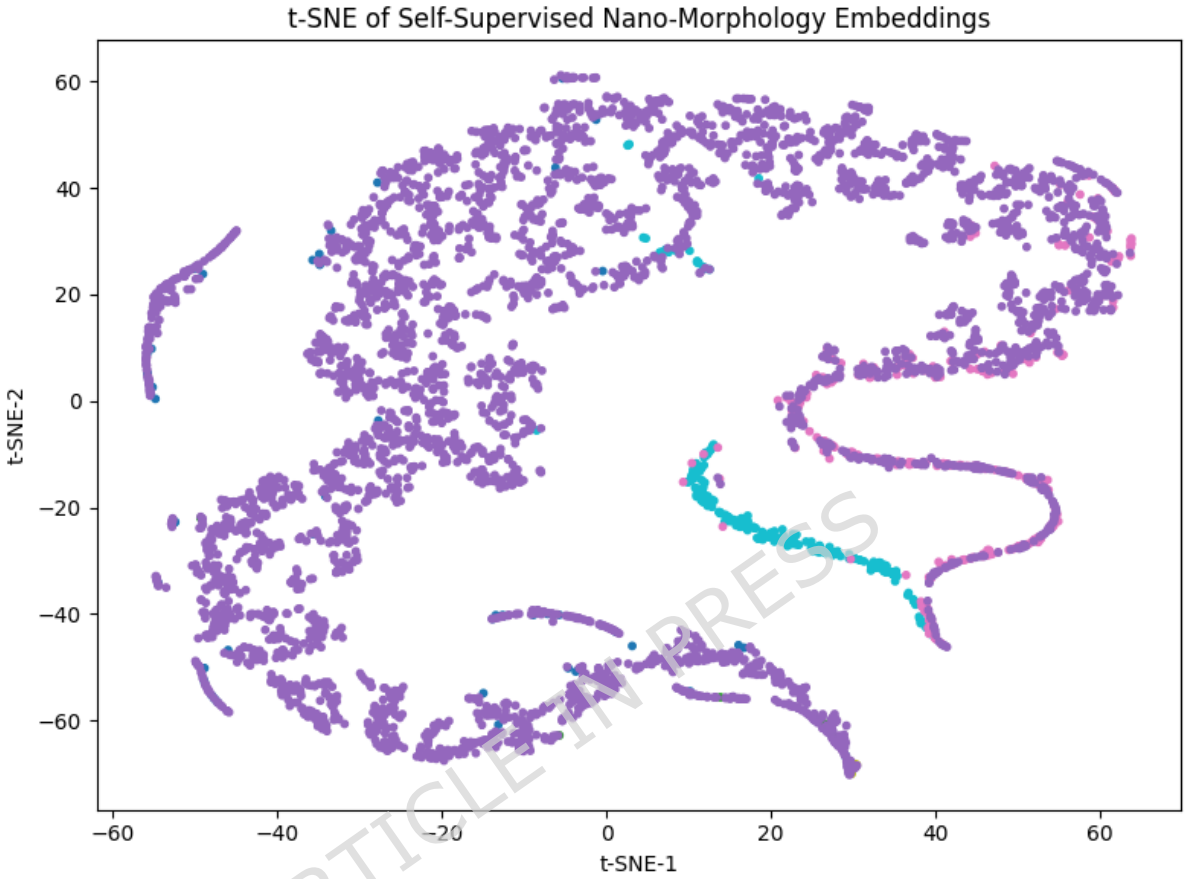

Figure 3. t-SNE visualization of self-supervised nano-morphology embeddings learned from SEM images.

