2021 32nd Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC) | 978-1-7281-8645-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ASMC51741.2021.9435673

A hybrid feature selection approach for virtual metrology: Application to CMP process

1st Taki Eddine KORABI
Mines Saint-Etienne,
Univ Clermont Auvergne,
CNRS, UMR 6158 LIMOS,
Center CMP, Department SFL,
F-13541 Gardanne France
takieddine.korabi@gmail.com

2nd Valeria BORODIN
Mines Saint-Etienne,
Univ Clermont Auvergne,
CNRS, UMR 6158 LIMOS,
Center CMP, Department SFL,
F-13541 Gardanne France
valeria.borodin@emse.fr

3rd Michel JUGE
Department of Data Science,
STMicroelectronics,
190 Avenue Célestin Coq,
13106 Rousset, Francey
michel.juge@st.com

4th Agnès ROUSSY
Mines Saint-Etienne,
Univ Clermont Auvergne,
CNRS, UMR 6158 LIMOS,
Center CMP, Department SFL,
F-13541 Gardanne France
agnes.roussy@emse.fr

***Abstract*—This paper presents a feature selection method for virtual metrology applied to a chemical mechanical planarization process in the semiconductor industry. The proposed approach is based on a filter method coupled with a wrapper method. The goal of this article is twofold: (i) to improve the prediction accuracy of the considered machine learning algorithms, and (ii) to reduce the computational time and the storage space in databases. Numerical experiments are conducted on an industrial dataset derived from the chemical mechanical planarization process of a semiconductor manufacturing facility.**

***Index Terms*—Feature selection, hybrid approach, genetic algorithm, virtual metrology, semiconductor manufacturing**

I. Introduction

One of the most important challenges of smart manufacturing lies in the exploitation and extraction of information contained in large volumes of data (big data) [1][2]. In the semiconductor manufacturing industry, big data mainly correspond to the raw data collected on production machines. These data are generated by internal equipment sensors and are often called *Fault Detection and Classification* data (or FDC data in short). The number of sensors can easily reach a hundred and the measurements are often sampled at very high frequencies, which exponentially increase the volume of data.

In semiconductor manufacturing, multiple industry-oriented tasks are confronted with large volumes of data and deal with the extraction of knowledge from collected data, necessary for improving key performance indicators, while minimizing industrial risks, reducing cycle time and improving the product quality. In the semiconductor industry, the three main data-intensive tasks are: **(i)** Virtual Metrology (VM), **(ii)** equipment Fault Detection, Classification (FDC) and diagnosis, and **(iii)** predictive maintenance.

This paper focuses on VM, which consists in using supervised regression algorithms to build a model capable of predicting post-process metrology variables without going through a metrology measurement step, often costly and time consuming. There is an abundant literature on VM in the semiconductor manufacturing field including algorithms ranging from simple (linear regression, simple decision trees, etc.) to more sophisticated ones (neural networks, support vector regression, random forest, etc.). For VM literature, the interested reader is referred to [3][4][5] and references therein.

A big part of the VM task consists in performing the so-called data preparation, including sorting, ordering, rearranging, selecting, and filtering of the input data. It has been shown recently that the choice, the form and especially the selection of the most relevant features often has a greater impact than the algorithm itself in obtaining relevant prediction results [5]. In this sense, the first step is to build data structures (often constructed or projected in two-dimensional arrays), which contain the observations of each run (wafer or batch of wafers) for each sensor at each sampling time. Thus, each row corresponds to a run and is called an *instance*. Each column corresponds to an attribute, and is called a *feature*. However, in many cases solving a supervised regression problem using a high-dimensional input becomes almost intractable because of the too large dimension of these data structures.

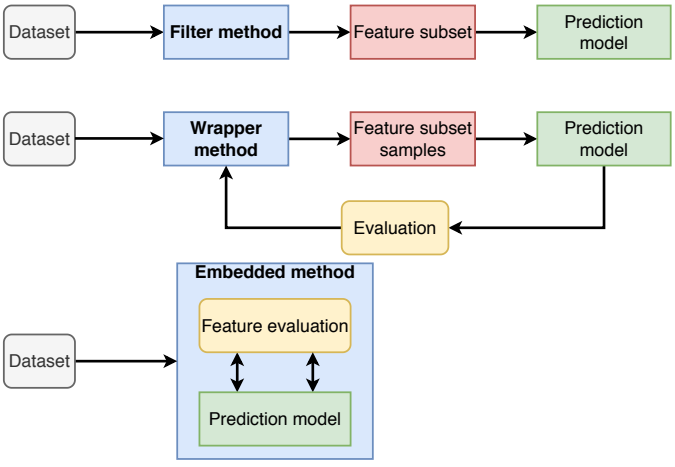

Fig. 1. Feature selection categories [6]

In data science, the size of features is often reduced and not that of instances. The reason is that instances provide information on different batches and therefore an important plus for the construction of the model. However, it is common to use analytical methods for deleting instances that contain

outliers and missing values [7]. We can also use statistical approximation methods for replacing them, so that their impact on the result becomes minimal. In turn, features usually contain redundant information, correlations, noises, constant values or even attributes with no impact on the prediction of the output, which can mislead the Machine Learning (ML) algorithm resulting in poor prediction results.

