Applications of Artificial Intelligence in Silicon / Wafer Manufacturing

UNIVERSITÄT MANNHEIM

Leon Knorr[1]
[1]Mannheim Master of Datascience

Student ID:1902854
Keywords: Manufacturing, AI, Silicon, Waver
22nd May, 2023

1 Introduction

The semiconductor industry plays a pivotal role in driving technological advancements across various sectors, including electronics, communications, and computing. The production of high-quality semiconductor wafers is a critical step, as these wafers serve as the foundation for manufacturing integrated circuits and microchips [1]. As the demand for smaller, faster, and more powerful devices continues to grow, the need for efficient and accurate wafer manufacturing processes becomes increasingly important. In recent years, Artificial Intelligence (AI) has emerged as a transformative technology, offering unprecedented opportunities for enhancing the semiconductor manufacturing workflow [1, 2]. Especially machine learning (ML) algorithms, have the potential to revolutionize semiconductor wafer manufacturing by providing intelligent solutions for quality control and process optimization. For example, traditionally, wafer inspection and defect detection have been labor-intensive and time-consuming tasks performed by human operators, leading to potential errors and inconsistencies. In addition, human operators operate without using every bit of the vast amounts of data available in a wafer manufacturing fab. Each machine of the manufacturing process is equipped with hundreds of sensors, that measure the conditions of the current process [3]. However, the integration of AI into the manufacturing process can significantly improve these aspects, enabling real-time, automated, and highly accurate detection and classification of defects [4]. Aside from defect detection, semiconductor manufacturing poses a lot of potential applications where AI can thrive and enhance the capabilities of a manufacturing plant. This paper aims to explore various applications of AI and machine learning in a semiconductor manufacturing process by explaining them and discussing their impact on the industry.

2 Approach

First, the manufacturing process of semiconductors is defined and the performance indicators for semiconductor manufacturing factories are laid out. Afterwards, a survey of current applications of AI in the manufacturing process is presented. During the survey, each field of application is explained, and different approaches are laid out and discussed.

3 Semiconductor Manufacturing

In this section, an overview of the process of semiconductor manufacturing, as well as important definitions are presented.

3.1 The Manufacturing Process

The manufacturing of semiconductors is a highly delicate and complicated procedure [5]. It consists of hundreds of processes, where the input materials are processed according to the given recipes. These processes can be summarized into four stages:

- Wafer Formation
- Front End Processing
- Testing
- Packaging

Each of these stages is presented in detail in the following section.
It is important to note, that this section only presents the basics of the manufacturing process and that it could've been altered or optimized for different recipes or to improve yields. As a result, some steps mentioned later on during the discussion of AI applications in this industry may not be listed here but are explained as needed.

3.1.1 Wafer Formation

During Wafer Formation, a so-called wafer is created. A wafer is a thin slice, often of a round shape, of extremely pure crystalline material. Usually silicon crystals are used. These wafers serve as the foundational material for microelectronic devices [2]. Each Wafer will hold multiple dies, where a die refers to one specific microchip. After a wafer is formed, it is organized into "Lots". Each lot contains up to 25 wafers, which are processed together, using the same recipe and configuration at each step in the applied recipe [2]. After Wafer Formation, the formed wafers will continue to Front End Processing.

3.1.2 Front End Processing

Front End Processing, encompasses the creation of transistors on a silicon wafer. It is performed in special "clean rooms", where dust, vapor and other pollutants are kept at a minimum. This is achieved through air filtering and strict access policies [2]. Each step in the Front End Process is repeated multiple times, to produce multiple interconnected layers on a wafers surface.
First, a Lot of wafers runs through *Wafer-cleaning*. In this step, each wafer is cleaned, such that is has a surface which is as smooth as possible. This is important, as following processes have strict requirements concerning surface smoothness and particle contamination. If one of these requirements are not fulfilled as good as possible, it can lead to defective dies later on in the process and reduce yields [2].
After Wafer-cleaning, each wafer is run through a process called *deposition*. During deposition, a dielectric and polysilicon film is applied to each individual wafer [2]. This film serves as the foundation for *Photo Lithography*. During Photo Lithography, an ultraviolet radiation is transmitted through a *photomask* to create circuit patterns according to the recipes. The radiation travels through the non-opaque parts of the mask and reacts with the previously applied film, which is coated onto the wafers surface as a result [2]. As this mask essentially burns in the required circuits into the film, the mask has to be aligned as accurately as possible, with only very small tolerances.
After Photo Lithography, the lot of wafers are passed on to *Etching*. During Etching, unused chemicals and layers are removed from a wafer. In the etching process, the circuits patterned onto the wafer using the resistant film, during Photo Lithography, serve as a masking material