5.4 Interpretation of findings

A significant contribution of the work is the explicit incorporation of multi-resolution information within the SSL framework. The outcome of the SSL model is improved cluster separability, as reflected in the highest silhouette score compared to classical methods. This depicts the model's ability to capture structural irregularities across spatial resolutions resulting from the cross-scale consistency enforcement. Classical methods demonstrated a comparatively lower silhouette score. This depicts that manual feature design is sensitive to imaging conditions, magnification, and noise. Further, the proposed SSL model learns from internal consistency and morphology rather than label agreement, whereas

conventional method often concentrates on classifying defects based on labels. This is reflected in the lower NMI value for the proposed SSL method compared to classical methods. SSL models do not use label frequencies during training, so the model's predictions are not affected by class imbalance. This is depicted from the UMAP and t-SNE distributions. In Figure 2, each point represents a SEM image mapped to a 2D latent space, and colors indicate the defect class labels used only for post-hoc analysis. Figure 2 shows that the UMAP projections exhibit a continuous, non-random manifold structure rather than a discrete, tightly separated cluster. This reveals variations in the surface morphologies of SEM images. This is consistent with the nature of semiconductor thin films, which exhibit a smooth rather than an abrupt change in morphology. Further, the absence of scattered or collapsed points confirms that the representations are stable and non-degenerate, and that the SSL model captures them. Elongated and curved regions visible across the embedding space reveal that the model organizes SEM images according to morphological similarities, such as texture density, surface roughness, or grain organisation, rather than strict defect labels. However, this behaviour supports the fact that the SSL framework concentrates on morphological representations rather than labels. Although no specific labels are used for training, localized regions show higher concentrations of a particular color, indicating that certain defects correspond to a distinct morphological pattern. However, label overlapping is observed across most regions, which indicates that (i) different defect categories have similar surface morphologies, (ii) the model is not optimized to separate predefined defect classes. This observation is supported by quantitative analysis results. Specifically, a higher silhouette score indicates good intrinsic separability, and a moderate NMI indicates non-random alignment with defect labels.

The t-SNE plot emphasizes local similarity relationships. The plot obtained for the proposed SSL predictions shows that morphologically similar SEM images form a compact local neighborhood. This indicates the preservation of fine-grain morphology characteristics at the local level.

Further, elongated structures represent a gradual transition between morphological states. The cyan-colored trajectory reveals a morphological signature of a specific defect category. However, a significant overlap in colors can be seen in the plot. This signifies that multiple defect categories share a common surface morphology. At the same time, the distinction in classes reveals that (i) the model predictions are unbiased in nature, and (ii) defect labels do not uniquely define morphologies. The absence of sharp cluster boundaries reveals that (i) the model is trained without labels in an unsupervised manner, (ii) the model is not overfitted and is label-agnostic in nature. These results are supported by a moderate value of NMI. The UMAP and t-SNE plots obtained for various classical methods exhibit clear boundaries between classes, revealing their label-dependent classification nature. The UMAP and t-SNE plots for classical methods are presented in Figures S1 to S8 in the supplementary file.

A higher silhouette score and moderate NMI indicate that the proposed SSL framework prioritises morphological coherence and physical structure over artificial label separation. In real-time industrial defect detection, datasets are discrete, process-defined, and continuously vary across spatial scales. Under this scenario, strict adherence to labels reduces the generalisation ability. In addition to defect detection, this property is useful for anomaly detection, process drift analysis, and exploratory inspection where predefined defect labels are incomplete, noisy, or unavailable. The visualisations, along with metrics, confirm that the proposed multi-resolution SSL framework learns morphology-aware scale-invariant representations from unlabelled SEM images. At the same time, preserve both global and local similarities without relying on labels.

5.5 Significance of the proposed method

The experimental results demonstrate that the proposed method holds numerous benefits for defect detection, which include,

- The proposed SSL framework predictions are robust to resolution changes, hierarchical surface structures, noises, and artifacts. Further, the SSL predictions are label-agnostic in nature and thus,

could be applied across various thin film fabrication processes for defect detection. These results are supported by NMI scores.

- The model demonstrated non-degenerative learning with stable convergence. This is supported by meaningful clustering structure and non-zero silhouette scores.
- In addition, the model is robust to class imbalances as no label information is used during training and clustering. These results are supported by UMAP and t-SNE visualisations.
- The novel design considerations that contributed to enhanced performance over classical methods are the multi-resolution representations, enforcing cross-consistency learning across the three resolution views.

From an industrial perspective, the proposed framework offers several benefits: (i) minimisation of annotation efforts, as the model could make predictions on unlabeled data, (ii) improved scalability as the model is generalised and could be deployed across different processes and materials, and (iii) robust interpretations make it suitable for exploratory analysis and monitoring tasks. The learned embeddings could be utilized for applications like defect, anomaly detection, unsupervised process clustering, and early-stage yield monitoring.

