Machine Learning on Multiplexed Optical Metrology Pattern Shift Response Targets to Predict Electrical Properties

Thomas J. Ashby, Vincent Truffert, Dorin Cerbu, Kit Ausschnitt, Anne-Laure Charley, Wilfried Verachtert, and Roel Wuyts

***Abstract*—Doing high throughput high accuracy metrology in small geometries is challenging. One approach is to build easily measurable proxy targets onto dies and make a predictive model based on those signals. We use optical Pattern Shift Response (PSR) proxy targets to build predictive models of the electrical characteristics of devices in the Back End Of Line (BEOL). Given the wide choice of PSR targets, we explore how to select combinations of them to maximise the utility of the features for building an accurate Machine Learning (ML) model; we call this approach Multiplexed Optical Metrology. We also explore the trade-off between chip area dedicated to targets and achievable accuracy. We run ML experiments using different selections of targets measured at different stages of BEOL processing: post-lithography and post-Chemical-Mechanical-Planarisation (CMP). Our results show that a) reasonable predictive performance can be achieved for a reasonable area budget; b) ML model performance across target families varies significantly, thus justifying the need for careful selection of targets; c) longitudinal measurements of targets increases accuracy for no extra area penalty; d) increasing the number of targets gives some improvement in accuracy for a dataset of this size, but relatively small compared to the increase in area budget needed. Ultimately we aim to do die-level yield prediction using these techniques. We discuss how collecting a larger dataset with appropriate yield information is the logical next step to achieving this.**

***Index Terms*—Metrology, high throughput, machine learning, optical target, semiconductors, yield prediction, BEOL, XGBoost.**

Manuscript received 15 September 2023; accepted 28 November 2023. Date of publication 5 December 2023; date of current version 5 February 2024. This work was supported in part by the MADEin4 Project and in part by the FlandersAI Project. MADEin4 has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under Grant 826589. This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation programme and France, Germany, Austria, Italy, Sweden, The Netherlands, Belgium, Hungary, Romania, and Israel. FlandersAI is the Flemish regional government's multi-year Artificial Intelligence Research Programme. *(Corresponding author: Thomas J. Ashby.)*

Thomas J. Ashby and Wilfried Verachtert are with the Exascience Lab, Imec, 3001 Leuven, Belgium (e-mail: ashby@imec.be).

Vincent Truffert, Dorin Cerbu, Kit Ausschnitt, and Anne-Laure Charley are with STS/Advanced Patterning, Imec, 3001 Leuven, Belgium.

Roel Wuyts is with the Exascience Lab, Imec, 3001 Leuven, Belgium, and also with Imec-Distrinet Computer Science, KU Leuven, 3000 Leuven, Belgium.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSM.2023.3339330.

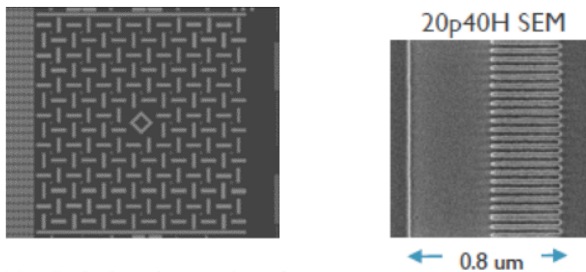
Digital Object Identifier 10.1109/TSM.2023.3339330

I. Introduction

As THE size of semiconductor process technology nodes shrinks, implementing metrology that is capable of predicting the outcome of processing (as opposed to monitoring) with both sufficient precision and sufficient throughput becomes very challenging. The complex three dimensional nature of modern devices is posing severe challenges for all existing metrology techniques. High-resolution technologies such as e-Beam have throughput limitations, and high throughput technologies such as optical metrology have insufficient sensitivity for directly predicting device performance. One approach to solving this challenge is to start from a high throughput metrology and create proxy targets that are designed to be easily measurable and which provide multiple indirect measurements that somehow reflect the condition of the process and devices nearby the proxy targets. Such approaches retain low cost and high throughput advantages whilst circumventing dimensional or sensitivity limits. However, they entail making a number of decisions about how the proxy targets should be designed, how many there should be, where they should be placed and understanding how the signals from the targets relate to the properties of the devices that we wish to indirectly measure.

We consider proxy targets for optical metrology. There are many possible designs for such an optical proxy target. One approach is the PSR targets developed at Imec [1]. The technique can track pattern designs longitundinally through any processing step [2]. The footprint can be tailored depending on the microscope capabilities. The unit channel of information without any attempt to optimize the footprint is currently 7.3 µm × 7.3 µm (a 31 × 31 µm area can contain 18 targets). A picture of a group of targets is shown in Figure 1(a). The technique works by measuring asymmetrical targets composed on one hand of a process insensitive zone (typically a large feature) and on the other hand a process or device-performance sensitive zone (typically the device pattern itself or fine design-rule-compatible geometrical shapes) as shown in Figure 1(b). For a given design, any substantial change in a parameter of the process will affect the density of the fine patterns and hence the measured optical centroid position, as depicted in Figure 2. Thanks to the amplification obtained in the design, the centroid between the two zones is detected with a precision below 0.2 nanometers whilst its variation within

© 2023 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/



(a) Optical micrograph of a collection of PSR targets, showing both vertical and horizontal targets. The targets are laid out in a 7 × 7 grid of pin-wheel shapes giving 48 pin-wheels (one is removed at the centre of the grid), each giving two readings from two aligned pairs of features.

(b) Scanning Electron Microscope (SEM) micrograph of one half of a PSR target, showing the process-insensitive area on the left and the patterned area on the right. This is half of a vertically aligned pair in a pin-wheel.

Fig. 1. Micrographs of PSR targets.

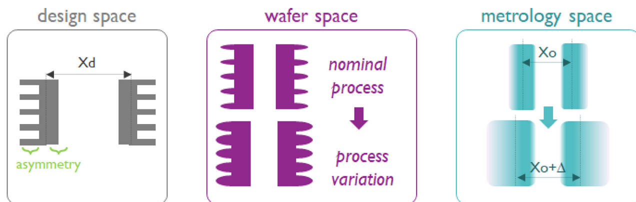


Fig. 2. The basic principle of the Pattern Shift Response technique in an image-based metrology example (optical microscope). The sketch describes the example of a single channel mark made of horizontal fine patterns represented in the 3 relevant spaces of design (GDS), wafer (top-down post exposure and development) and metrology (top-down sub-resolved image with visible wavelength). X_0 is the distance measured by the microscope between the effective centroids of the mark image. Δ is the measured shift of that distance indicating some aspects of the process have changed.

the process window is often up to tens of nanometers. To give an example, natural process Critical Dimension (CD) variation on a monitor uniform wafer (Scanning Electron Microscope (SEM) wafer CD) is approx. 0.2 nm, and the detected PSR centroid position variation (Δ in Figure 2) range spans 18.9 nm for some targets.

The detection is made by using state of the art in-line optical microscopes (in this case a KLA ARCHER700 overlay metrology tool). Imec has developed various different families of patterns and variations within a family [3]. For a given type of device however it is an open question as to what pattern or combination of patterns would work best to monitor and predict variation, how the signals from different PSR patterns should be combined, and what the achievable precision on final chip performance is.

In this paper, we give a first answer to these questions. We use the Imec N7 (iN7) BEOL platform [4] and analyse data from wafers with a large number of different PSR targets and a long meander electrical target. We then investigate how to combine the different signals provided by the PSR targets using machine learning models in order to predict the final resistance of the meander. We also consider how having a relatively low number of wafers with which to train the machine learning models impacts these experiments, and give an indication of the potential accuracy of the approach when based on a larger dataset, such as may be collected during high-volume manufacturing.

A. Related Work

1) Metrology: In [5], [6], the authors develop a method for Optical CD (OCD) metrology that is related to the standard scatterometry approach. They apply machine learning techniques in order to avoid the model development effort and approximations associated with standard OCD. They measure devices on the die directly, and target SRAM as an application. The machine learning model used is a single-layer neural network, i.e., a linear transform with one element-wise nonlinearity, and as such is significantly simpler than the ML models that we use. Their work differs from ours in that they measure physical device characteristics such as profile height rather than directly predicting electrical performance, and they develop their model based on Focus-Exposure Matrix wafers rather than Process Of Record wafers. The proposed application is dose and focus control of lithography. It is not clear how well this approach would apply to the BEOL applications that we target and prediction of electrical measurements. Their discussion of inter-die, inter-wafer and inter-lot variation is limited.