In this context, two main approaches for the dimensionality reduction have emerged: feature selection and feature projection. The feature projection consists in building a new set of features based on the initial set. Among the most popular methods of feature reduction, one can find the principal component analysis. The main issue related to the feature projection is the loss of information that can result from the linear or non-linear transformation of the initial set of features and the lack of interpretability in the context, for instance, of a root cause analysis. The Feature Selection (FS), in its turn, reduces the initial set of features to a smaller subset, while keeping as much relevant information as possible, thereby eliminating noise and irrelevant features.

The remainder of this paper is organized as follows. The problem under study is introduced and formalized in Section II. The proposed hybrid feature selection approach is given in Section III. Numerical experiments are conducted in Section IV. Concluding remarks and perspectives are provided in Section V.

II. Problem statement

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ be a set of n features, where $n \in \mathbb{N}$ is the total number of features in set $\mathcal{X}$. Denote by $f(\mathcal{X}')$ an evaluation function of subset $\mathcal{X}'$, where $\mathcal{X}' \subseteq \mathcal{X}$. The aim of the feature selection problem is to find a subset $\mathcal{X}^* \subseteq \mathcal{X}$ such that:

$$f(\mathcal{X}^*) = \min_{\mathcal{X}' \subseteq \mathcal{X}} f(\mathcal{X}') \tag{1}$$

Let n' be the cardinality of $\mathcal{X}^*$, where $n' \leq n$. The cardinality of $\mathcal{X}^*$ can be imposed via a constraint.

The impact of solving the feature selection problem on the pertinence and the efficiency of a prediction model is well established [8][9]. First, it reduces computational time, which can be crucial in certain applications like VM. Moreover, it makes easier cause-effect analyses since the reduced set contains only features that have an impact on the target output. Several papers dealing with the feature selection problem have been published in last decades [10]. There are three main classes of FS methods: filter methods, wrapper methods and embedded methods (see Figure 1).

In the case of filters, the evaluation criterion is applied only on the input data, without taking into account the performance of the ML algorithm. In the case of wrappers, the evaluation criterion takes into account the performance of the ML algorithm based on the selected subset of features. According to [11], evaluation criteria can be classified as follows:

- *Informational criterion:* is based on the difference of uncertainty (i.e. *a priori* and *a posteriori* uncertainty). Then, the variables that provide the most information are chosen.
- *Distance-based criterion:* This criterion reveals the most discriminating variables.
- *Independence criterion:* is based on correlation and association functions.
- *Consistence criterion:* The aim of this criterion relies on the elimination of redundant variables. The related feature selection problem consists in finding the smallest subset, which has the smallest *inconsistency rate*.
- *Precision criterion:* is mainly used in the framework of wrappers methods. It is used to evaluate the accuracy of the prediction provided by machine learning algorithms for a selected set of features. For instance, in the case of regression machine learning algorithms, the criterion could be the mean squared error of the difference between the target and the prediction values.

It has been shown that the use of raw data summaries, such as statistical moments and geometrical features, are more suitable for prediction and correlation extraction tasks than the direct use of raw data [5]. Moreover, the *feature-based prediction* is considered to be the *next generation* of regression and classification algorithms notably in the semiconductor manufacturing industry [5]. The first step therefore consists in extracting the right features from the time series. However, one can ask the following question: How to choose the right features? The answer to this question is not obvious, because it can be difficult even to define what a right feature means. The answer to this question often requires in-depth domain knowledge.

One of the solutions to remedy the aforementioned issue would be to calculate only a few indicators like the first two statistical moments. This course of action saves a lot of computational time, and has the advantage of lightening the databases. However, its major drawback is the loss of information, resulting after the summarizing of an entire multivariate time series in only one or two features. In this paper, another solution is proposed, the idea is to get around the problem by calculating as many features as possible, and then selecting the most relevant among them.

