Combining Machine Learning with Advanced Outlier Detection to Improve Quality and Lower Cost

Advanced Processor Control Smart Manufacturing Conference
October 2020

Christian Sendner
christian.sendner@pdf.com
PDF Solutions

Jeffrey D. David
jeff.david@pdf.com
PDF Solutions

Tomonori Honda
tomo.honda@pdf.com
PDF Solutions

Keith Arnold
keith.arnold@pdf.com
PDF Solutions

Vaishnavi Reddipalli
vaishnavi.reddipalli@pdf.com
PDF Solutions

Greg Prewitt
greg.prewitt@pdf.com
PDF Solutions

Richard Burch
richard.burch@pdf.com
PDF Solutions

Pragya Moharatha
pragya.moharatha@pdf.com
PDF Solutions

Abstract: In semiconductor manufacturing, a low defect rate of manufactured integrated circuits is crucial. To minimize outgoing device defectivity, thousands of electrical tests are run, measuring tens of thousands of parameters, with die that are outside of specified parameters considered as fails. However, conventional test techniques often fall short of guaranteeing acceptable quality levels. Given the large number of electrical tests, it can be difficult to determine which electrical test to rely upon for die quality screening. To address these issues, semiconductor companies have recently begun leveraging artificial intelligence and machine learning to better identify defective devices while minimizing the fallout of good die from electrical tests. To implement these advanced machine learning applications, a novel remote inference capability is also proposed. By placing an inference engine and corresponding machine learning models at the assembly and test house, inferences can be made without any sensitive data leaving the assembly and test house. The result is faster turnaround times on inferences, reduced data loss, increased security, and the enablement of advanced machine learning capabilities for real-time solutions such as adaptive testing.

Keywords: data mining, artificial intelligence

Motivation

In semiconductor manufacturing, a low defect rate of manufactured integrated circuits is crucial. To minimize outgoing device defectivity, electrical tests typically measure thousands, or even tens of thousands of parameters, e.g. voltage, current, or time delay. Univariate statistical techniques are leveraged to identify statistical outliers in various ways among these electrical tests performed, i.e. each test parameter is considered individually. One such commonly applied univariate techniques is Part Averaging Testing (PAT) [1]. For each test parameter, an upper and lower limit is chosen. Die that are outside of these limits are considered as fails. These limits can either be fixed statically for all wafers (SPAT) or dynamically for each wafer based on the measurement values' mean and standard deviation (DPAT) [2]. PAT is best applied if the measurements follow a Gaussian, or normal, distribution. Other techniques commonly used today contemplate how these test measurements are distributed across the wafer, and the failure rate of die within a given die's vicinity on the wafer.

Recently, however, the above-mentioned conventional test techniques often fall short of guaranteeing acceptable quality levels. In today's advanced semiconductor environment, thousands of electrical tests are performed to determine the quality of the die. Given this large number of electrical tests, it can be difficult to determine which electrical test to rely upon for die quality screening. Furthermore, many of the above-mentioned conventional screening techniques assume a normal distribution of the electrical test measurement results across a die population. Finally, device quality may not be a function of any single electrical test result, but could be multi-variate in nature.

To address the issues of a massively-dimensional test measurement space, the possibly non-normal distribution of electrical tests, and the multi-order interactions of individual electrical test measurements, fabless semiconductor manufacturers and integrated device manufacturers (IDM's) have recently begun leveraging artificial intelligence and machine learning to better identify defective devices while minimizing the fallout of good die from electrical tests. The end result is a reduction in testing costs and improved product quality. In this publication, we will explore the application of modern machine learning techniques to predict devices at risk while concurrently expediting and/or reducing tests for die with little risk of defectivity. By employing more advanced, multivariate outlier screening techniques powered by machine learning, defective chips can be identified more efficiently with less fallout. Additionally, the residual cost of test can be invested to more thoroughly screen devices exhibiting marginal quality to increase overall outgoing quality.

To implement these advanced machine learning applications, a novel remote inference capability is also proposed. By placing an inference engine and corresponding machine learning models at the assembly and test house, inferences can be made without any data leaving the assembly and test house, where much of the data used to make inferences already resides. Furthermore, machine learning models that contain sensitive intellectual property remain secure within the assembly and test house. The result is faster turnaround times on inferences, reduced data loss, increased security, and the enablement of advanced machine learning capabilities for real-time solutions such as adaptive testing.