In [7], the authors use optical scatterometry to develop a model to predict electrical properties for a Litho-Etch-Litho-Etch process, considering both resistance and capacitance of certain test structures. They achieve good prediction scores for the electrical properties, however, there is no description of the machine learning technique used, and the training and test data come from a single batch, so it's not clear how well inter-wafer and inter-lot variation is taken into account. By contrast, we are targeting a smaller geometry, we use more wafers, two lots, and we provide more analysis of the possible impact of inter-wafer variation on the building of machine learning models. We also explain the machine learning techniques we are using.

Reference [8] is similar to [7], in that the authors develop a method to directly measure structures on the die using scatterometry readings as an input to a machine learning model. They also provide no information about the machine learning technique used, and few details about how the method is trained. In contrast to [7], they are targeting processing for DRAM and the target is critical dimension control through dose and focus adjustment rather than prediction of final electrical performance. The experimental work uses seven wafers.

2) Machine Learning: ML has been used for many applications in semiconductor manufacturing, with a history going back more that 30 years (e.g., [9]). Major areas include defect classification from SEM measurements [10], run-to-run control of tools [11], identification of problems with and improvement of lithography [12], as well as quality control, predictive maintenance, virtual metrology, decision support, and production planning and scheduling [13]. This use of ML has important overlap with other advanced manufacturing use-cases such as the production of cars [14]. The area that

is most relevant to the work presented in this article is yield prediction.

Semiconductor manufacturing can be roughly split into 4 steps: Wafer Production, Wafer Test (WT), Die Packaging and Final Test. At the end of wafer production, there is usually a Wafer Acceptance Test (WAT) to check some simple electrical test structures associated with a small number of dies on the wafer, to ensure the wafer was processed correctly in the fab. WT usually tests all dies for circuit functionality with test patterns and marks failed dies, which are then not packaged. Final test is applied as the last check after dicing and packaging, and again tests for circuit functionality. Manufacturing thus gives rise to 3 different types of yield loss: line yield loss (wafers discarded before or at WAT), die yield loss (dies discarded after WT), and final test yield loss (entire packages discarded after final testing).

Yield prediction is usually applied to predict aggregate die yield loss, and sometimes final test yield loss. Yield prediction models are useful for a number of reasons. Firstly, WT can be expensive and time consuming. If a wafer will have high die yield, it will be cheaper to skip WT and package all the dies as the waste of packaging a small number of bad dies will be less that the cost of WT itself. If the wafer will have a lower yield, then it will be cheaper to filter first using WT. Secondly, if an individual die can be predicted to fail Final Test then it shouldn't be packaged. Thirdly, a predictive model for yield can help with a) doing Root Cause Analysis (RCA) for unexpected poor yield and b) the closely related task of yield optimization, by identifying which model inputs have the biggest effect on yield and using that information as a basis to start RCA or yield optimization.

There have been many previous works on predicting aggregate WT yield for a wafer. The most commonly given motivation is RCA/optimization, but skipping WT is also discussed [15]. The use of yield prediction in RCA is somewhat similar to the large mount of work on the spatial aspects of failure patterns for RCA (e.g., [16], [17]), but we do not consider spatial patterns here. Some models are purely based on WAT measurements (e.g., [15], [18], [19], [20], [21]). Some work has been done on predicting final test yield [17], [22], [23]. Whilst WAT measurements are the most common set of input features, the logs of processing machines and output of metrology steps can also be used [24], [25], [26], [27]. Note that machine logs often consist of time series data, which is much harder to model, and that metrology is often (very) sparse. There has also been work on feature selection to improve the accuracy of yield prediction [18], [20], [26].

Theoretically, if a wafer can be predicted to fail WAT early enough then the cost of further processing can be skipped by discarding it early. This is rarely mentioned in the literature, probably because catastrophic whole wafer failure is a) unlikely in a reasonably mature process and b) would probably be caught by existing concrete metrology steps (e.g., [28]). Reliable die (rather than wafer) level yield prediction during wafer processing would allow a more fine-grained approach: if the die is shown to be bad after a step that prevents reworking, then further lithography steps could be skipped to relieve the pressure on litho machinery, the low throughput of which is problematic, especially for Extreme Ultraviolet (EUV). However, building reliable die-level models for steps before WT has been hard until now due to the limits on existing metrology, which make it either sparse or low accuracy, as discussed above.

Fig. 3. Examples of the geometrical shapes used in the PSR mark designs. The large process insensitive zone is on the left of each design, and the process-sensitive finely patterned zone is shown on the right.

This work describes the selection of physical targets to provide features that are full coverage (rather than sparse) and that can be measured longitudinally (i.e., measurements of the same target taken at multiple different points in processing), aren't time series features, and whose measurements are related to WAT features that are used in various current yield prediction models. Unlike WAT they are not sparse, and thus should give better aggregate WT yield prediction, and also better per die WT yield prediction for further fine-grained WT optimization. Furthermore they give much more insight into the evolution of the wafer during processing to enable better RCA and yield optimization. Also, PSR information is available before the end of wafer processing, to enable early skipping of bad dies, by for example skipping the exposures necessary for subsequent metal layers.

II. Method

A. TITAN Platform Data

The TITAN platform consists of a die design containing various test structures in two metal layers. The first of these layers contains standard test structures such as short wires, long wires and meanders. The layer is produced in several macro steps, being deposition of the stack, using EUV lithography with a 42-nanometer pitch to pattern, followed by etching and metallisation, i.e., copper fill followed by CMP. In between these processing steps, various metrology measurements are made, including the PSR measurements.

The organisation of the PSR targets on a TITAN die is as follows: there are five sub-die locations, with each location containing 13 different target design families, with approximately 18 design variations each. This leads to a total of $5 \times 13 \times 18 = 1170$ possible PSR readings per die. The 13 families contain fine patterns such as trenches, elongated contact holes (short trenches) or elongated triangles; cf. Figure 3. Tone reversed versions are also present in the input readings. The key pitches at stake range from 42 nanometers to 120 nanometers, including various pattern width variations. The initial selection of PSR targets used in this study is based on a wide selection of design shapes and dimensions from the available PSR designs developed by Imec, but with a limited total number of designs.

The target values to predict, being the electrical measurements, occur once per die as the platform only has one copy of each electrical test structure. In these experiments

we report on only one electrical value to avoid verbosity: the resistance of a 1-centimeter-long meander at the minimum 42-nanometer pitch. In our experiments the device resistance ranged approx. from 60 to 140 $\Omega.\mu m^{-1}$. The TITAN platform has a limited number of electrical measurements, and we considered meander resistance to be the most useful of the available ones as the resistance of long wires in the BEOL can have a major impact on a design. We performed experiments with other electrical targets, including resistance of a short wire and a short meander, but these either produced less good predictive models, or were considered to be a less useful quantity to predict. Test device resistance is also one of the WAT values used to predict aggregate wafer yield (mostly at WT) in other articles (see references in Section I-A) and so should be relevant to yield. Yield prediction is our ultimate aim, but, to build such models we need datasets with WT results which we don't currently have, so meander resistance is the best proxy on this platform to illustrate our ideas.

The TITAN die design is copied out on multiple wafers, with multiple dies per wafer, to make the dataset. Our dataset consists of two lots, with the first lot having all 20 wafers with measurements, and the second lot having 8 out of 20 wafers with measurements, for a total of 28 wafers.

A smaller number of wafers were available for the second lot due to the processing requirements of the project in which the wafer processing was done; the majority of the wafers in the second lot were shipped to partners for other metrology experiments before electrical measurements could be made, and so cannot be used for modelling as the dependent variable of interest was not known.

Of the 28 wafers, 23 have 178 dies with measurements, and the remaining 5 have 165 dies with measurements. The steps at which PSR measurements were made are after lithography (which we refer to as ADI) and after CMP for our experiments.

TABLE I
DATA SET SIZE AND FEATURE COMPLETENESS

Measurement: max. number available	Lot 1	Lot 2	Lot 1+2
Total number of Wafers	20	8	28
Max PSR familes at ADI	13	13	13
Max PSR familes at CMP	11	13	11
Max sub-die locations at ADI	5	1	1
Max sub-die locations at CMP	2	1	1

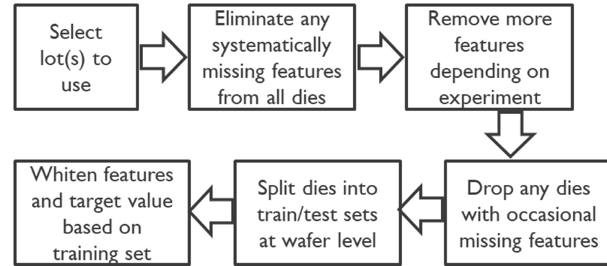


Fig. 4. Flow chart summarizing data selection and preparation before the training set is fed to the model training procedure.