which resits the etchant. As a result, the patterned circuits are transferred onto the wafer by removing unnecessary material, layers and chemicals [2].
After Front End Processing is complete, and all needed circuits are transferred onto the actual wafers, each wafer lot proceeds to Testing.

3.1.3 Testing

During Testing, each circuit is tested for functionality and defects. Two different types of tests are carried out:

Parametric Tests Parametric Tests are performed to monitor the efficiency of the process and the quality of the design. It is measured on ad-hoc structured prepared onto the device, and consists of electrical measurements of physical quantities, such as impedance, capacitance and resistance [2].

Electrical Tests Electrical Tests, verify that each die works within specifications, and that its behavior is consistent. If a die performs out of spec, it is marked with a small dot of ink and the passing / non-passing information is stored in a *wafermap* [2].

After testing, the manufactured wafers proceed to *Packaging*.

3.1.4 Packaging

In Packaging, the wafers are sawed into pieces, which only include one die at a time. Then, electrical connections, such as Pins etc., and Integrated Heat Spreaders (IHS) are added to the chip to protect it from mechanical and environment stress, as well providing a proper thermal path for the heat the chip generates under load [2].

3.2 Performance Indicators of Semiconductor Fabs

As with any manufacturing process, semiconductor factories have certain indicators, that represent their efficiency, competitiveness and success. The two main Indicators are the Yield and costs of the manufacturing process [1, 4, 5]. The yield of the process refers to the total amount of chips produced for a given recipe for a given amount of raw material. Yields of a certain ordered chip can vary a lot depending on the state of the equipment, the raw materials, procedures applied etc. Improving Yield, will also decrease costs, as the process gets more efficient and less material has to be used to produce the same number of chips, or if seen the other way around, more chips can be produced for the same amount of material and costs. Improving yield, will also increase the competitiveness of a given factory, as offered pricing can be adjusted accordingly [1]. As yield is influenced by a lot of measurable factors, which in return involves a lot of data, applications around yield optimization are the primary ares of use for AI applications [6]. There are two types of yield:

Line Yield Which refers to the fraction of wafers that reach the final electrical test [5].

Die Yield Die Yield on the other hand is the fraction of dice on yielding wafers, that are not discarded before reaching assembly and final test [5].

Besides Yield, an important quantity is called the Cycle Time. Cycle time refers to the total amount of time a wafer needs to be processed from start to finish. The shorter the cycle time, the better the yields and lower the costs [7].

In reality however, a lot more indicators are present to define the competitiveness of a semiconductor factory, such as the smallest available node size, which influences the power efficiency of a chip and allows the packaging of a lot more transistors onto the same die size.

4 AI Applications

In this section current applications of AI in semiconductor manufacturing are outlined and discussed.
As mentioned earlier, most AI applications that are specific to the domain of semiconductor manufacturing are applied during front end processing. The goal during this stage is to optimize yields by detecting defects as early as possible, minimizing cycle times, optimizing maintenance of the used equipment and predicting the yield of a configuration early on [1, 3, 6–8]. In summary then, all applications influence the yield management strategy of a semiconductor factory in a certain way, in order to make it as efficient as possible.

4.1 Defect Detection

The first application of AI in semiconductor manufacturing is defect detection. As the manufacturing process involves hundreds of sequential and recurring processes, it is essential to detect a fault as early as possible. Even the tiniest scratch or particle contamination of a wafer can lead to lower yields [3]. Over the years, three main categories for defect detection during the manufacturing process have emerged:

- Virtual Metrology
- Machine Learning approaches
- Deep Learning Applications

4.1.1 Virtual Metrology

One way to detect faults is to employ a metrology step after each step in the manufacturing process. During this step, metrology equipment is used to measure physical variables of a given wafer. However, this step is time-consuming and expensive. Thus, it increases Cycle Times and costs [9]. As a result, semiconductor factories use the similarity property of wafers in a lot in order to reduce the amount of time spend on metrology. Instead of measuring each wafer in a lot, a fraction of wafers is sampled from the lot and inspected accordingly [3]. However, non-inspected wafers may still have quality issues, even though the sampled wafer(s) passed the inspection. To overcome this gap, research introduced the concept of Virtual Metrology (VM). VM estimates metrology values for all wafers in a lot, using a predictive regression model [3]. These models are build using previous real metrology data, as well as data from the preceding process, to infer the actual metrology values with reduce costs [3, 9]. Common techniques in Virtual metrology are domain adaptive active learning, Neural Networks. Neural Networks are used to model metrology as a non-linear combination between the tool and logistic data and the target, which is the unlabeled wafer. They have been shown to be powerful approximators for VM but are hard to train because of VMs high dimensionality [2]. In addition, the results often lack interpretability.
In domain adaptive active learning aims at reducing the problem of class imbalance for VM models. As only a fraction of wafers is actually inspected using metrology equipment, the number of accurately labeled wafers is small. The amount of unlabeled wafers on the other hand is huge. In order to reach the desired prediction accuracy, it is necessary to increase the amount of labeled wafers by annotating additional wafers, which again is time-consuming and expensive [9]. During adaptive learning, the VM model iteratively selects wafers to be annotated and uses them to increase the prediction accuracy. However, as manufacturing processes can change on a daily basis, and each change can alter the distribution of process

measurements, a new model has to be built every time new equipment is introduced, or a significant change is applied to the manufacturing process [9]. In this situation no labeled wafers are readily available, and thus, the adaptive learning process can't be started. To overcome this problem, Shim and Kang [9] proposed domain-adaptive active learning. This approaches uses data from existing equipment, as the source domain. Then applied, unsupervised domain adaption to unlabeled wafers as the target and labeled wafers as the source domain to extract domain invariant features. This process allows the construction of a base VM model, which can be used to initialize active learning [9]. Active learning then proceeds to iteratively update the model and tune it for the current domain (new process and new equipment). From a business perspective VM models have to have two properties. They have to have low compute times, as they always need to be updated and their primary task is to reduce cycle times. And, the results have to be interpretable. This is important, as it allows the identification of the most important variables, which can be used to enhance the manufacturing process and reduce defects. According to Susto et al. especially the latter is challenging for current research [2].

4.1.2 Machine Learning

Besides Virtual Metrology, another common approach to detect defects is the use of machine learning. These methods, try to predict if a wafer has a defect based on data from the hundreds of sensors of a process step. This includes physical variables such as the air temperature or the exposure time. Such data is also referred to as *Fault Detection and Classification Data* (FDC). Key to understand the limitations of such a machine learning approach is that FDC data does not measure wafer quality directly [3].
The task is often modelled as a binary classification task, as a wafer is either defective or not. According to a study by Susto et al. [2], the most common models to be used are K-Nearest Neighbor, Support Vector Machines, Principal Component based K-Nearest Neighbor and Decision Trees. An example for a K-Nearest Neighbor classifier for Defect detection based on FDC data is He and Wangs *FD-KNN* [10]. FD-KNN uses FDC data to make decisions based on small local neighborhoods of similar batches. FD-KNN calculates the sum of the squared distance between the k the nearest data points. If this sum exceeds a set threshold, the Wafer is marked as defective [10]. This approach poses several advantages:

1. It is able to capture non-linear patterns in the process data
2. It can work on process data directly without the need for special data preprocessing
3. It copes well with class imbalance
4. It is simple, time and compute resource efficient.