Approach

Multivariate Anomaly Detection

In order to implement the pass/fail classification of chips correctly, it is ideal to generate a multiclass classifier for each of the failure types. However, there usually isn't enough training samples for each type of failure, as is in the case of failed field returns, or RMA's. These types of situations are well-suited for a branch of machine learning called Anomaly Detection. Anomaly Detection defines a boundary of what is normal and treats anything outside of this boundary as abnormal.

Many univariate outlier screening techniques, such as PAT, are used today in the semiconductor industry for the purpose of outlier screening and can be considered within the field of Anomaly Detection. Some multivariate Anomaly Detection techniques already exist in the semiconductor industry to find outliers in wafer sort data [3]. However, these techniques typically use a principal component analysis (PCA) to transform the measurement parameters into a reduced set of new parameters with removed correlations. The same univariate method is then used to find outliers. A limitation of PCA is that it can only remove the linear dependence between parameters.

Multivariate Anomaly Detection defines normal ranges, while allowing for correlated multimodal distributions for normal chips. For Multivariate Anomaly Detection to work well, it is optimal to rely on selected features that have the most predictive power. There are a number of ways to accomplish this input parameter selection. One technique is to employ univariate feature selection using a failure label (e.g. a field return of burn-in failure). By doing so, we isolate the measurements that are more critical to predicting a failure.

It can also be important to choose methods that contemplate non-Gaussian distributions of the parameter population. These methods can find outliers that are not seen in a univariate analysis, e.g. for non-linearly dependent features.

It should be noted that output from univariate Anomaly Detection methods can be used as input to multivariate approaches (rather than just the raw test parameter values). Table 1 describes the Multivariate Anomaly Detection techniques proposed, describing the pros, cons, and benefits.

Table 1: Pros and Cons of Multivariate Anomaly Detection Algorithms

Algorithms	***Pros***	***Cons***	***Additional Benefits***
MV-1	Able to identify outliers in a dataset that would not be outliers in another area of the data set.	Sensitive to distance definition, especially for higher dimensional space.	Could output nearest neighbors to particular test sample.
MV-2	Population does not need to follow Gaussian Distribution, just needs to be part of bigger set of clusters.	Very sensitive to scaling of input, especially for higher-dimensional space.	Can assign cluster id to normal cluster as well.
MV-3	Not sensitive to scaling of input. Can handle mixed continuous and discrete inputs.	May be harder to explain the result.	Determine overall Variable Importance easily.
MV-4	Less sensitive to correlation and coupling of input variables.	Easy to overfit. Difficult to understand why particular sample is considered outlier.	Hidden layer could be used as nonlinear dimensional reduction.
MV-5	Designed to capture nonlinearity.	Anomaly sample needs to be small.	Can obtain decision hyperplane (boundaries). Could be extended to classification coherently.

Modeled Yield Outlier Screening

A technique is proposed which identifies chips that are considered higher risk for failure by the machine learning model compared to typical die on the same spatial locations. The basic algorithm approach, herein referred to as Modeled Yield, is to develop two sets of models:

a) A model that considers only spatial information.
b) A model that utilizes test parameter values and features generated from these parametric values.

An ensemble of these two models identifies which die are likely defective, or low yielding. Yielding die with low predicted yield are identified as likely candidates for early lifetime failure. If the predicted yield considering only spatial information is high while the predicted yield including parametric values is low, the die are extremely likely to become failures. Experience has shown that die with low predicted yield can be an order of magnitude more likely to become field returns. If the spatial only model predicts that the die should be high yielding, the increased likelihood of a field return increases to nearly two orders of magnitude.

Results

The above-mentioned multivariate anomaly detection and modeled yield techniques were evaluated using an actual production dataset which contained roughly 20,000 total chips of which roughly 50 chips were field returns (not simulated). To predict field returns, input parameters were obtained from a wafer sort test insertion. There were roughly 10,000 raw input parameters. As a baseline to compare against, the industry standard DPAT outlier screening technique was used.

The comparison metrics are False Positive Rate (FPR) and True Positive Rate (TPR). FPR and TPR are common machine learning metrics and are calculated as follows:

$$TPR = True\ Positives\ /\ (True\ Positives\ +\ False\ Negatives) \quad (1)$$
$$FPR = False\ Positives\ /\ (False\ Positives\ +\ True\ Negatives) \quad (2)$$

Where Positive = field returned chip, and Negative = good die. Thus TPR is the percent of all actual field returns that are correctly predicted as defective, and FPR is the percent of die predicted as defective but are actually good die. FPR can be interpreted as the amount of “overkill” population sacrificed to screen defective die for a given defect capture rate.

