

# Artificial Intelligence for Silicon Wafer Production Monitoring

Luca Frittoli<sup>1</sup>, Nicolò Folloni<sup>1</sup>, Diego Carrera<sup>2</sup>, Beatrice Rossi<sup>2</sup>, Pasqualina Fragneto<sup>2</sup>, Giacomo Boracchi<sup>1</sup>

<sup>1</sup>Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano

<sup>2</sup>System Research and Applications, STMicroelectronics

<sup>1</sup>firstname.lastname@polimi.it, <sup>2</sup>firstname.lastname@st.com

## Abstract

The chips contained in every electronic device are manufactured over circular silicon wafers. The growing demand of semiconductors in nearly all the industrial sectors has made human quality inspection of wafers infeasible. Thus, electronics and semiconductors manufacturers require advanced Artificial Intelligence techniques to automatically monitor their entire production. Here we present the research projects established in a collaboration between Politecnico di Milano and STMicroelectronics, in which we design deep learning models to *i*) recognize and interpret defect patterns in silicon wafers during the manufacturing process, and *ii*) to retrieve similar images from a large and partially annotated database collecting Transmission Electronic Microscopy images of wafer parts. Our CNNs for classifying defect patterns are currently employed in several production sites of STMicroelectronics, and our image retrieval solution is being tested at the STMicroelectronics factory in Agrate Brianza.

## 1 Introduction

The demand for semiconductors has been increasing at an astonishing rate in the latest years due to the growth and technological development of sectors such as automotive and Internet-of-Things. Silicon wafers represent the base upon which every chip is built and require a long and high-tech manufacturing process. Nowadays, the huge production volumes prevent operators from visually controlling 100% of wafers at each production step, thus it is of key importance to use automatic techniques to facilitate quality inspection.

Over the past few years, Politecnico di Milano and STMicroelectronics have established multiple research projects to develop Machine Learning tools for automatically monitoring silicon wafers in different settings and production stages. This partnership has involved STMicroelectronics researchers and production engineers, an Associate Professor and three PhD students (whose scholarships have been sponsored by STMicroelectronics) from Politecnico di Milano. In 2019, one of the PhD students has joined the STMicroelectronics research team, after graduating, while the others are

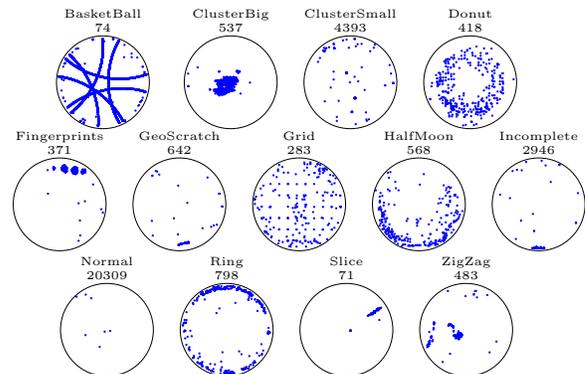


Figure 1: Examples of WDMs from the classes included in the ST dataset, and total number of samples in each class.

currently working towards their PhDs. In four years, these projects involved six MSc students from Politecnico di Milano through curricular internships at STMicroelectronics.

The primary research goal we addressed was the automatic detection of problems and failures in the production process. Current quality inspection machines can in fact exclusively identify localized defects, returning a list of defective locations in each wafer, namely a Wafer Defect Map (WDM). Production issues such as a robot accidentally scratching the wafer surface, can be instead detected by analyzing patterns over the WDM. Not surprisingly, the classification of defect patterns has been widely investigated in the literature [Huang e Pan, 2015] but in rather simplistic settings where *i*) all the classes of defect patterns are assumed to be known and represented in the training set and *ii*) WDMs are transformed into low-resolution images.

The major outcome of the research collaboration is an ad-hoc Convolutional Neural Network (CNN) that both classifies known patterns in WDMs, corresponding to specific production issues, and also detects unknown defect patterns as in *Open-Set Recognition* (OSR) [Geng *et al.*, 2021]. Remarkably, our CNN, which is currently deployed in several STMicroelectronics production sites, can process full resolution WDMs despite their huge size (defect coordinates span a  $20,000 \times 20,000$  grid, corresponding to a precision of  $10\mu m$ ). This research work is presented in two articles [di Bella *et al.*, 2019; Frittoli *et al.*, 2021] and a patent [Moioli *et al.*, 2021].