B. Lot-Specific Data

Not all measurements were made in the 28 wafers used for the experiment due to time limits for making metrology measurements and limited access to metrology tools, tool downtime etc. More specifically, for the first lot, for the ADI measurements, there are two target families that don't have full coverage (at some sub-die locations in some wafers they are either entirely missing or with only a subset of variants available). We refer to this as *occasional* missing data (probably due to a transient tool failure) as it is there for most dies. For the CMP measurements, there is by contrast *systematic* missing data: only 11 of the 13 target families were measured, and only at 2 die locations out of the 5 (for all dies on all wafers in this lot). In addition, there is also some occasional missing data for CMP: there was a failure to measure a target family on one of the wafers, which thus has only 10 target families at 2 sub-die locations.

For the second lot, there are large parts of systematic missing data for ADI: 2 wafers of the 8 where only one sub-die location was measured, and 3 wafers with no ADI PSR measurements at all. For the CMP measurements there is also systematic missing data: there are 5 wafers with only one sub-die location measured, with the other 3 having all sub-die locations measured. This means that the second lot can only be used with one sub-die location for both ADI and CMP, and that the wafers with no ADI measurements had to be dropped completely when ADI features were being used. In the second lot however, there is no occasional missing data.

For occasional missing data, we drop any dies that don't have the full set of features (i.e., PSR measurements). For systematic missing data, we trim the set of features used in the dataset to the largest common subset defined by what features are available across all lots/wafers/dies to avoid the dataset becoming too small (i.e., we drop features to make sure there are enough records available with full coverage of the remaining features), unless a wafer has no features of a certain type available in which case the wafer is dropped to avoid the subset becoming empty. For experiments involving both lots, this means for example only using the 11 target families available in lot 1 for CMP when doing experiments that rely on CMP features. Experiments using only ADI features can use all target families as they are all available in both lots. Similar logic applies for the available sub-die sites; this means that for experiments using data from both lots only one subdie location is used in ADI and CMP experiments. Available data and feature types are given in Table I. The selection process, train/test splitting and data pre-processing (whitening) is summarized in Figure 4.

C. Data Hierarchy and Flattening

The dataset forms quite a deep hierarchy, as shown in Figure 5. The existence of these hierarchical levels is important due to the possibility of correlations between the properties of devices that are grouped together at a given level of the hierarchy. For example, for many silicon processing steps, it is normal for there to be a wafer-level effect. Many processing steps operate on an entire wafer at once (e.g., etching, CMP etc.) and there are slight differences in the processing applied to different wafers. This can result in

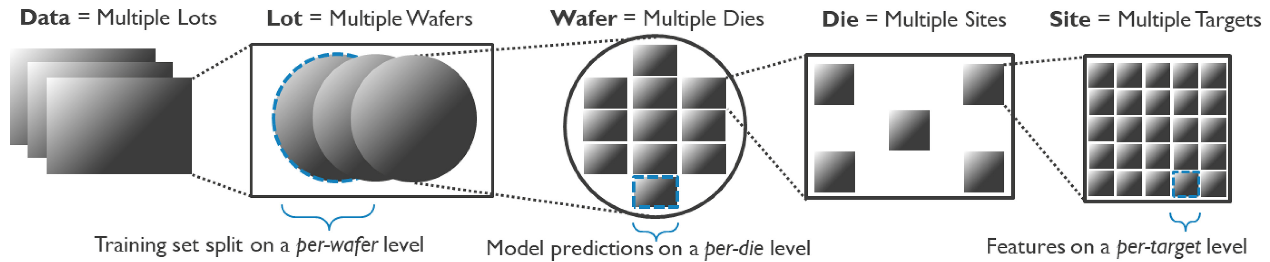


Fig. 5. The hierarchy in the dataset, also showing the grouping levels where train-test split and data flattening occur, and the level where individual features are found.

a global effect for all devices (in all dies, at all sub-die locations) on that wafer. The simplest example of such a result would be a shift in the mean for a given target measurement, e.g., resistance, for a given wafer; different wafers end up with different mean resistance values. Other changes are also possible, including a change in the variance or some distortion of the expected wafer-level signature of the processing step. Similarly, there can be some drift visible at the lot level if there is a sufficiently long gap between the processing of two lots or if some major change happens to the tools between the two lots, and there can be die-to-die variations within a wafer due to variations in lithography, e.g., exposure and focus.

Such a hierarchical dataset requires certain decisions about how to model it. Mainstream supervised machine learning techniques are designed to work with flat datasets, where the data consists of one separate record per item with a number of features describing that item and a class label or target value that is to be predicted. Consequently, we must choose an approach to remove the hierarchy and flatten the dataset whilst taking care not to introduce any data leaks into the model training procedure by incorrectly ignoring the parts of the hierarchy that we do not represent but which carry correlations.

The simplest approach to avoiding leaks would be to flatten the data at the top level of the hierarchy: each object then becomes an entire lot, and all the properties of all dies in each new lot would be predicted in one go. However, this introduces other problems. Given that the data is collected from advanced experimental silicon processing platforms where we can only do a limited number of runs, the number of items at the higher level of the dataset hierarchy is on the low or very low side for these experiments. Typical datasets for building machine learning models run to tens of thousands of items, and certain techniques such as deep learning can easily require four or five orders of magnitude more than this. In this dataset, the top two levels of the hierarchy have 2 and 28 items respectively, which is far too small. In addition, making predictions at the level of an entire lot or an entire wafer entails predicting multiple outputs per item, as there is at least one electrical value to predict per die (and many dies per wafer/lot). Whilst predicting multiple class or regression outputs is possible with some machine learning techniques, this is already more challenging than the standard case of predicting a single output. As a result we decided to flatten the dataset at the level of dies, which results in a total of 4919 items, which is somewhat on the low side, but hopefully still large enough to learn a non-linear model with reasonable predictive performance. Also, this choice means there is only one value to predict per die, which sits well with standard ML regression algorithms that predict one output, and the model still corresponds to the ultimate use-case of die-level yield prediction.

As described earlier, when flattening the dataset we need to take care to take into account the removed hierarchy levels when handling the data. In a production environment, PSR targets would be used to predict electrical measurements for each die on a wafer for which there are no electrical measurements available at all (because the wafer hasn't yet reached the measurement step). As such, we cannot simply assign items (dies) from the dataset randomly to training and test sets as would usually be done, because there is a high likelihood that for any given die item in the test set, at least one die from the same wafer would also appear in the training set. This data leak would then bias the model and give an overoptimistic impression of the achievable performance because electrical information about all of the wafers in the dataset is leaking into the training set through the way the dies are split. For example, it is entirely possible that different wafers have a different global mean for a given electrical measurement. This global mean will be unknown in a real production environment, because when the prediction of electrical properties is made based on PSR targets, there are no electrical measurements available for that wafer at all. If, for every wafer that appears in the test set, there are also some number of items from that wafer in the training set, then the machine learning technique will immediately have access to an approximate global mean for the target value, which it would not normally be able to see. Thus the model prediction would seem more accurate than is possible in practice due to this biasing, which results from incorrectly ignoring correlations across a level of the dataset hierarchy.

In order to avoid this problem, we allocate items to training and test sets based on its membership to a higher level of the data hierarchy than the level at which we did the flattening. We have chosen to do train/test splitting at the wafer level, so all dies from any given wafer can only appear either in the training set or the test set, but not in both, to avoid leaking wafer-level information from the training set into the test set. Ideally we would extend this to the top of the data hierarchy, i.e., the lot level. However, there are only two lots, one of which is significantly larger than the other, and we estimate that lot level effects are likely to be minor, unlike wafer-level effects. Thus, we consider our assessment of model performance to be relatively unbiased. It is also probably somewhat pessimistic;

TABLE II
AUTOMATIC FEATURE SELECTOR METHOD DESCRIPTIONS

Name	Description	SciKitLearn routines
AFS0	Model-based: first fit the XGBoost model to the whole dataset, then select features based on its reported importance weights	`SelectFromModel`
AFS1	Sequential: pre-filter features down to 64 using mutual information regression, then do sequential feature selection	`SelectKBest`, `SequentialFeatureSelector`
AFS2	Recursive: pre-filter features down to 64 using F-regression, then do recursive feature elimination	`SelectKBest`, `RFE`

given that our dataset is relatively small even at the die level, we would expect improved model performance if we had a larger dataset to train the model with.