5.6 Limitations and Future Extensions

Despite its advantages, numerous limitations remain that warrant future research interventions. This includes,

- The model evaluations were carried out using a single SEM dataset acquired from one production layer.
- The present framework does not incorporate explicit domain-specific physical constraints.
- Quantitative evaluations were based on clustering metrics rather than evaluations on task performance.

However, these limitations do not undermine the validity of the results but highlight opportunities for further investigation. In addition, the present work can be extended to include (i) predictions from multi-modal microscopy data such as AFM or EDX, (ii) integrating with physics-

informed constraints into self-supervised objectives, (iii) evaluations across multiple semiconductor processes and materials, and (iv) extension to anomaly detection and process monitoring tasks.

6. Conclusion

This study presented a self-supervised, multi-resolution learning framework for label-agnostic morphology representation learning from semiconductor thin film SEM images. The framework includes a domain-aware preprocessing pipeline and defines a multi-resolution image pyramid for learning fine-grained and coarse morphologies from semiconductor thin films. The framework employed a shared-weight encoder along with forced cross-scale consistency to learn scale-invariant embeddings from SEM images. Further, stable, non-degenerate learning is achieved through the timely introduction of an asymmetric loss function. The model is evaluated for its clustering ability employing qualitative visualizations (UMAP and t-SNE) and quantitative metrics (silhouette score and NMI). Experimental results support the superiority of the proposed SSL framework for learning and clustering morphological representations in a label-agnostic manner, compared to classical methods. Specifically, higher silhouette scores (0.5) and lower NMI value (0.203) compared to classical methods, along with smoothly transitioning, slightly overlapped classes in visualisations, reveal that the model prioritizes morphological coherence rather than exact label replication. This enables the proposed SSL framework to be effectively adapted for industrial-scale SEM morphology analysis in a semiconductor manufacturing environment. Further, the model can be rapidly adapted for exploratory analysis and process monitoring tasks.

Declarations:

Conflict of interests:

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Funding: No funding was received for conducting this study.

Acknowledgement: NIL

Data Availability statement:

The data that support the findings of this study are openly available in [zenodo] at https://zenodo.org/records/10715190 reference number [24].

Author Contributions:

U.K. conceived the study and contributed to the experimental design. C.K.C. assisted in methodology development and technical guidance. C.S. contributed to data analysis and interpretation of results. U.K and A.V. supervised the research work and coordinated the overall project. S.S. contributed to software work and validation of results. D.V. assisted in data collection, visualization, and manuscript preparation. All authors reviewed and approved the final manuscript.

References

[1] Ho, S., & Ejikeme, E. I. (2025). Investigation of nanostructured thin films using scanning electron microscopy technique. *International Journal of Engineering Trends and Technology*, 73(2), Article P103. https://doi.org/10.14445/22315381/IJETT-V73I2P103

[2] Sivaraj, S., Rathanasamy, R., Kaliyannan, G. V., Panchal, H., Jawad Alrubaie, A., Musa Jaber, M., Said, Z., & Memon, S. (2022). A Comprehensive Review on Current Performance, Challenges and Progress in Thin-Film Solar Cells. *Energies*, *15*(22), 8688. https://doi.org/10.3390/en15228688

[3] Ren, Z., Fang, F., Yan, N. *et al.* State of the Art in Defect Detection Based on Machine Vision. *Int. J. of Precis. Eng. and Manuf.-Green Tech.* 9, 661–691 (2022). https://doi.org/10.1007/s40684-021-00343-6.

[4] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural Features for Image Classification," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[5] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, July 2002, doi: 10.1109/TPAMI.2002.1017623.

[6] M. Unser, "Texture classification and segmentation using wavelet frames," in *IEEE Transactions on Image Processing*, vol. 4, no. 11, pp. 1549-1560, Nov. 1995, doi: 10.1109/83.469936.

[7] Brian L. DeCost, Elizabeth A. Holm, A computer vision approach for automated analysis and classification of microstructural image data, Computational Materials Science, Volume 110, 2015, Pages 126-133, ISSN 0927-0256, https://doi.org/10.1016/j.commatsci.2015.08.011.