We also developed a visual explanation tool to support decision making, which is presented in [Morbidelli *et al.*, 2020].

We also designed an image retrieval system based on deep neural networks to support production engineers in the search for similar *Transmission Electronic Microscopy* (TEM) images of wafer parts. TEM images are collected in a large database called IMAGO, and the network returns a set of images from IMAGO that are most similar to a given query image. This automatic retrieval system allows a fast and reliable access to IMAGO, helping engineers to diagnose defects in wafers at every production stage. The main challenge for the training of this system is the lack of annotations for the vast majority of the database, which we address in [Gatta, 2021] by training the network in a hybrid manner, using both siamese and autoencoder loss functions for labeled and unlabeled samples, respectively. The image retrieval system is planned to be deployed at STMicroelectronics, and is currently being tested at the production site in Agrate Brianza.

## 2 Deep Learning for Detecting Process Failures from Wafer Defect Maps

### 2.1 Industrial Scenario

In a wafer manufacturing pipeline there are multiple inspection machines operating at different stages. Inspection machines cannot directly detect failures in the manufacturing process, but only defective locations in each wafer, listed in a Wafer Defect Map (WDM). In normal conditions, WDMs contain a small number of randomly distributed defects. In contrast, production failures result in specific patterns appearing on WDMs. For example, when a robot manipulating wafers accidentally scratches their surface, WDMs might exhibit geometric patterns (see Figure 1). Recognizing these patterns as soon as they are generated allows to promptly identify and solve the issues in the production pipeline, thus preventing them from damaging the entire production.

### 2.2 Wafer Defect Maps Monitoring

A WDM  $w$  is a list of 2D coordinates of defects within a wafer. Since these coordinates belong to a grid defined by the resolution of the inspection machine,  $w$  can be seen as a sparse, binary image  $w \in \{0, 1\}^{K \times K}$ , where each pixel corresponds to an inspected location, which is set to 1 if there are defects and 0 otherwise. Like many industrial problems, there are no public annotated datasets to train a deep learning model. Thus, the first challenge to address was the acquisition and the labeling of a large dataset.

The label of each WDM can either be *Normal* (i.e., no defective patterns) or one of the twelve classes of defective patterns identified by production engineers, illustrated in Figure 1. Failures that had never been observed might result in an *Unknown* pattern over WDMs, and also this case must be handled by the classifier. Therefore, our goal is to train an open-set classifier  $\mathcal{K}$  that associates to each WDM  $w$  either a known class label or the *Unknown* label. A major challenge when handling WDMs is that traditional CNNs cannot be directly applied because images obtained from a full-resolution WDM are huge: in our case, WDMs span a  $20,000 \times 20,000$

grid, resulting in a 3 GB grayscale image. A second challenge to address is class imbalance, because the vast majority of WDMs are *Normal*, while some patterns are very rare and thus under-represented.

Although pattern recognition on WDMs can be performed automatically, operators must decide whether and how to intervene based on classifier predictions. Since deep neural models are not interpretable, we decided to pair the classifier with an explainability tool to support decision-making. In this case, our goal is to provide each prediction with a high-resolution *saliency map* highlighting the regions of the wafer that have mostly influenced the prediction.

### 2.3 Proposed Solutions

**Dataset Acquisition.** Inspection machines produce a WDM for each analyzed wafer. Annotating a large amount of WDMs is a tedious work that has to be carried out by experts, and as such is time-consuming and prone to fatigue-related errors. We therefore designed an intuitive and efficient *Graphic User Interface*, to allow production engineers to quickly annotate 31,893 WDMs, which form the ST dataset.

**Open-Set Recognition on WDMs.** We designed a deep learning pipeline for WDM monitoring, which we address as an Open-Set Recognition problem. We leverage Submanifold Sparse Convolutions [Graham *et al.*, 2018] to efficiently process full-resolution WDMs, and customized data augmentation procedures to overcome class imbalance. Class imbalance is rather typical in industrial monitoring scenarios where the vast majority of products is normal.