D. Modelling Approach and Experiments

a) Features: As well as the PSR measurements (18 per family per sub-die location, up to 5 sub-die locations within a die), the X, Y and radial coordinates of the die position on the wafer are provided as features. Lot and wafer numbers are not provided as features. Die coordinates are provided to allow the model to adjust for within wafer location-dependent effects; for example, the speed of etching may vary with the distance to the centre of a wafer depending on the process conditions [29].

b) Dataset Splitting and Cross Validation: As previously described, the data is split into training and test sets at the level of wafers. Lot membership is ignored. To generate predictive performance metrics we perform five-fold cross validation and take the mean average of the results across folds.

c) Whitening: For each fold, after the splitting of data into training and test sets, a whitening transformation is fitted to the PSR features and target values of the training set and used to transform both the training and test sets. This data transformation is inverted after making predictions so that model output is in the original units of the dataset.

d) Models: On the iN7 experimental platform, there are many more PSR targets than would be acceptable on a die for an actual chip product; each PSR target takes up some area, which translates into extra cost. To see what is achievable with a lower number of PSR targets, we apply feature selection. We apply machine learning models to the data to investigate different approaches to the selection of features and analysis of target design families.

The first main approach to feature selection is to group features together into their target families (i.e., groups of 18 variants) and select one or more target families to make up the PSR feature vector for an experiment. For the first method, we take target families individually. For the second method, a small number of families are chosen at random and combined. The chosen PSR feature vector is then always concatenated with the three wafer location coordinates. After this feature selection, XGBoost [30] with default settings is used to build a predictive model.

The rationale behind this approach is to try to understand whether there are dominant individual or combinations of target design style that lead to better predictive performance and to give some indication of the extent to which extra features can improve the predictive power, how quickly the performance increases as the number of families used increases, and whether there is a saturation in achievable performance. Note that even for the first round of experiments, three target families is already 50% bigger than we expect to be an acceptable number of PSR targets to embed within a die, and that these larger feature vectors are used to understand the possible improvements over a single family.

The second main approach is to apply an Automatic Feature Selector (AFS) to the entire collection of available features; that is, (up to) $13 \times 18 = 234$ PSR readings per sub-die location, with all the available sub-die locations for that experiment being concatenated, and the within wafer location features. The aim here is to search with a finer granularity to see if there are any combinations of individual target variants that happen to work well together, and as such is potentially more rigorous than choosing features at the level of whole target families. However, finding optimum feature combinations from the large number of PSR targets is hard, especially given the high levels of correlation of the PSR readings, and so there is no guarantee that this approach will outperform the simpler method of picking whole families.

For AFS we select 36 features, a number equivalent to 2 groups of 18 targets, which we think is a reasonable estimate for a number of PSR targets that can be put on a die without incurring a significant silicon area penalty. The three different styles of feature selection we use are described in Table II; apart from setting the number of features for pre-filtering and the pre-filter metric used, all function parameters were left at default settings. Unfortunately, the time required to apply sequential feature selection or recursive feature elimination to the full feature vectors was extremely long, with individual experiments sometimes taking multiple days. This is why we use pre-filtering. Whilst this speeds up those methods enough to make it usable, the quality of the feature selection is probably reduced. This is acceptable for a proof-of-concept work such as this, but a more stringent approach should be adopted when making a production model.

Once we have applied the AFS step, we again apply XGBoost with default settings to build the predictive model. In the rest of this article the names AFS0 etc. refer to the pipeline of the AFS method followed by the XGBoost model.

We report the R2 and root mean-squared error (RMSE) predictive performance metrics. R2 is dimensionless, and RMSE is in $\Omega.\mu m^{-1}$. We aim to do experiments on the combined data from both lots where possible, to maximise the size of the dataset and show that reasonable performance can be achieved across both lots. In addition, we aim to do experiments based on features available at ADI, and at CMP, to better understand how early in the silicon processing

flow it is possible to make predictions about the eventual electrical measurements. Being able to predict earlier is better; for example, prediction of problems after ADI could trigger rework of the litho on the wafer before any further irreversible processing is carried out. However, it is likely that there is more information later on in the processing flow which will make predictions more accurate. In addition, we also do a final set of experiments based on a combination of features available at ADI and CMP, to see whether the longitudinal evolution of features across different processing steps can help make predictive models more accurate. This combination doubles the number of PSR features available without increasing the number of targets used on the die.

Given the systematic missing data in the dataset, it is only possible to do experiments across lots and across ADI/CMP by using information at a single sub-die location that is measured everywhere. Hence, this is the approach that we take in the first round of experiments. Whilst this may seem like an artificial restriction on the amount of data that is available to build the machine learning model, as noted earlier the acceptable number of targets that can be added to a die in practice is limited. In addition, given that the same PSR targets are replicated across the multiple sub-die sites, it is not clear how much extra information will be gained by including extra sub-die sites as input to the machine learning model.

We do not include ADI + CMP results for the AFSs, because the standard feature selection algorithms work on a flat collection of features. To properly select a set of feature based on sub-die location, family and variant, and then include both the ADI and CMP readings for that PSR target would require a customised implementation of the feature selection algorithms. Whilst we think this avenue is promising, we did not have time to fully pursue it in these experiments.

In order to check whether using multiple sub-die sites may actually gain anything, we perform a second round of experiments using only the first lot, and using all available sub-die sites as part of the feature vectors (5 sub-die sites for ADI, 2 for CMP). We apply AFS in this case. We also apply all three methods (single family, multiple family and AFS) to a single sub-die site on the same single lot to compare with the multiple sub-die location experiment. Given that the dataset is smaller and comes from only one lot, the results from these experiment should be treated with more caution.

1) Single Target Family: For these experiments, all 18 variants of a single target family are used to make the PSR feature vector. For the experiments across both lots, the final feature vector is thus 21 long after adding wafer coordinates. For the single lot experiments with multiple sub-die sites, the final feature vector after adding coordinates is $(5 \times 18) + 3 = 93$ long (ADI) or $(2 \times 18) + 3 = 39$ long (CMP). For the experiments where both ADI and CMP features are combined, the figures are doubled.

2) Mix of Target Families: For these experiments we randomly pick between three and seven target families. Seven is an arbitrary upper bound to limit the amount of experiments being done, with the expectation that any gain would already be marginal for that size of feature vector. Three was taken as a lower bound as it is larger than both the single family and AFS experiments, and so takes into account the potential gain for feature vector sizes that they don't cover. This leads to a PSR feature vector that is between $3 \times 18 = 54$ and $7 \times 18 = 126$ long for the first round of experiments with a single sub-die site. For the experiments where both ADI and CMP features are combined, these figures are doubled. For the second round of experiments with either 5 (ADI) or 2 (CMP) sub-die sites, the feature vectors are thus between $2 \times 3 \times 18 = 108$ and $5 \times 7 \times 18 = 630$ long.

3) Automatically Selected Target Variants: The input for these experiments is all available PSR features, giving feature vectors between $13 \times 18 = 234$ and $5 \times 13 \times 18 = 1170$ long. Note though that the feature selection part of the pipeline only keeps 36 of these features to pass to the main model building step; this is equivalent to the size of two target families at one location, which we believe is a reasonable area budget to aim for.

E. Larger Training Sets

Process monitoring using PSR targets is intended for use in High-Volume Manufacturing (HVM). Fully developing the technique for such a setting would involve collecting data from a much larger number of wafers with PSR targets added, and we expect that achievable predictive performance would increase significantly with a multi-thousand wafer dataset as compared to the model that we currently have, which is built with a very modest size dataset.

The hierarchical nature of the dataset exacerbates the general problem of a lack of data. We have chosen to build our model using the die level of the dataset hierarchy. However, we strongly suspect that there are modellable wafer-level effects in processing. Being able to properly model these effects would require a much larger number of wafers in the dataset, *and* may also require wafer-level features and/or an explicitly hierarchical model that tries to separate wafer-level effects from die-level effects.