III. A hybrid feature selection approach

Embedded methods are out of the scope of this paper since they are integrated directly in some ML algorithms. Filter methods are much faster than wrapper methods. However, it has been shown that wrapper methods are more efficient in selecting features, by exploiting the bias of ML algorithms [8]. In this context, we propose a method which allows us to combine the advantages of both filter and wrapper methods, namely the speed of the filters and the efficiency of the wrappers. The proposed Hybrid Feature Selection (HFS) method allows us to obtain an informative subset of features in a more competitive computational time than the wrapper method alone. Figure 2 describes the proposed HFS method which includes two main stages:

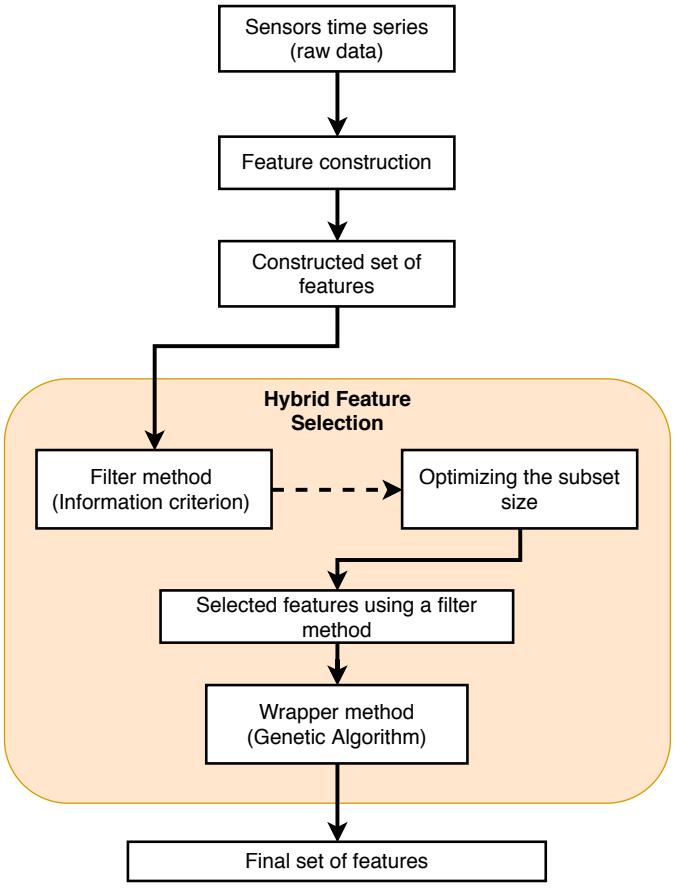

Fig. 2. Diagram of hybrid feature selection method

- *Filter stage:* dedicated to eliminating the irrelevant variables.
- *Wrapper stage*: used to exploit the bias of a given ML algorithm to find the best feature subset. This stage is based on a Genetic Algorithm (GA). Inspired from the natural selection and genetics, a genetic algorithm moves forward from an initial population (i.e. a set of feasible solutions) towards new generations, by applying a reproduction mechanism (materialized by operators such as crossover and mutation). Basically, genetic algorithms operate with chromosomes, which encode solutions of the studied problem, and a fitness function evaluating the quality of chromosomes. The reader is referred to [12] for a detailed description and a survey of genetic algorithms.

Note that, depending on the choice of the evaluation function, genetic algorithms can either be used as filter [13] or wrapper methods [14].

The proposed hybrid feature selection method is constituted of four main parts (see Figure 2):

- *Feature extraction:* In this part, a large number of indicators is used to extract features from time series.
- *Intermediate feature selection*: From the resulting set of constructed features, which contains thousands of features, a filter method is applied by using a distance or information criterion. This step considerably reduces the number of features. Thus, a partial feature selection performed during this stage allows us to quickly eliminate the redundant features, or those which do not explain the output variance.
- From the intermediate set, we apply a second feature selection method to further reduce the initial set. This second method is a wrapper method, which allows us to make more efficient the selection process, by taking into account the impact of selected features on the performance of the machine learning algorithm. Therefore, the wrapper method is used to obtain a final set, which has a significantly reduced size compared to the initial set. The application of the wrapper method on the intermediate set (i.e. after filter method) saves a lot of time.
- In order to avoid the saturation of the communication channels and databases, only the selected features of the final set could only be calculated and collected in several next executions. This also helps to reduce the calculation time.

IV. Numerical Experiments

A. Industrial case study

The proposed HFS method has been tested on a industrial dataset from the Chemical Mecanical Planarization (CMP) process of a semiconductor facility. CMP is a polishing process assisted by chemical reactions to remove surface materials. This process involves the joint action of the polishing pad and the chemical and abrasive effects of the polishing solution, commonly referred to as a slurry (see Figure 3). The wafer is pressed on the pad by the polishing head, on which it is held in place by a retaining ring. The polishing plate (on which the pad is mounted) has the effect of removing material and planarizing any topography. It thus polishes the surface of the wafer, making it smoother.