In [di Bella *et al.*, 2019] we address WDM classification in a *closed-set* scenario, i.e., assuming that all the possible classes are known and represented in the training set. Traditional CNNs for image classification cannot be directly applied to WDMs because of their huge resolution, so all the previous solutions adopt binning [Huang e Pan, 2015] to produce low-resolution images. However, binning might overlook important information contained in the original WDMs. For this reason, we implement a Submanifold Sparse Convolutional Network (SSCN) [Graham *et al.*, 2018], which can efficiently process sparse images at an arbitrary resolution by taking as input the list of coordinates of the *active sites* of the image (in our case, the defect coordinates). To tackle the severe imbalance [di Bella *et al.*, 2019] we propose class-specific transformations to augment training data. Moreover, to improve classification performance, we apply test-time augmentation using *label-preserving* transformations.

In [Frittoli *et al.*, 2021] we extend our SSCN for WDM classification to address Open-Set Recognition, where we solve the additional task of detecting *Unknown* defect patterns. We follow the simple yet effective approach of applying an outlier detector based on a *Gaussian Mixture Model* (GMM) to the latent representation of the SSCN, namely the output of its penultimate layer. Our intuition is that a classifier maps instances of the same class in the same region of the latent space. Thus, we model the distribution of the latent representations as a GMM. During testing, we classify as *Unknown* those WDMs having a low likelihood with respect to the GMM. Finally, we show that test-time augmentation can be safely adopted also in OSR, since our transformations

True Label \ Predicted Label	BasketBall	ClusterBig	ClusterSmall	Donut	Fingerprints	GeoScratch	Grid	HalfMoon	Incomplete	Normal	Ring	Slice	ZigZag
BasketBall	0.95	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ClusterBig	0.00	0.79	0.04	0.04	0.03	0.00	0.00	0.05	0.01	0.02	0.01	0.01	0.01
ClusterSmall	0.00	0.01	0.74	0.00	0.04	0.02	0.01	0.02	0.10	0.04	0.01	0.00	0.02
Donut	0.00	0.03	0.01	0.89	0.00	0.00	0.00	0.03	0.01	0.00	0.02	0.00	0.00
Fingerprints	0.00	0.02	0.04	0.00	0.86	0.01	0.00	0.02	0.02	0.00	0.00	0.01	0.02
GeoScratch	0.00	0.00	0.02	0.00	0.00	0.87	0.00	0.00	0.01	0.01	0.00	0.01	0.06
Grid	0.00	0.00	0.03	0.00	0.00	0.00	0.93	0.01	0.00	0.03	0.00	0.00	0.01
HalfMoon	0.00	0.03	0.02	0.01	0.03	0.00	0.00	0.79	0.07	0.02	0.03	0.00	0.00
Incomplete	0.00	0.00	0.04	0.00	0.01	0.00	0.00	0.01	0.90	0.02	0.03	0.00	0.00
Normal	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.02	0.93	0.00	0.00	0.00
Ring	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.03	0.10	0.01	0.82	0.00	0.00
Slice	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00
ZigZag	0.00	0.01	0.02	0.00	0.03	0.14	0.00	0.01	0.01	0.00	0.00	0.00	0.77

Figure 2: Confusion matrix obtained by our SSCN through 10-fold cross-validation on the ST dataset.

cannot turn an *Unknown* WDM into an instance of a known class.

**Explainability.** To get an explanation of network’s output, we designed a general framework that leverages data augmentation to compute high-resolution saliency maps, indicating those regions of the image that have mostly influenced a network in providing a specific output. To be meaningful, saliency maps have to be well localized around the object of the queried category (*class-discriminative*) and should capture fine-grained details (*high-resolution*).