Although we do not currently have enough wafers, and may not have the right features and/or machine learning techniques to model wafer-level effects, we consider it worthwhile to try to estimate how much performance we are losing due to these factors. To do this, we reran some of the experiments with one or both of the following changes. Firstly, we don't enforce full separation of wafers between training and test set, and allocate dies from all wafers to the training and test sets at random (i.e., *wafer split* is false rather than true). Secondly, we normalise both the features and target value on each wafer by subtracting the wafer mean before splitting the data into training and test sets (we call this *removing the offset*). The idea behind the first approach is to give the machine learning method access to data that would allow it to try to learn something about wafer-level differences of all wafers in the dataset. The idea behind the second approach is to assume that the wafer-level differences are very simple (i.e., just a mean shift) and then remove them entirely, leaving the machine learning method to learn how to model the differences at the die level. This is roughly equivalent to assuming that the machine learning method can factor out a simple wafer-level difference by learning how to

TABLE III
SINGLE PSR FAMILY RESULTS

Design Family	Step	Metric	
		R2	*RMSE*
Family A	ADI	0.24	7.0
Family D		0.17	7.3
Family E		0.16	7.3
Family F		0.15	7.3
Family D	CMP	0.57	5.0
Family B		0.55	5.3
Family C		0.54	5.4
Family F		0.52	5.3
Family D	ADI+CMP	0.61	4.8
Family C		0.57	5.1
Family A		0.55	5.0
Family B		0.53	5.4

TABLE IV
MULTIPLE PSR FAMILY RESULTS

Number of Design Families	Step	Metric	
		R2	*RMSE*
6	ADI	0.24	7.0
5		0.24	7.1
3		0.24	6.7
3		0.23	6.8
6	CMP	0.60	5.0
5		0.60	5.1
3		0.59	4.9
4		0.58	5.0
4	ADI+CMP	0.62	4.6
5		0.60	4.9
5		0.60	4.8
5		0.60	4.9

TABLE V
AFS RESULTS

Feature selection algorithm	Step	Metric	
		R2	*RMSE*
AFS0	ADI	0.12	7.2
AFS1		-0.03	7.7
AFS2		-0.15	7.8
AFS0	CMP	0.57	4.9
AFS2		0.53	5.1
AFS1		0.51	5.2

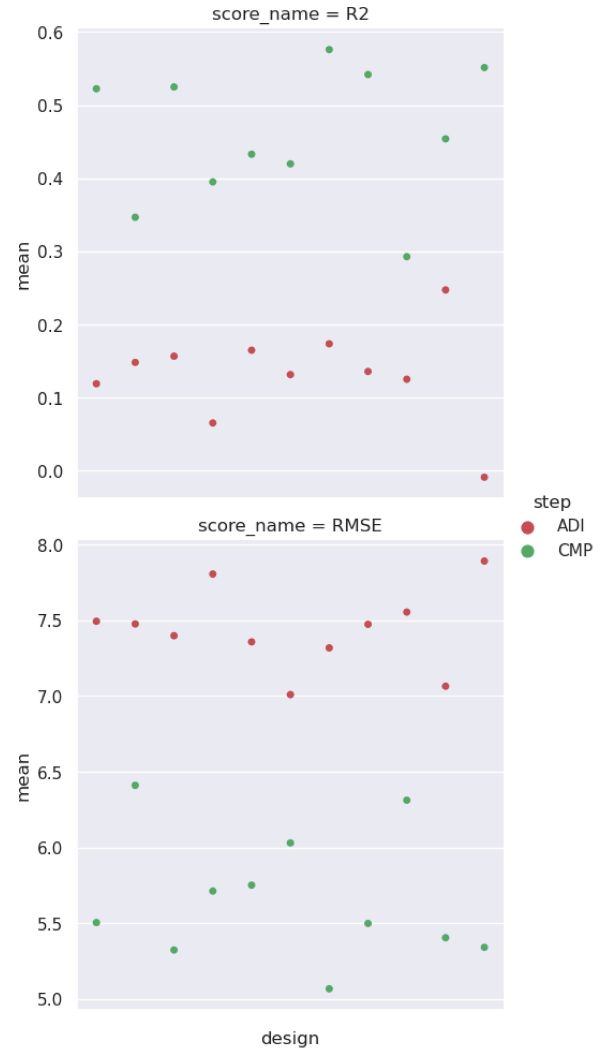


Fig. 6. Performance of models built on different individual PSR target families, using readings at different points in wafer processing (ADI/CMP). Both R2 (above, higher is better, dimensionless) and RMSE (below, lower is better, $\Omega.\mu m^{-1}$) are shown. The mean of the score across folds is used. Design name has been omitted for clarity.

predict it if it is given suitable wafer-level features and enough data.

III. RESULTS

A. Single Sub-Die Location, Both Lots

Tables III, IV and V show the top ranked results for the three different methods of single target family, multiple combined target families and AFS respectively. These results are on both lots, with a single sub-die location. Figure 6 shows the performance of different individual PSR families to show relative performance, and also gives the results across models built on only the ADI and CMP steps to show the degree of variation possible in the results and the general trend that models built on CMP readings are significantly better than those built on ADI readings. There is a similar level of variability and similar split across ADI and CMP steps for the models using multiple PSR families (not shown), although the absolute performance of multiple PSR family models is better.

Figure 7 shows how predictions correlate with actual values for individual dies, for the best type of model (i.e., Multiple PSR ADI+CMP with 4 design families). We can immediately see that a large proportion of the less good predictions (i.e., further from the line) come from the two wafers that are first in their lots. These two wafers appear to be harder to predict than the others and are bringing down the reported performance of the model.

B. Lot a Only, All Sub-Die Locations

Table VI shows the top ranked results for AFS for a single lot, using all available sub-die locations, being 5 for ADI, 2 for CMP.

C. Lot a Only, Single Sub-Die Location

Tables VII, VIII and IX show the top ranked results for the three different methods of single target family, multiple combined target family and AFS respectively for a single lot,

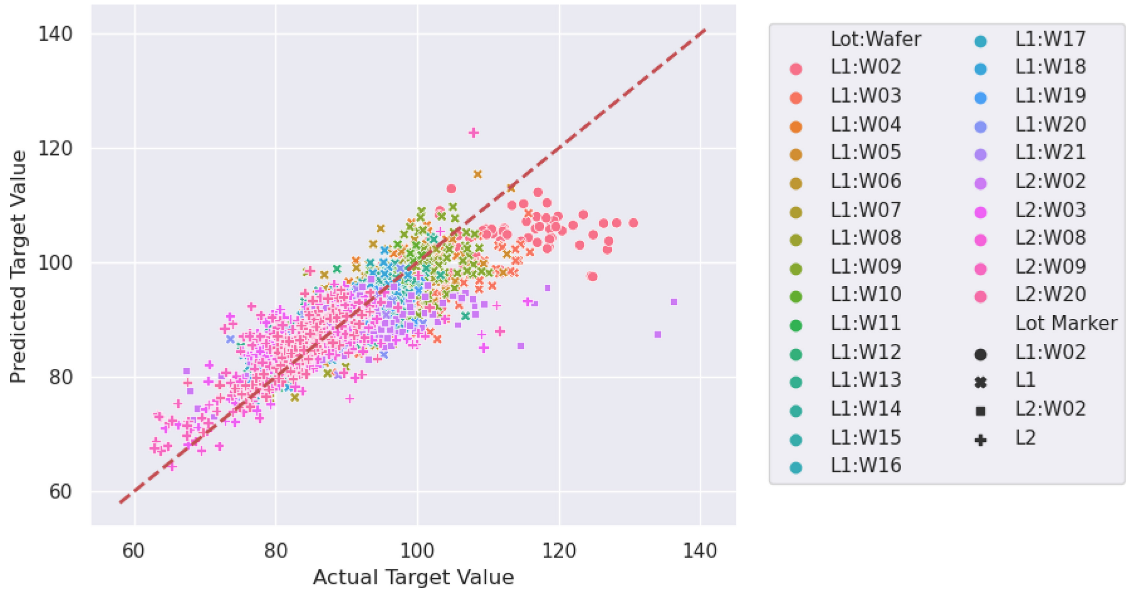


Fig. 7. A plot comparing predicted value of resistance ($\Omega.\mu m^{-1}$) with actual value (also ($\Omega.\mu m^{-1}$) for each die for the general best model (fold-averaged result) built across multiple wafers using ADI+CMP features as input. Die predictions have a different colour per wafer. The two lots have a point marker that are common to all wafers in that lot (circle for lot 1 and square for lot 2), except for the first wafer in each lot which has its own specific marker, namely a diagonal cross for wafer L1:W02 from lot 1 and a straight cross for wafer L2:W02 from lot 2). The dashed line shows the ideal relationship.