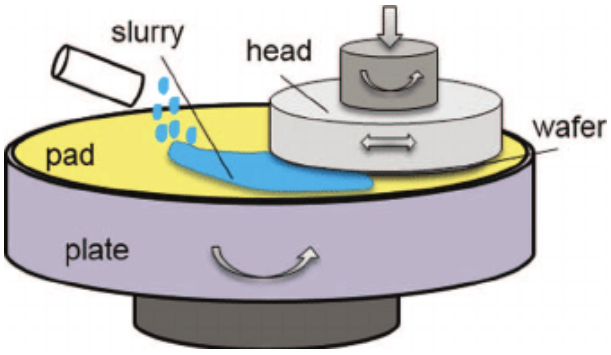

Fig. 3. Functional principle of CMP [15]

The dataset has been provided by STMicroelectronics (Rousset, France), an 8 inch wafer fab [16]. This dataset includes the following parameters:

- *Fault detection diagnosis data* from a Chemical-Mechanical Polishing (CMP) process collected by dedicated sensors e.g.: platen usage, head speed, condition-

ing pressure, etc. FDC data are collected at predefined sampling intervals (often 1 second).
- The *Polishing time* of the second platen associated to each wafer is considered as a controllable variable for the future run-to-run control.
- *Metrology measurements* taken after the CMP process: According to a deterministic sampling policy, two wafers per lot are sampled to recover the metrology measurements, denoted as the post-polishing thicknesses.

Hence, for each of the 546 wafers, we have all the time series provided by the sensors as well as the measurement obtained in metrology at the end of the CMP process. Using the provided FDC data and the extraction procedure described in Section III, more than 14,000 features have been extracted.

B. Genetic algorithm: Parameter tuning

The chromosomes, encoding solutions of the FS problem (i.e. the subset of selected features), are represented via a binary vector of genes. The size of this vector corresponds to the cardinality of the initial set of features. The value 1 indicates the presence (i.e. selection) of a given feature, while the value 0 indicates its absence (or elimination). The order of the binary vector corresponds to the order of the features in the initial dataset. The key parameters of the GA are presented in Table I. Note that these parameters have been chosen in an empirical manner, where several values and strategies have been tested for all parameters (see the column range). The used selection strategy is *tournement*. This method consists of organizing several tournaments among individuals (i.e. chromosomes) chosen randomly. The winner of each tournament is selected for the next stage, the *crossover*. For the crossover step, the uniform strategy is used. In the uniform crossover, we deal with every gene separately. Thus, crossover is carried out point by point and the child has equal chances of inheriting a gene from each of his parents.

TABLE I
SETUP PARAMETERS OF GA

Parameters	**Value**	**Range**
Selection strategy	Tournament	NA
Fraction of survivors	0.4	[0.10, 0.90]
Elitism rate	0.10	[0.05, 0.50]
Crossover strategy	Uniform	NA
Crossover rate	0.6	[0.10, 0.90]
Mutation rate	0.01	[0.01, 0.10]
Stop strategy	10 generations	[5, 35]
Population size	50	[20, 150]

C. Results analysis

The HFS approach has been compared to a wrapper method based on a genetic algorithm and different filters. Three machine learning algorithms have been considered to predict the post-polishing thickness values, namely: Lasso linear regression, support vector regression, and random forest. The considered performance metrics are: the execution time, the root mean squared error (RMSE) and the number of Selected Features (SF) in the final set. Computational experiments have been carried out on an Intel(R) Core(TM)i7-2720 QM CPU2.20 giga-hertz workstation.

Table II (resp. Table IV) provides the obtained results when the ML algorithms is a Lasso linear regression (resp. Support Vector Regression (SVR)). The best filter method for Lasso is the dependence filter (correlation filter), whereas the information filter is the best filter for SVR. Note that, for the correlation filter, several correlation coefficients have been tested and the one with the best results has been used in a grid-like validation process.

Table III summarizes the obtained results when the ML algorithm is a Random Forest (RF). It shows that the HFS approach has the lowest RMSE and it outperforms the GA in terms of both the execution time and the size of the final set. However, the filter method is quicker than the HFS method while the size of the final set is much smaller for HFS with only 65 selected features. In Table III, the best filter (information) has been chosen between variance, dependence, and information filters.