A very popular method to explain the output of a CNN is *Grad-CAM* [Selvaraju *et al.*, 2017], which however returns low-resolution saliency maps. In [Morbidei *et al.*, 2020] we present *Augmented Grad-CAM*, which leverages test time augmentation to upsample saliency maps returned by Grad-CAM. To this purpose, we formulate a specific optimization problem inspired by *Multi-Frame Super Resolution (MFSR)*, where we invert an unknown downsampling process by fusing several noisy low-resolution images. We model saliency maps computed from augmented versions of the same input as generated by a process that downsamples and degrades an ideal high-resolution saliency map that we want to recover. The procedure allows us to extract the slightly different information included in each saliency map from augmented images. The whole upsampling procedure can be efficiently performed on the GPU at inference time and our TensorFlow implementation has been publicly released.

## 2.4 Results and Industrial Impact

In our experiments we show that processing full-resolution WDMs by SSCN yields better classification performance than using traditional CNNs trained on low-resolution images [di Bella *et al.*, 2019]. In particular, our SSCN turns out to be more robust to class imbalance than standard CNNs, and our augmentation procedure substantially improves the accuracy of both SSCN and CNNs, especially on under-

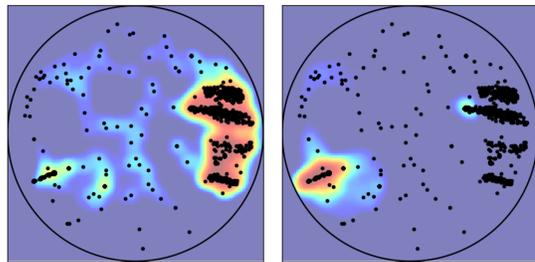


Figure 3: Saliency maps of a WDM referred to two defect patterns: *Fingerprints* (left), and *GeoScratch* (right). The saliency maps highlight the location of the patterns on the WDM.

represented classes, compared to traditional augmentation. Most remarkably, our experiments show that our OSR solution achieves superior unknown detection performance than alternatives from the literature, which we applied on top of the same SSCN for a fair comparison [Frittoli *et al.*, 2021].

Beside two publications in major international scientific venues, our research on WDM monitoring resulted in a US patent [Moioli *et al.*, 2021] and a prestigious corporate award at STMicroelectronics. Thanks to its excellent classification performance, illustrated by the confusion matrix in Figure 2, our solution is currently deployed in several STMicroelectronics production sites all over the world.

We also demonstrate that Augmented Grad-CAM yields high-resolution and class-discriminative saliency maps which are also particularly useful in WDM monitoring. Figure 3 reports a WDM containing two different defect patterns (*GeoScratch* and *Fingerprints*) where the saliency map correctly highlights the regions covering the target pattern.

## 3 Deep Learning for Retrieving Images of Wafer Parts

### 3.1 Industrial Scenario

Inspection machines acquire TEM (Transmission Electron Microscopy) images over specific parts of the analyzed wafer samples. In the Agrate Brianza production site, hundreds of these images are acquired every day and these are collected in a large database called IMAGO. These images represent very different structures corresponding to both fundamental steps in the manufacturing process and recurrent elements in integrated circuits. These images are acquired at very different scales (see Figure 4), colors and resolutions, and only a small fraction of the database has been labeled in the classes corresponding to the depicted structures. Only five structures have been annotated, but many others remain unlabeled. Therefore, an automatic retrieval system, able to efficiently scan the database and recover images belonging to the same class of a query image, is of paramount importance for quality control.

### 3.2 Retrieval of TEM Images

Given a database  $D$  containing  $n$  images, the problem of image retrieval consists in selecting the  $k$  images in  $D$  that are most *similar* to a query  $q \in D$ . In our framework, we consider images to be similar when they belong to the same class

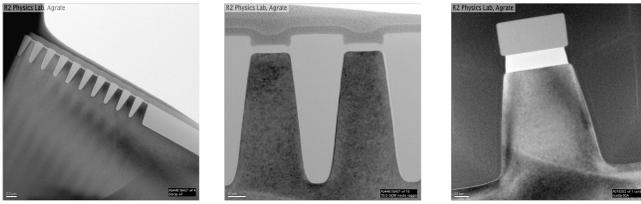


Figure 4: TEM images belonging to the same class of IMAGO, acquired at different scales. As can be seen, images belonging to the same class can be quite different one from the other, so a purely unsupervised approach is not suitable to perform effective retrieval.

in the IMAGO database. State-of-the-art methods to solve the image retrieval problem are mainly based on neural networks such as siamese networks or autoencoders. However, the challenges of the IMAGO database prevent these techniques, which are suited only for supervised or unsupervised problems, to perform an effective retrieval in the IMAGO database.