[8] Maxim Ziatdinov, Ondrej Dyck, Artem Maksov, Xufan Li, Xiahan Sang, Kai Xiao, Raymond R. Unocic, Rama Vasudevan, Stephen Jesse, and Sergei V.

Kalinin, Deep Learning of Atomically Resolved Scanning Transmission Electron Microscopy Images: Chemical Identification and Tracking Local Transformations, *ACS Nano* 2017 *11* (12), 12742-12752. DOI: 10.1021/acsnano.7b07504.

[9] Taekyeong Park, Yongho Son, Sanghyuk Moon, Seungju Han, Je Hyeong Hong, Long-tailed detection and classification of wafer defects from scanning electron microscope images robust to diverse image backgrounds and defect scales, Engineering Applications of Artificial Intelligence, Volume 162, Part A, 2025, 112342, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2025.112342..

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 149, 1597–1607. https://arxiv.org/abs/2002.05709.

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent a new approach to self-supervised learning. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 1786, 21271–21284.

[12] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 15745-15753, doi: 10.1109/CVPR46437.2021.01549.

[13] Ma J, Zhang T, Yang C, Cao Y, Xie L, Tian H, Li X. Review of Wafer Surface Defect Detection Methods. *Electronics*. 2023; 12(8):1787. https://doi.org/10.3390/electronics12081787.

[14] Kim, TY., Park, S., Lim, CK. *et al.* Deep Learning-Based Detection of Defects in Wafer Buffer Zone During Semiconductor Packaging Process. *Multiscale Sci. Eng.* 6, 25–32 (2024). https://doi.org/10.1007/s42493-024-00103-z

[15] Francisco López de la Rosa, José L. Gómez-Sirvent, Rafael Morales, Roberto Sánchez-Reolid, Antonio Fernández-Caballero, Defect detection and classification on semiconductor wafers using two-stage geometric transformation-based data augmentation and SqueezeNet lightweight convolutional neural network, Computers & Industrial Engineering, Volume 183, 2023, 109549, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2023.109549.

[16] Zheng J, Dang J, Zhang T. Wafer Surface Defect Detection Based on Feature Enhancement and Predicted Box Aggregation. *Electronics*. 2023; 12(1):76. https://doi.org/10.3390/electronics12010076.

[17] Zheng J, Zhang T. Wafer Surface Defect Detection Based on Background Subtraction and Faster R-CNN. *Micromachines*. 2023; 14(5):905. https://doi.org/10.3390/mi14050905

[18] Kim, M., Tak, J. and Shin, J. (2024), A Deep Learning Model for Wafer Defect Map Classification: Perspective on Classification Performance and

Computational Volume. Phys. Status Solidi B, 261: 2300113. https://doi.org/10.1002/pssb.202300113.

[19] Mayank Jariya, Parveen Kumar, Rekha Devi, Balwinder Singh, Silicon wafer defect pattern detection using machine learning, Materials Today: Proceedings, 2023, , ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2023.04.233.

[20] Chien J-C, Wu M-T, Lee J-D. Inspection and Classification of Semiconductor Wafer Surface Defects Using CNN Deep Learning Networks. *Applied Sciences*. 2020; 10(15):5340. https://doi.org/10.3390/app10155340.

[21] Lee J, Ju Y, Lim J, Hong S, Baek SW, Lee J. Enhancing Confidence and Interpretability of a CNN-Based Wafer Defect Classification Model Using Temperature Scaling and LIME. Micromachines (Basel). 2025 Sep 17;16(9):1057. doi: 10.3390/mi16091057. PMID: 41011946; PMCID: PMC12472186.

[22] Suhee Yoon, Seokho Kang, Semi-automatic wafer map pattern classification with convolutional neural networks, Computers & Industrial Engineering, Volume 166, 2022, 107977, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2022.107977.

[23] Siyamalan Manivannan, Semi-supervised imbalanced classification of wafer bin map defects using a Dual-Head CNN, Expert Systems with Applications, Volume 238, Part F, 2024, 122301, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.122301.

[24] Dataset available at: https://zenodo.org/records/10715190.