TABLE II
COMPARATIVE RESULTS FOR AN INITIAL SET OF 14,625 FEATURES AND LR AS AN ML ALGORITHM ON THE TRAIN SET

	GA with Lasso	Correlation filter	HFS with Lasso
# SF	4,857	3,265	88
RMSE	48.07	52.98	41.59
Time (min)	37	0.7	2.1

TABLE III
COMPARATIVE RESULTS FOR AN INITIAL SET OF 14,625 FEATURES AND RF AS ML ALGORITHM ON THE TRAIN SET

	GA with RF	Information filter	HFS with RF
# SF	3,891	1,974	65
RMSE (Train set)	32.85	44.98	29.04
Time (min)	37	0.7	4

TABLE IV
COMPARATIVE RESULTS FOR AN INITIAL SET OF 14,625 FEATURES AND SVR AS AN ML ALGORITHM ON THE TRAIN SET

	GA with SVR	Information filter	HFS with SVR
# SF	5,633	1,974	49
RMSE	35.12	41.3	27.17
Time (min)	29	0.7	2.9

TABLE V
HFS WITH SVR ON THE TEST SET: *Focus on the size of selected features*

	Solution 1	Solution 2	Solution 3	Solution 4
# SF	10	49	50	52
RMSE	31.12	30.04	27.17	29.18

We also analyzed several solutions of the hybrid method with different imposed cardinalities. It is found that the solution can degrade in the case where the cardinality is significantly reduced (see e.g. solutions 1 and 4 in Table V).

V. Conclusions

In this paper, a hybrid approach for feature selection has been proposed based on the combination of filter and wrapper methods. This method has been tested and validated on a real-life dataset from a CMP process. A comparative analysis with respect to conventional approaches show that the proposed hybrid method is competitive in terms of prediction accuracy and computational time. The analysis of the selected features has been provided valuable insights about the relevance of a number of features in the CMP process.

Acknowledgment

This paper is conducted in the framework of the project MADEin4, which has received funding from the ECSEL JU (Electronic Components and Systems for European Leadership Joint Undertaking) under grant agreement No 826589. The JU receives support from the European Union's Horizon 2020 research and innovation program and France, Germany, Austria, Italy, Sweden, Netherlands, Belgium, Hungary, Romania and Israel.

References

[1] J. Moyne and J. Iskandar, "Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing," *Processes*, vol. 5, no. 3, p. 39, 2017.

[2] M. S. Reis and G. Gins, "Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis," *Processes*, vol. 5, no. 3, p. 35, 2017.

[3] P. Kang, D. Kim, H.-j. Lee, S. Doh, and S. Cho, "Virtual metrology for run-to-run control in semiconductor manufacturing," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2508–2522, 2011.

[4] G. A. Susto, S. Pampuri, A. Schirru, A. Beghi, and G. De Nicolao, "Multi-step virtual metrology for semiconductor manufacturing: A multilevel and regularization methods-based approach," *Computers & Operations Research*, vol. 53, pp. 328–337, 2015.

[5] K. Suthar, D. Shah, J. Wang, and Q. P. He, "Next-generation virtual metrology for semiconductor manufacturing: A feature-based framework," *Computers & Chemical Engineering*, vol. 127, pp. 140–149, 2019.

[6] M. Babiker, E. Karaarslan, and Y. Hoşcan, "A hybrid feature-selection approach for finding the digital evidence of web application attacks," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, pp. 4102–4117, 05 2019.

[7] H. Kaneko, "Automatic outlier sample detection based on regression analysis and repeated ensemble learning," *Chemometrics and Intelligent Laboratory Systems*, vol. 177, pp. 74–82, 2018.

[8] R. Kohavi, G. H. John *et al.*, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[10] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[11] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[12] C. Reeves and J. E. Rowe, *Genetic algorithms: principles and perspectives: a guide to GA theory*. Springer Science & Business Media, 2002, vol. 20.

[13] P. L. Lanzi, "Fast feature selection with genetic algorithms: a filter approach," in *Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC'97)*. IEEE, 1997, pp. 537–540.

[14] L. Zhuo, J. Zheng, X. Li, F. Wang, B. Ai, and J. Qian, "A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine," in *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images*, vol. 7147. International Society for Optics and Photonics, 2008, p. 71471J.

[15] H. Li, Z. Qu, Q. Zhao, F. Tian, D. Zhao, Y. Meng, and X. Lu, "A reliable control system for measurement on film thickness in copper chemical mechanical planarization system," *The Review of scientific instruments*, vol. 84, p. 125101, 12 2013.

[16] W.-T. Yang, J. Blue, A. Roussy, J. Pinaton, and M. S. Reis, "A physics-informed run-to-run control framework for semiconductor manufacturing," *Expert Systems with Applications*, p. 113424, 2020.