Despite these advantages, Susto et al. also lists the same challenges for machine learning approaches as for VM, such as high dimensionality, the lack of interpretability etc. The exception is that machine learning approaches additionally suffer from missing unstructured data, which is often recorded manually by engineers during maintenance [2].
In order to combine the best of FDC based defect detection approaches and VM, Kim et al. [3] proposed a machine-learning based novelty detection method for faulty wafer detection, which uses both FDC data and the results of VM. Including VM data, is especially useful as faulty wafers are defined to have similar Metrology data. In a first step, the input data was cleaned, by removing meta-information about the equipment, categorical data that only has one value and variable which have zero standard deviation. The target variable of the system were derived from real world metrology data, which exposed four different target variables. Thus, the authors

build four different models, one for each target variable [3]. However, even after data cleaning, the data is still of high dimensionality, leading to the degradation of model performance. Thus, Kim et al. applied two dimensionality reduction methods, variable selection, which chooses input variables that contribute most to model performance, and Principal Component Analysis. In order to cope with class imbalance, a novelty detection approach has been applied, which allows the model to be trained only with data of non-defective wafers, with the goal to detect wafers which have different characteristics than the ones in the training data. This approach goes inline with the definition of defective wafers, as defective wafers do not have any special characteristics. Rather they are defined to be different from the majority of non-defective wafers by not being in spec [3]. In terms of novelty detection methods, they applied three kinds of methods:

- Density estimation methods, such as Gaussian Density Estimation, which computes the distance between a new data point and the center of the gaussian distribution. If it exceeds a set threshold, the wafer is classified as defective.
- Non-probalistic methods, such as k-means clustering, which works in a similar fashion as He and Wangs approach, or 1-Support Vector Machine.
- Reconstruction methods, such as PCA, where a data point can be reconstructed using the set of eigenvectors that defines the linear subspace the variables have been mapped to during PCA. For the reconstruction process a reconstruction error can be calculated. If it exceeds a set threshold for new test data, the wafer is classified as defective.

During Kim et al. performance comparison 1-SVM performed the best in a cross validation setting [3].

4.1.3 Deep Learning

In recent years, the equipment for defect detection has advanced to include automated visual inspections (AVI) [1]. This type of equipment, such as Scanning Electron Microscopes or X-Ray Analyzers yield images of the wafers, which expose their characteristics. Deep Learning methods set out to use the gathered image data of this equipment, as well as from wafer maps constructed during the manufacturing process to not only classify wafer defects, but to also try and determine the cause by assigning the wafer a defect type [1, 4]. This important, as most defects are results of processing errors or contamination during front-end processing. Thus, if the root cause of a defective can be identified correctly, it can be fixed improving the yield and efficiency of the factory, while also reducing costs. There are two primary types of defects:

1. **Random Defects**. Random defects are randomly scattered across the whole Wafer. They are often attributed to the manufacturing environment, hinting at possible particle contamination or equipment faults. Their elimination requires equipment replacement and intense maintenance of the manufacturing environment, both of which are extremely expensive [1].
2. **Systematic Defects**. Systematic Defects, are defects that occur in clusters or a specific pattern on a wafer. Their patterns are often correlated to their root cause. Thus, their correct identification and analysis is vital to improve product quality and die yield [1].

Most defect detection systems only targets systematic defects, as they encode useful information. Random defects are often just perceived as image noise and ignored. However, random and systematic defects often occur together on the same wafer, which makes correctly identifying

systematic defects more difficult [1]. Modern Deep Learning techniques for image processing, such as Convolutional Neural Networks have been proven to be highly effective in order to identify these defect patterns, all without the need for manual feature engineering and special data pre-processing. In addition, they cope well with low-resolution and noisy data [1]. The most used network architecture in this field are CNNs because of their simplicity. They are easy to train and yield good results, also while being highly adaptable. However, obtaining sufficient labeled data for CNN training is difficult, thus unsupervised approaches such as Generative Adversarial Networks (GANs) and Auto Encoders (AE), also received a lot of attention [1]. In terms of interpretability, reliability and quality, Deep Learning approaches have been proven to be superior to earlier machine learning approaches [1].