To demonstrate the true production decision making process, the validation methodology shown in Fig. 1 was employed to train and test each of the multivariate anomaly detection techniques.

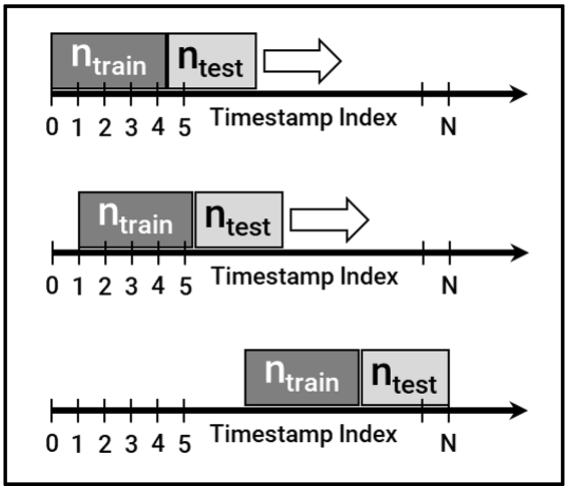

Figure 1 Validation approach to evaluate the multivariate anomaly detection techniques. A sliding window of training and testing data, partitioned by time, is applied to the entire dataset to simulate a real production scenario. nTrain = 20 timestamps, and nTest = 5 timestamps, where each timestamp contained virtually equivalent die count.

For modeled yield, there are no training/testing partitions as this approach does not rely on a rank ordering or selection of input parameters as determined by a correlation to failures.

For the baseline DPAT comparison, there were 38 total significant parameters identified by multivariate approach in each sliding window over the duration of the dataset. The reoccurrence of each significant parameter across the duration of the dataset was calculated, and those parameters that appeared at least

25% of the time were selected. 25% was chosen so that there would be a substantial amount of parameters to run DPAT.

Results for all the testing windows are shown in Table 2 for FPR = 1%, 2%, 4%, and 6%.

Table 2: Algorithm Comparison

	False Positive Rate (FPR)			
	1%	**2%**	**4%**	**6%**
MV-1	14.0%	21.7%	52.2%	76.5%
MV-2	NA	25.4%	52.7%	69.3%
MV-3	-16.9%	44.6%	70.0%	91.7%
MV-4	36.8%	36.9%	57.2%	78.0%
MV-5	NA	NA	31.9%	94.1%
Modeled Yield	46.9%	83.6%	137.4%	154.9%

Results in Table 2 show MV-1, MV-4, and Modeled Yield techniques are all able to capture more field returns in respective portions of the total chip population identified as at-risk (“overkill”). Overall, the Modeled Yield technique performed the best for this dataset. This would imply that for this dataset, yield at wafer sort is a good predictor for field returns. It is important to note that this is not always the case, as it has been observed that for some datasets Modeled Yield does not outperform the other multivariate techniques. Additionally, we have observed that different Multivariate Anomaly Detection techniques outperform others for different datasets and/or chip product lines.

Deployment

Once perfected and trained, machine learning models must be deployed to and integrated with the overall product manufacturing flow. This likely requires deployment of prediction models to multiple remote facilities in today’s distributed manufacturing ecosystem.

The term “Edge Prediction” as used in this paper refers to deployment of machine learning to facilities where production test and assembly operations are performed. Distributed machine learning requires reliable mechanisms to transport and update prediction models, compute infrastructure at remote facilities, timely access to test data, and potentially, integration with factory process automation and control systems.

Use Cases

Compelling reasons to employ edge prediction include:

1) Die Grading and Exclusion – where traditional statistical outlier detection, or better yet, more advanced machine learning models are used to grade individual die based on their performance in contrast to their test population or historical data.
2) Computation of Die Quality Metrics – based on available lot, wafer and/or die specific data such as lot equipment history, parametric test, visual inspection and electrical test operations to infer the likely quality of subject die and to prescribe either the avoidance of or requirement for further testing.

Deployment Challenges

There are numerous considerations and related challenges when contemplating the effective deployment of machine learning.

- What are the timing constraints for prediction?
- How will the necessary input data be collected, merged and sourced to the prediction model?
- What are the confidentiality and security requirements for the prediction model?
- When can and should predictions be run?