### 3.3 Proposed Solution

We design a retrieval system that is based on a specifically designed neural network. This network can be interpreted as a function  $f : \mathbb{R}^{N \times M} \rightarrow \mathbb{R}^d$ , which takes as input the image and returns a compact latent representation. We expect the distance between feature vectors of similar images to be small, while dissimilar images are supposed to fall far apart in the latent space. Therefore, first we extract the latent representations of all the images of the IMAGO dataset, then we perform the retrieval by similarity search. The network  $f$  is trained as a siamese network over labeled samples, where we minimize the distance between two similar samples while maximizing the distance between dissimilar ones. In contrast, over unlabeled samples we train  $f$  as an autoencoder, minimizing the reconstruction error of the input image. The training is performed alternating the optimization of the two losses, and this approach allows  $f$  to learn features even from the unknown classes and to prevent overfitting over the small percentage of labeled images available. To further improve the retrieval performance, we implement a query expansion procedure, which consists in averaging the latent representation of the query image with its  $k$  nearest neighbors over training data. The resulting representation is then issued to the network as new query to perform retrieval. This allows queries to return meaningful results primarily characterized by the structure and less biased by individual features of each image.

### 3.4 Results and Industrial Impact

Our experiments were conducted over a subset of the IMAGO database including 34,858 images, 10% of which is labeled. Results in [Gatta, 2021] show that our model outperforms both state-of-the-art solutions based on a single siamese network and a single autoencoder in terms of average precision and mean average precision. Our automatic retrieval system is currently being tested at the Agrate Brianza factory to replace the visual inspection of the TEM images, enabling faster and more effective retrieval.

## 4 Conclusion and Future Work

Artificial Intelligence is becoming increasingly important in semiconductor manufacturing. The fruitful partnership between Politecnico di Milano and STMicroelectronics has produced innovative and effective solutions for monitoring the silicon wafer manufacturing pipeline, resulting in scientific publications, a patent, and a major corporate award. Most remarkably, our WDM classification method is having a real impact, being deployed in several STMicroelectronics factories across the world. Future work will concern the application of our image retrieval system at industrial level and the extension of our explainability tool to our Submanifold Sparse Convolutional Network for WDM classification.

## References

- [di Bella *et al.*, 2019] Roberto di Bella, Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, e Giacomo Boracchi. Wafer defect map classification using sparse convolutional networks. In *ICIAP*, pages 125–136, 2019.
- [Frittoli *et al.*, 2021] Luca Frittoli, Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, e Giacomo Boracchi. Deep open-set recognition for silicon wafer production monitoring. *Pattern Recognition*, 124(2022):108488, 2021.
- [Gatta, 2021] Giuseppe Gianmarco Gatta. Deep learning content-based image retrieval for TEM images, 2021. M.Sc. Thesis, Politecnico di Milano.
- [Geng *et al.*, 2021] Chuanxing Geng, Sheng-Jun Huang, e Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3614–3631, 2021.
- [Graham *et al.*, 2018] Benjamin Graham, Martin Engelcke, e Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018.
- [Huang e Pan, 2015] Szu-Hao Huang e Ying-Cheng Pan. Automated visual inspection in the semiconductor industry: A survey. *Computers in Industry*, 66:1–10, 2015.
- [Moioli *et al.*, 2021] Lidia Moioli, Pasqualina Fragneto, Beatrice Rossi, Diego Carrera, Giacomo Boracchi, Mauro Fumagalli, Elena Tagliabue, Paolo Giugni, e Annalisa Aurigemma. Wafer manufacturing system, device and method, February 16 2021. US Patent 10,922,807.
- [Morbidei *et al.*, 2020] Pietro Morbidelli, Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, e Giacomo Boracchi. Augmented Grad-CAM: heat-maps super resolution through augmentation. In *ICASSP*, pages 4067–4071, 2020.
- [Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, e Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.