4.2 Yield Prediction

In addition to detecting defects, which marks the biggest field of AI applications in semiconductor manufacturing. Being able to reliably predict the different yields of a product is vital to a factories planning and strategic departments. In addition, it reduces the strain on supply chains by setting accurate expectations up front. In addition, yield prediction can also help to monitor the manufacturing process and determine points for Improvements in the process, by monitoring the predicted yield over the whole process [5, 6]. Therefore, yield can be predicted at various steps in the manufacturing process, such as after every process step and after every process stage. Systems can also predict different types of yields, such as line yield, die yield and final test yield [5, 6]. Yield prediction can also be used to classify wafers into different yield subpopulations, which assigns each wafer to a bin, depending on where the yield prediction system predicted the highest yield. Each bin has different performance characteristics [6]. For example, bin A might include fully functional chips, while bin C contains die which are only working up to 75%. If this is put into the domain of CPUs, for the first bin all cores work but for bin C only 3/4 of the cores are working. This allows to increase the overall yield of the factory by reusing defective dies to produce lower-spec products. Jiang et al. proposed a yield prediction framework which achieves this by using an ensemble classifier consisting of an SVM, K-NN, Gaussian Process, Logistic Regression, Tree Classifier and Gradient Boosting [6], where the top three performing models are used in a soft voting setting to form the final model. The final models input consists of data from the production stage where the model should be deployed in the manufacturing process. Originally, it has been deployed at the Wafer Acceptance Test. As Jiang et al. work is supposed to be used as a framework to build your own prediction model for a specific product, all applicable models have to be tested and can also be extended to feature other approaches if needed [6].

4.3 Cycle Time Optimization

Cycle Time Optimization is the industry specific adoption of production scheduling. This process allocates a limited number of machines to different processing wafer lots to optimize the on-time delivery rate and inventory holding costs [7]. In the domain of semiconductor manufacturing, these two objectives can be summarized by using only one primary objective, which is to reduce the Cycle-Times of a wafer lot. Traditional approaches to Cycle time optimization / wafer lot scheduling include the use of mathematical programming, heuristic rules and intelligent algorithms. All of them work well in theory, but in practice, they are difficult to implement because of frequent dynamic disturbance, changes in job processing, poor job quality or changes in the process technology [7]. Thanks to emerging technologies like the Internet of things (IoT), every piece of equipment and part of the process provides lots of

real time data, which can be collected and analyzed. Using this data, allows scheduling systems to adhere to dynamic and ever-changing environments as in semiconductor manufacturing [7]. Because many interactions with the environment are needed to properly estimate the effects of a certain action, Reinforcement learning poses a natural fit for modelling such an optimization problem. However, the characteristics of semiconductor manufacturing, such as the re-entrant process, where wafer lots are processed multiple times as they are build layer by layer and product batching, forms significant challenges to build accurate and good performing models [7].

4.4 Predictive Maintenance

Knowing at which point in time a piece of equipment will fail or lead to substantial defects during manufacturing is essential to keep up high yields, high quality, reduce costs and reduce downtimes. Instead of "running to failures" or performing recurring maintenance on a fixed schedule, predictive maintenance systems, allow factories to schedule maintenance actions only when necessary, while ensuring maximum equipment efficiency [2]. To assess, whether maintenance actions should be taken or not, predictive models are deployed, which monitor real-time data from the manufacturing equipment, and predict when maintenance actions are likely to be required. This tools can also provide data scheduling methods, which aim at decreasing cycle times, to account for maintenance operations in their optimizations [2]. To enable this sort of behavior, predictive maintenance systems usually employ a Health Factor. A Health Factor is a quantitative measure of the status of the equipment, which is calculated from observable parameters from the processing equipment. These parameters are often available as historical time series and sensor data. Because of the complexity of this task, no real state-of-the-art model has yet been found, that provides sufficient accuracy [2]. However, different common approaches such as regression tasks and SVM have been applied to predictive maintenance. Finding a common denominator between different environments and tasks for predictive maintenance is really hard. Because of that, currently every problem in that field is studied separately to achieve some-what presentable results [2]. Susto et al. were able to identify three main challenges in predictive maintenance [2]:

1. **Lack of data**: Predictive maintenance problems suffer a from the lack of sufficient amounts of observations to build a reliable statistical model. This is because the amount of observed maintenance interventions are very small, compared to the total number of processed wafers and current maintenance protocols are actively disturbing the measurement process, as the equipment might be maintained even if no indication for a failure is present.

2. **Non-trivial evaluation of the impact on the manufacturing process**: Measuring how the implementation of a predictive maintenance system affects the industrial environment is challenging. Thus, it is hard to compare different approaches of predictive maintenance against each other.

3. **How reactive should the model be?**: To determine an optimal balance between being reactive to a maintenance indicator and when to ignore is also challenging. Reacting too much, increases costs and down-times of the manufacturing equipment, while reacting less can reduce yields because of too many defective wafers.