TABLE VI
AFS RESULTS (MULTIPLE LOCATIONS, SINGLE LOT)

Feature selection algorithm	Step	Metric	
		R2	*RMSE*
AFS0	ADI	-0.14	6.5
AFS1		-0.19	6.5
AFS2		-0.28	6.7
AFS0	CMP	0.63	3.9
AFS1		0.58	4.1
AFS2		0.55	4.2

TABLE VII
SINGLE PSR FAMILY RESULTS (SINGLE LOCATION, SINGLE LOT)

Design Family	Step	Metric	
		R2	*RMSE*
Family G	ADI	-0.14	6.6
Family A		0.05	6.6
Family D		-0.15	6.7
Family F		-0.17	6.8
Family D	CMP	0.58	4.4
Family F		0.56	4.6
Family B		0.56	4.5
Family H		0.56	4.5

TABLE VIII
MULTIPLE PSR FAMILY RESULTS (SINGLE LOCATION, SINGLE LOT)

Number of Design Families	Step	Metric	
		R2	*RMSE*
3	ADI	0.08	6.1
4		0.07	6.2
6		0.07	6.2
4		0.00	6.2
4	CMP	0.67	4.0
6		0.65	3.9
3		0.65	3.9
5		0.63	3.9

TABLE IX
AFS RESULTS (SINGLE LOCATION, SINGLE LOT)

Feature selection algorithm	Step	Metric	
		R2	*RMSE*
AFS0	ADI	-0.08	6.3
AFS1		-0.15	6.4
AFS2		-0.28	6.8
AFS0	CMP	0.53	4.2
AFS1		0.48	4.3
AFS2		0.47	4.3

using a single sub-die location, to compare with the multi-location results in Table VI.

D. Discussion

Figure 6 shows that the achievable results using different families can vary significantly, and as such selecting the right targets to use is important. Furthermore, Table III shows that certain families occur repeatedly in the top rankings across different measurement points (ADI, CMP etc). However, comparing with the results on multiple families in Table IV shows that there is no single family that contains all available information as combining across families gives slightly better results.

The results for features taken at different steps show some clear trends. CMP features seem to result in much better predictive models than ADI features. For the single and multiple family results, the longitudinal combination of ADI and CMP gives some reasonable extra improvement.

The trend for choice of features to use is much less clear. The results with a single target family are almost as good as the results using multiple families. When using multiple families, there is also no clear trend towards a larger number of families giving better results. Although the top-ranked result for both

ADI and CMP uses six families, the next best use less than six, and sometimes as low as three, and achieve quite similar performance. For ADI + CMP, the overall best model when modelling both lots uses only four families.

The automatic feature selector performs surprisingly poorly for ADI. In theory, a good feature selector should at least be able to match the results of the single target family as it has a feature budget that is larger than the number of features used by that approach. For CMP, it provides a very slight benefit over single family models, but it is not quite as good as the multiple family models.

Overall it seems that adding PSR features give slightly more information for this size of data set, but the benefit is low for the extra silicon area used. The absolute best number of PSR targets for this experiment appears to be larger than the "reasonable" area budget we allowed the AFS to use. However, the total gain of adding extra features is limited, and may be an artefact of a suboptimal approach to feature selection and/or model inaccuracies due to the low amount of training data.

Our results on this dataset give some evidence that there is a benefit to being able to combine information from multiple locations. To investigate this, we performed experiments using data from the first lot only, which has 5 locations available for ADI, and 2 for CMP. We forced the single family and multiple family approaches to use data from a single location only, as before. For the AFS, we allowed it to select features from all available locations. We also restricted it to one location only, as a control.

The results are not easy to interpret. For AFS, the results with multiple locations are significantly better than the results with a single location for CMP when looking at the single lot. When comparing against the multiple and single family results at CMP, the multi location AFS model is clearly better than the single family model (at a single location), but still not as good as the multiple family approach (at a single location). For ADI the results for multi locations could also be viewed as worse, but it is rather academic as both models are bad. The difference in benefit for CMP and ADI suggests that it may not be different readings *at different die locations* that are giving the benefit to the multi location model – we can't rule out that AFS is selecting multiple readings of the same design but at different locations, and that this is what is giving a more stable reading, with the placement at different locations being coincidental.

We can also compare the single lot results against the model built on both lots. For CMP, the single lot multi-location AFSs are better than the AFS result modeling both lots (at a single location). However, the single lot single location AFS results for CMP are worse than the multi lot single location model. For ADI, both single lot AFSs are worse than the model for both lots, and also worse than the multiple family approach. We can also look at the other methods of feature selection when changing from modeling both lots to modelling a single lot. The single family results (single lot) are worse for ADI, and marginally better for CMP. The multiple family results (single lot) are worse for ADI, and quite a bit better for CMP.

The fact that the CMP results for all models get better when changing to a single lot, except for the (single location) AFS,

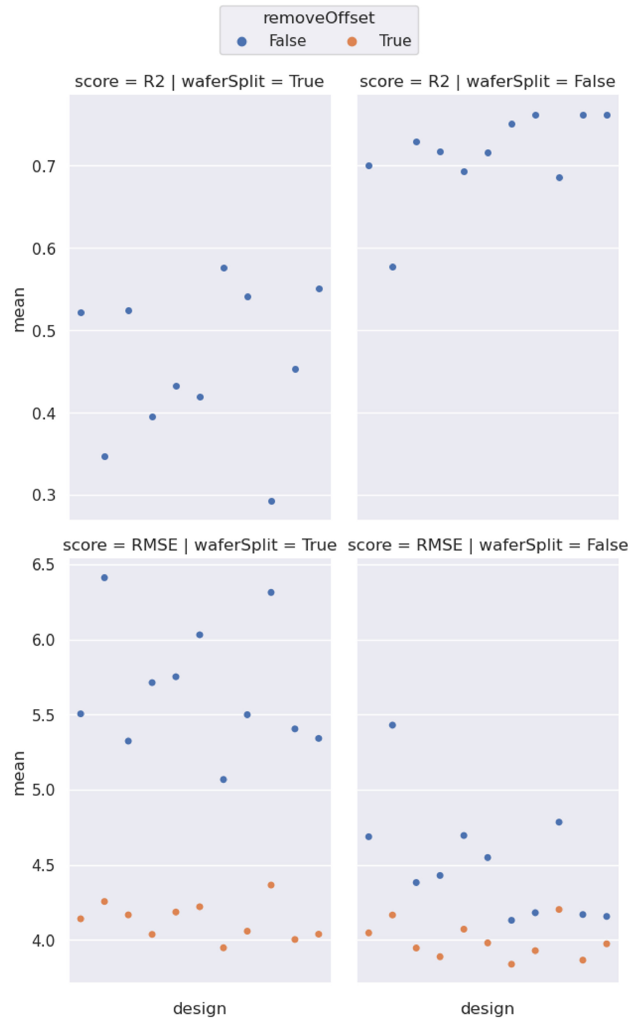


Fig. 8. Single PSR target family: comparison of the predictive power (different scores in different rows) of the original model compared to models built across wafers (*wafer split* is respectively *true* or *false*, in different columns) and/or with the wafer mean subtracted (*remove offset* is respectively *false* for the original model, or *true*, shown by different coloured points).

and the ADI results get *worse* is interesting, especially as ADI has more locations available than CMP for the automatic selector. Also interesting is the fact that the AFS (with multiple locations) leapfrogs the single family approach in performance when going from both lots to one lot. It's not clear what is driving this, and may be an artifact of small dataset sizes.

Across all the experiments, the drop in performance from the training set to the test set tends to be quite large; the R2 result on the training set is often 1.0 or very close to it. Normally this would indicate that the model is overfitting. However, our limited attempts to reduce possible overfit, including a hyperparameter search, have so far only resulted in lowering the performance on the test set. This should be further investigated.

E. Larger Training Sets

The results from these experiments are shown in Figure 8 for the single target family approach, and Figure 9 for the feature selection approach. In both cases we show the results for CMP targets only, to give a consistent reference point.

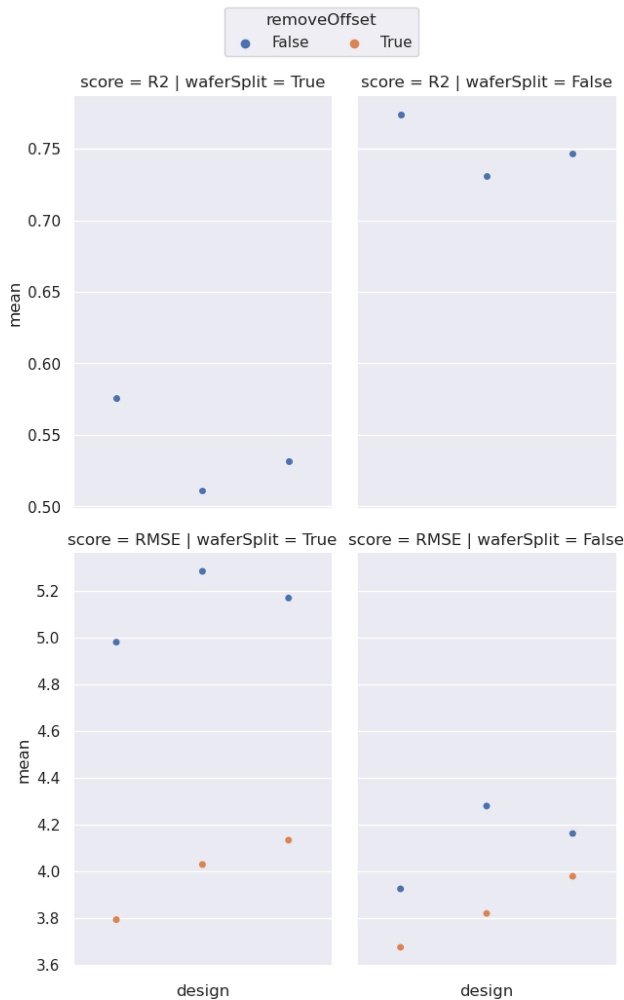


Fig. 9. PSR automatic feature selection: comparison of the predictive power (different scores in different rows) of the original model compared to models built across wafers (*wafer split* is respectively *true* or *false*, in different columns) and/or with the wafer mean subtracted (*remove offset* is respectively *false* for the original model, or *true*, shown by different coloured points).