1) Prediction Time Domains

Different machine learning applications correspond with different turnaround time and input data requirements. The most stringent of turnaround times are those executed during device test in real-time. Inline predictions are based on the current data stream and any precomputed feed forward data such that the turnaround time represents a fraction of the overall device test time. These real-time predictions must be tightly integrated with either or both test program and test platform. Alternatively, other predictions are better suited to be computed by a post-process server-side implementation where the model can consume data from multiple wafers, lots, dates and test operations. While these server-side predictions typically have a somewhat relaxed requirement for turnaround time, they frequently consist of more computationally exhaustive models, require larger data sets (e.g. lot, lots or date range) and may need to communicate the prediction results to other systems (e.g. update a wafer assembly map). Real-time computations most commonly execute directly on test equipment control computers, but care must be taken to not slow the test operation. Post-process computations should clearly be offloaded to local servers which can provide significantly more memory and processing power without impacting test operations. This suggests that an effective machine learning deployment requires compute servers at each facility.

2) Prediction Input Data

Most predictions require data sets integrated across multiple die, wafers, lots, or dates. These data sets must be collected, merged and conformed before a useful prediction can be made. For instance, consider the example of a split-lot tested on more than a single tester. To satisfy the prediction model data requirements, the data must be collected from more than one test system. Further, prediction models are normally only applicable to data points from the last die touchdown or device insertion when a device has been retested. This requires a system to collect and merge data at the die level. Lastly, for prediction models that consider data across multiple test steps, data may need to be stored for weeks to be available when the subsequent test or assembly operation is performed. Taken together, these data requirements demand that a data store with intelligence to merge data is required at each facility where machine learning will be deployed.

3) Model Confidentiality and Security

Implementation of industry standard outlier detection algorithms may not raise much concern for confidentiality or security, but more advanced machine learning models and their key parameters may well be considered highly sensitive by the device manufacturer. There are several potential platforms upon which machine learning computations can be based. Some of these are inherently more opaque than others. For instance, a model compiled from C-language will be binary in nature and would require significant effort to reverse engineer. Other machine learning platforms are, by default, much less opaque. For example, the greater Python data science ecosystem is perhaps the most popular machine learning platform, and due to its Python heritage of open source, is difficult to produce an intractably opaque executable model. For device manufacturers concerned with the confidentiality of their prediction models, deployment must also include forms of strong code obfuscation, encryption and/or server security.

4) Prediction Timing

Timing of real-time predictions is rather obvious; compute as the data is produced by the test system. Timing the execution of post-process predictions is not as straight forward. Post-process computations are usually needed at manufacturing operation boundaries. Examples include, when a wafer test is complete, after partial wafer tests have been merged, when a wafer lot has completed the wafer sort test operation or visual inspection operations, or before the start of the assembly operation. Simple observation of incoming data does not provide a reliable trigger for when prediction models should be executed. Machine learning can be thought of as a virtual test operation with associated yield loss, and therefore, requires integration with the overall manufacturing flow much like physical test operations. This implies that a successful edge prediction deployment must also have integration points with the test and assembly facility manufacturing execution system (MES).

Conclusion

In this publication, we have demonstrated the possible benefits of employing more advanced, multivariate outlier screening techniques powered by machine learning. In order to make multivariate screening work well, prescreening of test measurements is required to reduce the noise. We have demonstrated a second approach based on creating a proxy to the target variable when a target variable is difficult to obtain. By using these approaches, the residual cost of test can be re-invested to more thoroughly screen devices exhibiting marginal quality to increase overall outgoing quality.

Additionally, this paper discussed an approach for implementing a novel remote inference engine. Inferences can be made without any data leaving the assembly and test house by placing an inference engine and corresponding machine learning models at the OSAT. This approach insures that prediction will be faster by reducing unnecessary data transfer over internet. Furthermore, sensitive intellectual property including test data and machine learning models remain secure within the assembly and test house.

Acknowledgment

The authors of this paper would like to thank Kenneth Harris of PDF Solutions, Dennis Ciplickas of PDF Solutions, Hanna Bullata of Progineer, and Rauf Afifi of Progineer for their contributions and support in making the findings in this paper possible.

References

[1] Manuel J. Moreno-Lizaranzu and Federico Cuesta, Sensors 2013, 13, 13521-13542; doi:10.3390/s131013521

[2] Automotive Engineering Council AEC-Q001

[3] Jeff Tikkanen, Nik Sumikawa, Li-C. Wang, Magdy S. Abadir (2014). Multivariate Outlier Modeling for Capturing Customer Returns – How Simple It Can Be. http://mtv.ece.ucsb.edu/licwang/PDF/2014-IOLTS.pdf