Because of these challenges, predictive maintenance remains to be an active research field. It has a lot of potential to revolutionize the industry, but is hard to get right. It also needs a lot of work in establishing a common benchmark to be able to compare different approaches.

5 Conclusion

In this work, different applications of AI in semiconductor manufacturing have been laid out. First, the manufacturing process, as well as common and important performance indicators has been explained. Then AI applications in defect detection, yield prediction, cycle time optimization and predictive maintenance have been discussed. Defect Detection turned out to be the biggest and most impactful research field for AI applications in semiconductor manufacturing. A lot of different methods, ranging from the prediction of results from physical equipment in virtual metrology, to the application of image based deep learning methods have been applied to detect wafer defects as early as possible. Detecting them early is vital to increase yields and reduce costs. One important factor in the utilization of deep learning methods, is identification of the defect type, allowing engineers to draw conclusions about the root cause of the defect and optimizing the manufacturing process.
In Yield prediction, models aim at predicting a factories yield based on real-time data from the current step in the manufacturing process. Aiming at supporting the planning process of a factory or to increase the turn-over rate of raw materials, by classifying them into yield subpopulations.
Another application of AI, is the optimization of cycle times through Reinforcement Learning. Which make use of the connectivity of the manufacturing equipment by monitoring real-time data to adhere to the changing environments in semiconductor manufacturing.
Lat but not least, the field of predictive maintenance has been covered. In this field, the development of applicable methods turned out to be still in the early stages. Current research has to cope with significant challenges because of the characteristics of the environment. This includes the lack of observation data, non-trivial evaluation of developed systems and finding an optimum as to how reactive the system should be.
Even though AI applications still face a lot of challenges in every field of application, they are set to revolutionize the industry of semiconductor manufacturing. They have the potential to substantially reduce costs, increase yields and help factories to keep up with the ever-growing demand for semiconductors and their technological advancements.

6 References

1. Batool, U., Shapiai, M. I., Tahir, M., Ismail, Z. H., Zakaria, N. J. & Elfakharany, A. A Systematic Review of Deep Learning for Silicon Wafer Defect Recognition. *IEEE Access* **9.** Conference Name: IEEE Access, 116572–116593 (2021).
2. Susto, G. A., Pampuri, S., Schirru, A., De Nicolao, G., McLoone, S. F. & Beghi, A. *Automatic Control and Machine Learning for Semiconductor Manufacturing: Review and Challenges* en. in *Proceedings of the 10th European Workshop on Advanced Control and Diagnosis (ACD 2012)* (Technical University of Denmark, Kgs. Lyngby, Denmark, 2012).
3. Kim, D., Kang, P., Cho, S., Lee, H.-j. & Doh, S. Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. en. *Expert Systems with Applications* **39,** 4075–4083 (Mar. 2012).
4. Yuan-Fu, Y. *A Deep Learning Model for Identification of Defect Patterns in Semiconductor Wafer Map* in *2019 30th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)* ISSN: 2376-6697 (May 2019), 1–6.
5. *A machine learning approach to yield management in semiconductor manufacturing* en.

6. Jiang, D., Lin, W. & Raghavan, N. A Novel Framework for Semiconductor Manufacturing Final Test Yield Classification Using Machine Learning Techniques. *IEEE Access* **8.** Conference Name: IEEE Access, 197885–197895 (2020).
7. Wang, J., Gao, P., Zheng, P., Zhang, J. & Ip, W. H. A fuzzy hierarchical reinforcement learning based scheduling method for semiconductor wafer manufacturing systems. en. *Journal of Manufacturing Systems* **61,** 239–248 (Oct. 2021).
8. Irani, K., Cheng, J., Fayyad, U. & Qian, Z. Applying machine learning to semiconductor manufacturing. *IEEE Expert* **8.** Conference Name: IEEE Expert, 41–47 (Feb. 1993).
9. Shim, J. & Kang, S. Domain-adaptive active learning for cost-effective virtual metrology modeling. en. *Computers in Industry* **135,** 103572 (Feb. 2022).
10. He, Q. P. & Wang, J. Fault Detection Using the k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes. *IEEE Transactions on Semiconductor Manufacturing* **20.** Conference Name: IEEE Transactions on Semiconductor Manufacturing, 345–354 (Nov. 2007).