The RMSE results are easiest to interpret here as they denote absolute error, which won't be changed by subtracting the mean unless the model has got better at predictions. Removing the offset gives a definite improvement, and so there is clearly some inter-wafer variation that the original model is not capturing. However, it doesn't say whether that is due to a lack of training data, a lack of wafer-level features to learn the relationship, or both. Removing the wafer split gives almost as much improvement to RMSE as removing the offset. Given that training examples aren't labelled with which wafer they come from, this indicates that the existing features are enough to get a pretty good estimate of the wafer-level effects; hence it would appear that the biggest problem is a lack of training data rather than a lack of informative features or model capacity to handle inter-wafer variation.

R2 is more complicated to analyse as the metric measures the global difficulty of regression as compared to the mean prediction. Flattening the wafer data by subtracting the offset will reduce the total range that the dataset covers, and so make it harder to get a good R2 score as compared to the original data. Hence we don't show R2 when the offset is removed. The results without removing the offset show that R2 is significantly improved when changing wafer split to *false*, with the best results being above 0.75. Given that RMSE is slightly better when removing offset as opposed to removing the wafer split, we would expect the final R2 of a model that can successfully model wafer-level variation to be even better than the best R2 in Figures 8 and 9.

IV. Conclusion and Future Work

This paper presents a proof of concept for selecting and using multiple PSR targets to predict electrical properties of structures on a die. The main overall result is that it is possible to build predictive models with a reasonable predictive accuracy using relatively few PSR targets and relatively little data, and that the choice of target types matters; this validates the main ideas of PSR Multiplexed Optical Metrology.

The best RMSE when using both lots is 4.6, which is approx. 5% average error with respect to the global target value mean (91.1, std dev. 7.9). The best R2 is 0.62. Whilst these results are not outstanding, given the limits of the experiments they are at the very least an indication that our approach has potential. Furthermore, the results in Section III-E suggest that this will improve with larger datasets and/or more sophisticated modelling techniques. Longitudinal measuring of PSR targets at several different stages of wafer processing adds useful information without costing extra area. We note that it is possible to get close to the absolute best results using techniques that require relatively few PSR targets, and that none of the better models used all the targets available. However, the current data set is too small to make any strong claims about the relationship between the number of targets and achievable accuracy. The results for fully automated feature selection are somewhat mixed. Finding a good AFS for this kind of problem will require further research.

The multi-location results are encouraging but not conclusive. For CMP, having 2 different locations appears to help, but there isn't any obvious benefit to have 5 locations to choose from for ADI. Further experiments are needed to investigate whether it is just having multiple copies of a particular PSR target that helps, or whether it is comparing PSR targets placed at different positions in the die that adds information. The multi-location experiments are worth expanding on as variation across a die can help to capture aspects of processing signature that varies across the wafer, and thus given a better informed model than just using die coordinates.

We believe that the measurements after CMP give better models as they incorporate more of the effects of the chain of processing steps. The fact that longitudinal ADI+CMP models do even better is interesting but we are not certain what is driving this. We don't currently know what is causing the poor predictive performance for the first wafers in the lots, and suspect that it might be an artefact of small training set size that will be alleviated with larger training sets.

We consider the best ways to extend this work to be the following:

A. Data

The main problem with the current state of the experiments is lack of sufficient appropriate data. We do not yet build full yield prediction models using PSR features as we don't have such a dataset, but it is a logical next step, and we have shown that PSR can be linked to electrical properties that resemble the WAT features currently used to build WT yield prediction models. Also, HVM data will probably look different from the less well tuned silicon processing steps available in an experimental fab such as the one we used at Imec, and collecting data from the context in which multi-PSR would be used is important to see how well it scales to HVM. The total amount of available data is also clearly important. Again, this points towards collecting HVM data; an experimental fab wouldn't usually run 1000s of the same lots, which is what will be needed to train the models properly. Of course, to do this an appropriate number of PSR targets would need to be added to an HVM mask, and this work is intended to help in scoping how many and which ones. If adding targets on each die is unacceptable, approaches such as putting targets on otherwise un-used areas on the periphery of the wafer can also be explored, to get better features for, e.g., aggregate WT yield prediction. Collecting such HVM data would be best done via an industrial partnership.

B. Early Prediction and Combining Stages

In this work we have compared ADI, CMP and ADI + CMP (albeit without AFS for the combined case). Given that the ADI results are weak and the CMP results much better, there is an open question about whether measuring the PSR targets at an intermediate point in processing between these two would add even more information without needing space for extra targets. The obvious place to start is at After Etch Inspection (AEI), but there may be other points in the flow worth trying. Similarly, the improvement achieved with ADI + CMP suggests that (as mentioned) AFS should be applied to this, but also points to the possibility of combining multiple measurements at points earlier in the flow to get a good predictive model. For example, ADI + AEI may give a good enough result at an earlier point than CMP. Having said that, for yield prediction it is probably enough to work with all information including CMP as the prediction would be most useful for skipping the next lithography step if the die can already be seen to be bad. Current predictive yield models use WAT data for aggregate yield, so any early individual die skipping for lithography or WT is already an advance on that.

C. Data Hierarchy

Based on this work, we assume that the most appropriate level to flatten the data hierarchy is likely to be the level at which the target to predict exists (here it is the die level). Flattening at a lower level is unlikely to make sense due to the assumptions the modelling algorithm will make about noise etc., and the fact that it amounts to a kind of crude feature selection by hand. Flattening at a higher level means picking an appropriate ML technique to predict multiple regression outputs, and having enough data to train such an algorithm. The possible role of hierarchical ML modelling should be explored here.

D. Models

XGBoost is a natural choice for this type of dataset. Nonetheless, a proper comparison against other techniques ought to be done for completeness, and a hyperparameter search should be done for XGBoost (and the AFSs). We have performed a limited auto-ML search for better algorithms using TPOT [31], and a limited hyperparameter search for XGBoost, but did yet find anything sufficiently better to be worth reporting; the goal of this article is to show a proof-of-concept rather than to establish the best performance achievable on a data set that is too small for such tuning to be meaningful.

In conclusion, PSR with Multiplexed Optical Metrology is a promising technique for predictive modeling use-cases like yield prediction, and further research on larger datasets will establish the achievable accuracy.

Acknowledgment

The authors would like to thank Stéphane Larivière for his integration work, Philippe Leray for his continued support and Bart Baudemprez for the advice on the optical microscope.

References

[1] V. Truffert, K. Ausschnitt, V. V. Nair, and K. D'Havé, "Novel monitoring of EUV litho cluster for manufacturing insertion," in *Extreme Ultraviolet (EUV) Lithography XI*, vol. 11323, N. M. Felix and A. Lio, Eds. Bellingham, WA, USA: SPIE, 2020, pp. 192–204. [Online]. Available: https://doi.org/10.1117/12.2551881

[2] K. Ausschnitt, V. Truffert, K. D'Have, and P. Leray, "Pattern shift response metrology," in *Photomask Technology 2018*, vol. 10810, E. E. Gallagher and J. H. Rankin, Eds. Bellingham, WA, USA: SPIE, 2018, pp. 126–136. [Online]. Available: https://doi.org/10.1117/12.2502353

[3] C. Ausschnitt and V. Truffert, "Metrology method for a semiconductor manufacturing process," U.S. Patent US10 656 535B2, May 19, 2020.

[4] S. Larivière et al., "Electrical comparison of iN7 EUV hybrid and EUV single patterning BEOL metal layers," in *Extreme Ultraviolet (EUV) Lithography IX*, vol. 10583, K. A. Goldberg, Ed. Bellingham, WA, USA: SPIE, 2018, pp. 206–218. [Online]. Available: https://doi.org/10.1117/12.2299389

[5] S. Pandev et al., "Signal response metrology (SRM): A new approach for lithography metrology," in *Metrology, Inspection, and Process Control for Microlithography XXIX*, vol. 9424, J. P. Cain and M. I. Sanchez, Eds. Bellingham, WA, USA: SPIE, 2015, pp. 570–583. [Online]. Available: https://doi.org/10.1117/12.2086056

[6] F. Fang et al., "Improving OCD time to solution using signal response metrology," in *Metrology, Inspection, and Process Control for Microlithography XXX*, vol. 9778, M. I. Sanchez, Ed. Bellingham, WA, USA: SPIE, 2016, pp. 51–60. [Online]. Available: https://doi.org/10.1117/12.2219775

[7] S. Das et al., "Machine learning for predictive electrical performance using OCD," in *Metrology, Inspection, and Process Control for Microlithography XXXIII*, vol. 10959, V. A. Ukraintsev and O. Adan, Eds. Bellingham, WA, USA: SPIE, 2019, pp. 71–79. [Online]. Available: https://doi.org/10.1117/12.2515806

[8] H. Lee et al., "Clean focus, dose and CD metrology for CD uniformity improvement," in *Metrology, Inspection, and Process Control for Microlithography XXXII*, vol. 10585, V. A. Ukraintsev, Ed. Bellingham, WA, USA: SPIE, 2018, pp. 545–553. [Online]. Available: https://doi.org/10.1117/12.2299976

[9] K. B. Irani, J. Cheng, U. M. Fayyad, and Z. Qian, "Application of machine learning techniques to semiconductor manufacturing," in *SPIE Proceedings*, M. M. Trivedi, Ed. Bellingham, WA, USA: SPIE, 1990. [Online]. Available: https://doi.org/10.1117/12.21147

[10] F. L. De La Rosa, R. Sánchez-Reolid, J. L. Gómez-Sirvent, R. Morales, and A. Fernández-Caballero, "A review on machine and deep learning for semiconductor defect classification in scanning electron microscope images," *Appl. Sci.*, vol. 11, no. 20, p. 9508, Oct. 2021. [Online]. Available: https://doi.org/10.3390/app11209508

[11] K. Liu, Y. Chen, T. Zhang, S. Tian, and X. Zhang, "A survey of run-to-run control for batch processes," *ISA Trans.*, vol. 83, pp. 107–125, Dec. 2018. [Online]. Available: https://doi.org/10.1016/j.isatra.2018.09.005

[12] M. Phute, A. Sahastrabudhe, S. Pimparkhede, S. Potphode, K. Rengade, and S. Shilaskar, "A survey on machine learning in lithography," in *Proc. Int. Conf. Artif. Intell. Mach. Vis. (AIMV)*, 2021, pp. 1–6. [Online]. Available: https://doi.org/10.1109/aimv53313.2021.9670977

[13] P. Espadinha-Cruz, R. Godina, and E. M. G. Rodrigues, "A review of data mining applications in semiconductor manufacturing," *Processes*, vol. 9, no. 2, p. 305, Feb. 2021. [Online]. Available: https://doi.org/10.3390/pr9020305

[14] D. Stanisavljevic and M. Spitzer, "A review of related work on machine learning in semiconductor manufacturing and assembly lines," in *Proc. SAMI@iKNOW*, 2016, pp. 1–5.

[15] P. Lenhard, A. Kovalenko, and R. Lenhard, "Integrated circuit die level yield prediction using deep learning," in *Proc. 33rd Annu. SEMI Adv. Semicond. Manuf. Conf. (ASMC)*, 2022, pp. 1–6. [Online]. Available: https://doi.org/10.1109/asmc54647.2022.9792526

[16] K. Nakata, R. Orihara, Y. Mizuoka, and K. Takagi, "A comprehensive big-data-based monitoring system for yield enhancement in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 4, pp. 339–344, Nov. 2017. [Online]. Available: https://doi.org/10.1109/tsm.2017.2753251

[17] S. Kang, S. Cho, D. An, and J. Rim, "Using wafer map features to better predict die-level failures in final test," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 3, pp. 431–437, Aug. 2015. [Online]. Available: https://doi.org/10.1109/tsm.2015.2443864

[18] H. Xu, J. Zhang, Y. Lv, and P. Zheng, "Hybrid feature selection for wafer acceptance test parameters in semiconductor manufacturing," *IEEE Access*, vol. 8, pp. 17320–17330, 2020. [Online]. Available: https://doi.org/10.1109/access.2020.2966520

[19] C.-F. Chien, P.-C. Lee, R. Dou, Y.-J. Chen, and C.-C. Chen, "Modeling collinear WATs for parametric yield enhancement in semiconductor manufacturing," in *Proc. 13th IEEE Conf. Autom. Sci. Eng. (CASE)*, 2017, pp. 739–743. [Online]. Available: https://doi.org/10.1109/coase.2017.8256192

[20] K. Chen, P.-Y. Chang, and C.-H. Yeh, "Wafer die yield prediction by heuristic methods," in *Proc. 40th Int. Conf. Comput. Ind. Eng.*, 2010, pp. 1–4. [Online]. Available: https://doi.org/10.1109/iccie.2010.5668273

[21] Q. Xu, C. Xu, and J. Wang, "Forecasting the yield of wafer by using improved genetic algorithm, high dimensional alternating feature selection and SVM with uneven distribution and high-dimensional data," *Autonomous Intelligent Systems*, vol. 2, no. 1, p. 24, Sep. 2022. [Online]. Available: https://doi.org/10.1007/s43684-022-00041-3

[22] D. Jiang, W. Lin, and N. Raghavan, "A novel framework for semiconductor manufacturing final test yield classification using machine learning techniques," *IEEE Access*, vol. 8, pp. 197885–197895, 2020. [Online]. Available: https://doi.org/10.1109/access.2020.3034680

[23] S. H. Park, C.-S. Park, J. S. Kim, S.-S. Kim, J.-G. Baek, and D. An, "Data mining approaches for packaging yield prediction in the post-fabrication process," in *Proc. IEEE Int. Congr. Big Data*, 2013, pp. 363–368. [Online]. Available: https://doi.org/10.1109/bigdata.congress.2013.55

[24] A. Anaya, W. Henning, N. Basantkumar, and J. Oliver, "Yield improvement using advanced data analytics," in *Proc. 30th Annu. SEMI Adv. Semicond. Manuf. Conf. (ASMC)*, 2019, pp. 1–5. [Online]. Available: https://doi.org/10.1109/asmc.2019.8791752

[25] Y. Kong and D. Ni, "A practical yield prediction approach using inline defect metrology data for system-on-chip integrated circuits," in *Proc. 13th IEEE Conf. Autom. Sci. Eng. (CASE)*, 2017, pp. 744–749. [Online]. Available: https://doi.org/10.1109/coase.2017.8256193

[26] K.-J. Kim, K.-J. Kim, C.-H. Jun, I.-G. Chong, and G.-Y. Song, "Variable selection under missing values and unlabeled data in semiconductor processes," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 1, pp. 121–128, Feb. 2019. [Online]. Available: https://doi.org/10.1109/tsm.2018.2881286

[27] C.-F. Chien, Y.-J. Chen, and J.-Z. Wu, "Big data analytics for modeling WAT parameter variation induced by process tool in semiconductor manufacturing and empirical study," in *Proc. Winter Simul. Conf. (WSC)*, 2016, pp. 2512–2522. [Online]. Available: https://doi.org/10.1109/wsc.2016.7822290

[28] M. P. McLaughlin et al., "Enhanced defect detection in after develop inspection with machine learning disposition," in *Proc. 32nd Annu. SEMI Adv. Semicond. Manuf. Conf. (ASMC)*, 2021, pp. 1–5. [Online]. Available: https://doi.org/10.1109/asmc51741.2021.9435721

[29] V. Georgieva, S. Tinck, and A. Bogaerts, "Optimizing the operating conditions for radially uniform plasma and etch rates," in *Proc. 10th Plasma Etch Strip Process. Micro- Nano-Technol. Workshop (PESM2017)*, 2017, pp. 1–3.

[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[31] T. T. Le, W. Fu, and J. H. Moore, "Scaling tree-based automated machine learning to biomedical big data with a feature set selector," *Bioinformatics*, vol. 36, no. 1, pp. 250–256, 2020.